STATISTICAL METHODS FOR SEMICONDUCTOR MANUFACTURING

Semiconductor manufacturing increasingly depends on the use of statistical methods in order to develop and explore new technologies, characterize and optimize existing processes, make decisions about evolutionary changes to improve yields and product quality, and monitor and maintain well-functioning processes. Each of these activities revolves around data: Statistical methods are essential in the planning of experiments to efficiently gather relevant data, the construction and evaluation of models based on data, and decision-making using these models.

Formal statistical assessment is an essential complement to engineering knowledge of known and suspected causal effects and systematic sources of variation. Engineering knowledge is crucial to specify data collection plans that ensure that statistical conclusions are defensible. If data is collected but care has not been not taken to make "fair" comparisons, then the results will not be trusted no matter what statistical method is employed or amount of data collected. At the same time, correct application of statistical methods are also crucial to correct interpretations of the results. Consider a simple scenario: The deposition area in a production fab has been running a standard process recipe for several months, and has monitored defect counts on each wafer. An average of 12 defects per 8 in. wafer had been observed over that time. The deposition engineer believes that a change in the gasket (from a different supplier) on the deposition vacuum chamber can reduce the number of defects. The engineer makes the change and runs two lots of 25 wafers each with the change (observing an average of 10 defects per wafer), followed by two additional lots with the original gasket type (observing an average of 11 defects per wafer). Should the change be made permanently? A "difference" in output has been observed, but a key question remains: Is the change "statistically significant"? That is to say, considering the data collected and the system's characteristics, has a change really occurred or might the same results be explained by chance?

Overview of Statistical Methods

Different statistical methods are used in semiconductor manufacturing to understand and answer questions such as

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*.

that posed above and others, depending on the particular problem. Table 1 summarizes the most common methods and for what purpose they are used, and it serves as a brief outline of this article. In all cases, the issues of sampling plans and significance of the findings must be considered, and all sections will periodically address these issues. To highlight these concepts, note that in Table 1 the words "different," "same," "good," and "improve" are mentioned. These words tie together two critical issues: significance and engineering importance. When applying statistics to the data, one first determines if a statistically significant difference exists; otherwise the data (or experimental effects) are assumed to be the same. Thus, what is really tested is whether there is a difference that is big enough to find statistically. Engineering needs and principles determine whether that difference is "good," the difference actually matters, or the cost of switching to a different process will be offset by the estimated improvements. The size of the difference that can be statistically seen is determined by the sampling plans and the statistical method used. Consequently, engineering needs must enter into the design of the experiments (sampling plan) so that the statistical test will be able to see a difference of the appropriate size. Statistical methods provide the means to determine sampling and significance to meet the needs of the engineer and manager.

In the first section, we focus on the basic underlying issues of statistical distributions, paying particular attention to those distributions typically used to model aspects of semiconductor manufacturing, including the indispensable Gaussian distribution as well as binomial and Poisson distributions (which are key to modeling of defect and yield related effects). An example use of basic distributions is to estimate the interval of oxide thickness in which the engineer is confident that 99% of wafers will reside; based on this interval, the engineer could then decide if the process is meeting specifications and define limits or tolerances for chip design and performance modeling.

In the second section, we review the fundamental tool of statistical inference, the hypothesis test as summarized in Table 1. The hypothesis test is crucial in detecting differences in a process. Examples include: determining if the critical dimensions produced by two machines are different; deciding if adding a clean step will decrease the variance of the critical dimension etch bias; determining if no appreciable increase in particles will occur if the interval between machine cleans is extended by 10,000 wafers; or deciding if increasing the target doping level will improve a device's threshold voltage.

In the third section, we expand upon hypothesis testing to consider the fundamentals of experimental design and analysis of variance, including the issue of sampling required to achieve the desired degree of confidence in the existence of an effect or difference, as well as accounting for the risk in not detecting a difference. Extensions beyond single factor experiments to the design of experiments which screen for effects due to several factors or their interactions are then discussed, and we describe the assessment of such experiments using formal analysis of variance methods. Examples of experimental design to enable decision-making abound in semiconductor manufacturing. For example, one might need to decide if adding an extra film or switching deposition methods will improve reliability, or decide which of three gas distribution plates (each with different hole patterns) provides the most uniform etch process.

In the fourth section, we examine the construction of response surface or regression models of responses as a function of one or more continuous factors. Of particular importance are methods to assess the goodness of fit and error in the model, which are essential to appropriate use of regression models in optimization or decision-making. Examples here include determining the optimal values for temperature and pressure to produce wafers with no more than 2% nonuniformity in gate oxide thickness, or determining if typical variations in plasma power will cause out-of-specification materials to be produced.

Finally, we note that statistical process control (SPC) for monitoring the "normal" or expected behavior of a process is a critical statistical method (1,2). The fundaments of statistical distributions and hypothesis testing discussed here bear directly on SPC; further details on statistical process monitoring and process optimization can be found in SEMICONDUCTOR FACTORY CONTROL AND OPTIMIZATION.

STATISTICAL DISTRIBUTIONS

Semiconductor technology development and manufacturing are often concerned with both continuous parameters (e.g., thin-film thicknesses, electrical performance parameters of transistors) and discrete parameters (e.g., defect counts and yield). In this section, we begin with a brief review of the fundamental probability distributions typically encountered in semiconductor manufacturing, as well as sampling distributions which arise when one calculates statistics based on multiple measurements (3). An understanding of these distributions is crucial to understanding hypothesis testing, analysis of variance, and other inferencing and statistical analysis methods discussed in later sections.

Table 1. Summary of Statistical Methods Typically Used in Semiconductor Manufacturing

Topic	Statistical Method	Purpose
1	Statistical distributions	Basic material for statistical tests. Used to characterize a population based upon a sample.
2	Hypothesis testing	Decide whether data under investigation indicates that elements of concern are the "same" or "different."
3	Experimental design and analysis of variance	Determine significance of factors and models; decompose observed variation into constituent elements.
4	Response surface modeling	Understanding relationships, determine process margin, and optimize process.
5	Categorical modeling	Use when result or response is discrete (such as "very rough," "rough," or "smooth"). Understand relationships, determine process margin, and optimize process.
6	Statistical process control	Determine if system is operating as expected.

Descriptive Statistics

"Descriptive" statistics are often used to concisely present collections of data. Such descriptive statistics are based entirely (and only) on the available empirical data, and they do not assume any underlying probability model. Such descriptions include histograms (plots of relative frequency ϕ_i versus measured values or value ranges x_i in some parameter), as well as calculation of the mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

(where n is the number of values observed) calculation of the median (value in the "middle" of the data with an equal number of observations below and above), and calculation of data percentiles. Descriptive statistics also include the sample variance and sample standard deviation:

$$s_x^2 = \text{Sample Var}\{x\} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2}$$

$$s_x = \text{Sample std. dev.}\{x\} = \sqrt{s_x^2} \tag{3}$$

A drawback to such descriptive statistics is that they involve the study of observed data only and enable us to draw conclusions which relate only to that specific data. Powerful statistical methods, on the other hand, have come to be based instead on probability theory; this allows us to relate observations to some underlying probability model and thus make inferences about the population (the theoretical set of all possible observations) as well as the sample (those observations we have in hand). It is the use of these models that give computed statistics (such as the mean) explanatory power.

Probability Model

Perhaps the simplest probability model of relevance in semiconductor manufacturing is the Bernoulli distribution. A Bernoulli trial is an experiment with two discrete outcomes: success or failure. We can *model* the *a priori* probability (based on historical data or theoretical knowledge) of a success simply as p. For example, we may have aggregate historical data that tells us that line yield is 95% (i.e., that 95% of product wafers inserted in the fab successfully emerge at the end of the line intact). We make the leap from this descriptive information to an assumption of an underlying probability model: we suggest that the probability of any one wafer making it through the line is equal to 0.95. Based on that probability model, we can predict an outcome for a new wafer which has not yet been processed and was not part of the original set of observations. Of course, the use of such probability models involves assumptions—for example, that the fab and all factors affecting line yield are essentially the same for the new wafer as for those used in constructing the probability model.

Normal (Gaussian) Distribution. In addition to discrete probability distributions, continuous distributions also play a crucial role in semiconductor manufacturing. Quite often, one is interested in the probability density function (or pdf) for some parametric value. The most important continuous distribution (in large part due to the central limit theorem) is the Gaussian or normal distribution. We can write that a random variable x is "distributed as" a normal distribution with mean μ and variance σ^2 as $x \sim N(\mu, \sigma^2)$. The probability density function for x is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-1/2[(x-\mu)/\sigma]^2} \tag{4}$$

which is also often discussed in unit normal form through the normalization $z = (x - \mu)/\sigma$, so that $z \sim N(1, 0)$:

$$f(z) = \frac{1}{\sqrt{2\pi}}\, e^{-1/2z^2} \tag{5}$$

Given a continuous probability density function, one can talk about the probability of finding a value in some range. For example, if oxide thickness is normally distributed with $\mu = 100$ Å and $\sigma^2 = 10$ Å^2 (or a standard deviation of 3.16 Å), then the probability of any one such measurement x falling between 105 Å and 120 Å can be determined as

$$\begin{aligned}
&\Pr(105 \le x \le 120) \\
&\quad = \int_{105}^{120} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-1/2[(x-\mu)/\sigma]^2}\, dx \\
&\Pr\left(\frac{105-\mu}{\sigma} \le \frac{x-\mu}{\sigma} \le \frac{120-\mu}{\sigma}\right) \\
&\quad = \Pr(z_l \le z \le z_u) = \int_{z_l}^{z_u} \frac{1}{\sqrt{2\pi}}\, e^{-1/2z^2}\, dz \\
&\quad = \left(\int_{-\infty}^{z_u} \frac{1}{\sqrt{2\pi}}\, e^{-1/2z^2}\, dz\right) - \left(\int_{-\infty}^{z_l} \frac{1}{\sqrt{2\pi}}\, e^{-1/2z^2}\, dz\right) \\
&\quad = \Phi(z_u) - \Phi(z_l) = \Phi(6.325) - \Phi(1.581) = 0.0569
\end{aligned} \tag{6}$$

where $\Phi(z)$ is the cumulative density function for the unit normal, which is available via tables or statistical analysis packages.

We now briefly summarize other common discrete probability mass functions (pmf's) and continuous probability density functions (pdf's) that arise in semiconductor manufacturing, and then we turn to sampling distributions that are also important.

Binomial Distribution. Very often we are interested in the number of successes in repeated Bernoulli trials (that is, repeated "succeed" or "fail" trials). If x is the number of successes in n trials, then x is distributed as a binomial distribution $x \sim \text{B}(n, p)$, where p is the probability of each individual "success." The pmf is given by:

$$f(x, p, n) = \binom{n}{x} p^x (1-p)^{n-x} \tag{7}$$

where "n choose x" is

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{8}$$

For example, if one is starting a 25-wafer lot in the fab above, one may wish to know what is the probability that some number x (x being between 0 and 25) of those wafers will survive. For the line yield model of $p = 95\%$, these proba-

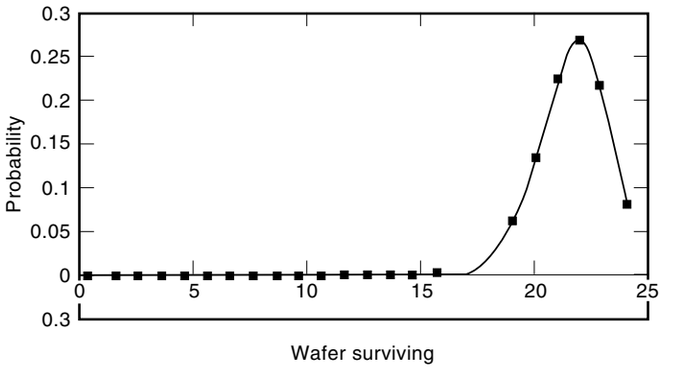

Figure 1. Probabilities for number of wafers surviving from a lot of 25 wafers, assuming a line yield of 95%, calculated using the binomial distribution.

bilities are shown in Fig. 1. When n is very large (much larger than 25), the binomial distribution is well approximated by a Gaussian distribution.

Poisson Distribution. A third discrete distribution is highly relevant to semiconductor manufacturing. An approximation to the binomial distribution that applies when n is large and p is small is the Poisson distribution:

$$f(x, \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{9}$$

for integer $\lambda = 0, 1, 2, \ldots$ and $\lambda \cong np$.

For example, one can examine the number of chips that fail on the first day of operation: The probability p of failure for any one chip is (hopefully) exceedingly small, but one tests a very large number n of chips, so that the observation of the mean number of failed chips $\lambda = np$ is Poisson-distributed. An even more common application of the Poisson model is in defect modeling. For example, if defects are Poisson-distributed with a mean defect count of 3 particles per 200 mm wafer, one can ask questions about the probability of observing x [e.g., Eq. (9)] defects on a sample wafer. In this case, $f(9, 3) = e^{-3}3^9/9! = 0.0027$, or less than 0.3% of the time would we expect to observe exactly 9 defects. Similarly, the probability that 9 or more defects are observed is $1 - \sum_{x=0}^{\infty} f(x, 3) = 0.0038$.

In the case of defect modeling, several other distributions have historically been used, including the exponential, hypergeometric, modified Poisson, and negative binomial. Substantial additional work has been reported in yield modeling to account for clustering and to understand the relationship between defect models and yield (e.g., see Ref. 4).

Population Versus Sample Statistics

We now have the beginnings of a statistical inference theory. Before proceeding with formal hypothesis testing in the next section, we first note that the earlier descriptive statistics of mean and variance take on new interpretations in the probabilistic framework. The mean is the expectation (or "first moment") over the distribution:

$$\begin{aligned}\mu_x = E\{x\} &= \int_{-\infty}^{\infty} x f(x)\, dx \\ &= \sum_{i=1}^{n} x_i \cdot p_r(x_i)\end{aligned} \tag{10}$$

for continuous pdf's and discrete pmf's, respectively. Similarly, the variance is the expectation of the squared deviation from the mean (or the "second central moment") over the distribution:

$$\begin{aligned}\sigma_x^2 = \mathrm{Var}\{x\} &= \int_{-\infty}^{\infty} (x - E\{x\})^2 f(x)\, dx \\ &= \sum_{i=1}^{n} (x_i - E\{x_i\})^2 \cdot p_r(x_i)\end{aligned} \tag{11}$$

Further definitions from probability theory are also highly useful, including the covariance:

$$\begin{aligned}\sigma_{xy}^2 = \mathrm{Cov}\{x, y\} &= E\{(x - E\{x\})(y - E\{y\})\} \\ &= E\{xy\} - E\{x\}E\{y\}\end{aligned} \tag{12}$$

where x and y are each random variables with their own probability distributions, as well as the related correlation coefficient:

$$\rho_{xy} = \mathrm{Corr}\{x, y\} = \frac{\mathrm{Cov}\{x, y\}}{\sqrt{\mathrm{Var}\{x\}\mathrm{Var}\{y\}}} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y} \tag{13}$$

The above definitions for the mean, variance, covariance, and correlation all relate to the underlying or assumed *population.* When one only has a sample (that is, a finite number of values drawn from some population), one calculates the corresponding *sample statistics.* These are no longer "descriptive" of only the sample we have; rather, these statistics are now estimates of parameters in a probability model. Corresponding to the population parameters above, the sample mean $\bar{x}$ is given by Eq. (1), the sample variance s_x^2 is given by Eq. (2), the sample std. dev. s_x is given by Eq. (3), the sample covariance is given by

$$s_{xy}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{14}$$

and the sample correlation coefficient is given by

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{15}$$

Sampling Distributions

Sampling is the act of making inferences about populations based on some number of observations. Random sampling is especially important and desirable, where each observation is independent and identically distributed. A statistic is a function of sample data which contains no further unknowns (e.g., the sample mean can be calculated from the observations and has no further unknowns). It is important to note that a sample statistic is itself a random variable and has a "sampling distribution" which is usually *different* than the underlying population distribution. In order to reason about the "likelihood" of observing a particular statistic (e.g., the mean of five measurements), one must be able to construct the underlying sampling distribution.

Sampling distributions are also intimately bound up with *estimation* of population distribution parameters. For example, suppose we know that the thickness of gate oxide (at the center of the wafer) is normally distributed: $T_i \sim N(\mu, \sigma^2) = N(100, 10)$. We sample 5 random wafers and compute the

mean oxide thickness $\overline{T} = \frac{1}{5}(T_1 + T_2 + \cdots + T_5)$. We now have two key questions: (1) What is the *distribution* of $\overline{T}$? (2) What is the probability that $a \le \overline{T} \le b$?

In this case, given the expression for $\overline{T}$ above, we can use the fact that the variance of a scaled random variable ax is simply $a^2\text{Var}\{x\}$, and the variance of a sum of independent random variables is the sum of the variances:

$$\overline{T} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{16}$$

where $\mu_{\overline{T}} = \mu_T = \mu$ by the definition of the mean.

Thus, when we want to reason about the likelihood of observing values of $\overline{T}$ (that is, averages of five sample measurements) lying within particular ranges, we must be sure to use the distribution for $\overline{T}$ rather than that for the underlying distribution T. Thus, in this case the probability of finding a value $\overline{T}$ between 105 Å and 120 Å is

$$\begin{aligned}\Pr(105 \le \overline{T} \le 120) &= \Pr\left(\frac{105-\mu}{\sigma/(\sqrt{n})} \le \frac{\overline{T}-\mu}{\sigma/(\sqrt{n})} \le \frac{120-\mu}{\sigma/(\sqrt{n})}\right) \\ &= \Pr\left(\frac{105-100}{(\sqrt{10})/(\sqrt{5})} \le \frac{\overline{T}-\mu}{\sigma/(\sqrt{n})} \le \frac{120-100}{\sqrt{2}}\right) \\ &= \Pr(z_l \le z \le z_u) \\ &= \Phi(14.142) - \Phi(3.536) = 0.0002 \end{aligned} \tag{17}$$

which is relatively unlikely. Compare this to the result from Eq. (6) for the probability 0.0569 of observing a single value (rather than a five sample average) in the range 105 Å to 120 Å.

Chi-Square Distribution and Variance Estimates. Several other vitally important distributions arise in sampling, and are essential to making statistical inferences in experimental design or regression. The first of these is the chi-square distribution. If $x_i \sim N(0, 1)$ for $i = 1, 2, \ldots, n$ and $y = x_1^2 + x_2^2 + \cdots + x_n^2$, then y is distributed as chi-square with n degrees of freedom, written as $y \sim \chi_n^2$. While formulas for the probability density function for χ^2 exist, they are almost never used directly and are again instead tabulated or available via statistical packages. The typical use of the χ^2 is for finding the distribution of the variance when the mean is known.

Suppose we know that $x_i \sim N(\mu, \sigma^2)$. As discussed previously, we know that the mean over our n observations is distributed as $\bar{x} \sim N(\mu, \sigma^2/n)$. How is the sample variance s^2 over our n observations distributed? We note that each $(x_i - \bar{x}) \sim N(0, \sigma^2)$ is normally distributed; thus if we normalize our sample variance s^2 by σ^2 we have a chi-square distribution:

$$\begin{aligned} s^2 &= \left(\sum_i (x_i - \bar{x})^2\right) \Big/ (n-1) \\ \frac{(n-1)s^2}{\sigma^2} &\sim \chi_{n-1}^2 \\ s^2 &\sim \left[\frac{\sigma^2}{(n-1)}\right] \cdot \chi_{n-1}^2 \end{aligned} \tag{18}$$

where one degree of freedom is used in calculation of $\bar{x}$. Thus, the sample variance for n observations drawn from $N(\mu, \sigma^2)$ is distributed as chi-square as shown in Eq. (18).

Student *t* Distribution. The Student t distribution is another important sampling distribution. The typical use is when we want to find the distribution of the sample mean when the true standard deviation σ is not known. Consider

$$\begin{aligned} x_i &\sim N(\mu, \sigma^2) \\ \frac{\bar{x}-\mu}{s/(\sqrt{n})} &= \frac{\left(\frac{\bar{x}-\mu}{\sigma/(\sqrt{n})}\right)}{s/\sigma} \sim \frac{N(0,1)}{\sqrt{\frac{1}{n-1}\chi_{n-1}^2}} \sim t_{n-1} \end{aligned} \tag{19}$$

In the above, we have used the definition of the Student t distribution: If z is a normal random variable, $z \sim N(0, 1)$, then $z/\sqrt{y/k}$ is distributed as a Student t with k degrees of freedom, or $z/\sqrt{y/k} \sim t_k$, if y is a random variable distributed as χ_k^2. As discussed previously, the normalized sample variance s^2/σ^2 is chi-square-distributed, so that our definition does indeed apply. We thus find that the normalized sample mean is distributed as a Student t with $n - 1$ degrees of freedom when we do not know the true standard deviation and must estimate it based on the sample as well. We note that as $k \to \infty$, the Student t approaches a unit normal distribution $t_k \to N(0, 1)$.

***F* Distribution and Ratios of Variances.** The last sampling distribution we wish to discuss here is the F distribution. We shall see that the F distribution is crucial in analysis of variance (ANOVA) and experimental design in determining the significance of effects, because the F distribution is concerned with the probability density function for the ratio of variances. If $y_1 \sim \chi_u^2$ (that is, y_1 is a random variable distributed as chi-square with u degrees of freedom) and similarly $y_2 \sim \chi_v^2$, then the random variable $Y = (y_1/u)/(y_2/v) \sim F_{u,v}$ (that is, distributed as F with u and v degrees of freedom). The typical use of the F distribution is to compare the *spread* of two distributions. For example, suppose that we have two samples $x_1, x_2, \ldots, x_n$ and $w_1, w_2, \ldots, w_m$, where $x_i \sim N(\mu_x, \sigma_x^2)$ and $w_i \sim N(\mu_w, \sigma_w^2)$. Then

$$\frac{s_x^2/\sigma_x^2}{s_w^2/\sigma_w^2} \sim F_{n-1,m-1} \tag{20}$$

Point and Interval Estimation

The above population and sampling distributions form the basis for the statistical inferences we wish to draw in many semiconductor manufacturing examples. One important use of the sampling distributions is to estimate population parameters based on some number of observations. A *point estimate* gives a single "best" estimated value for a population parameter. For example, the sample mean $\bar{x}$ is an estimate for the population mean μ. Good point estimates are representative or unbiased (that is, the expected value of the estimate should be the true value), as well as minimum variance (that is, we desire the estimator with the smallest variance in that estimate). Often we restrict ourselves to linear estimators; for example, the best linear unbiased estimator (BLUE) for various parameters is typically used.

Many times we would like to determine a confidence interval for estimates of population parameters; that is, we want to know how likely it is that $\bar{x}$ is within some particular range of μ. Asked another way, to a desired probability, where will

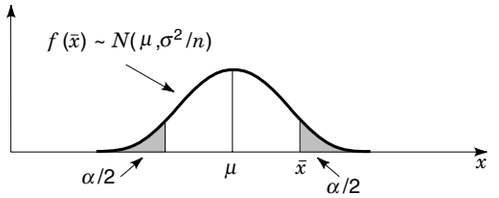

Figure 2. Probability density function for the sample mean. The sample mean is unbiased (the expected value of the sample mean is the true mean μ). The shaded portion in the tail captures the probability that the sample mean is greater than a distance $|\bar{x} - \mu|$ from the true mean.

μ actually lie given an estimate $\bar{x}$? Such interval estimation is considered next; these intervals are used in later sections to discuss hypothesis testing.

First, let us consider the confidence interval for estimation of the mean, when we know the true variance of the process. Given that we make n independent and identically distributed samples from the population, we can calculate the sample mean as in Eq. (1). As discussed earlier, we know that the sample mean is normally distributed as shown in Eq. (16). We can thus determine the probability that an observed $\bar{x}$ is larger than μ by a given amount:

$$z = \frac{\bar{x} - \mu}{\sigma/(\sqrt{n})}$$
$$Pr(z > z_\alpha) = \alpha = 1 - \Phi(z_\alpha) \tag{21}$$

where z_α is the alpha percentage point for the normalized variable z (that is, z_α measures how many standard deviations greater than μ we must be in order for the integrated probability density to the right of this value to equal α). As shown in Fig. 2, we are usually interested in asking the question the other way around: To a given probability [$100(1 - \alpha)$, e.g., 95%], in what range will the true mean lie given an observed $\bar{x}$?

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \le \mu \le \left(\bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$
$$\mu = \bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \tag{22}$$

For our oxide thickness example, we can now answer an important question. If we calculate a five-wafer average, how far away from the average 100 Å must the value be for us to decide that the process has changed away from the mean (given a process variance of 10 Å^2)? In this case, we must define the "confidence" $1 - \alpha$ in saying that the process has changed; this is equivalent to the probability of observing the deviation by chance. With 95% confidence, we can declare that the process is different than the mean of 100 Å if we observe $\overline{T} < 97.228$ Å or $\overline{T} > 102.772$ Å:

$$\frac{\alpha}{2} = \Pr\left(z_{\alpha/2} \le \frac{\overline{T} - \mu}{\sigma/(\sqrt{n})}\right) = \Pr\left(-1.95996 \le \frac{\overline{T} - \mu}{\sigma/(\sqrt{n})}\right)$$
$$|\overline{T} - \mu| = 1.95996\sigma/(\sqrt{n}) = 2.7718 \tag{23}$$

A similar result occurs when the true variance is not known. The $100(1 - \alpha)$ confidence interval in this case is determined using the appropriate Student-t sampling distribution for the sample mean:

$$\left(\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right) \le \mu \le \left(\bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right)$$
$$\mu = \bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} \tag{24}$$

In some cases, we may also desire a confidence interval on the estimate of variance:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \tag{25}$$

Many other cases (e.g., one-sided confidence intervals) can also be determined based on manipulation of the appropriate sampling distributions, or through consultation with more extensive texts (5).

HYPOTHESIS TESTING

Given an underlying probability distribution, it now becomes possible to answer some simple, but very important, questions about any particular observation. In this section, we formalize the decision-making earlier applied to our oxide thickness example. Suppose as before that we know that oxide thickness is normally distributed, with a mean of 100 Å and standard deviation of 3.162 Å. We may know this based on a very large number of previous historical measurements, so that we can well approximate the true population of oxide thicknesses out of a particular furnace with these two distribution parameters. We suspect something just changed in the equipment, and we want to determine if there has been an impact on oxide thickness. We make a new observation (i.e., run a new wafer and form our oxide thickness value as usual, perhaps as the average of nine measurements at fixed positions across the wafer). The key question is: What is the probability that we would get this observation if the process has *not* changed, versus the probability of getting this observation if the process has indeed changed?

We are conducting a hypothesis test. Based on the observation, we want to test the hypothesis (label this H_1) that the underlying distribution mean has increased from μ_0 by some amount δ to μ_1. The "null hypothesis" H_0 is that nothing has changed and the true mean is still μ_0. We are looking for evidence to convince us that H_1 is true.

We can plot the probability density function for each of the two hypotheses under the assumption that the variance has not changed, as shown in Fig. 3. Suppose now that we observe

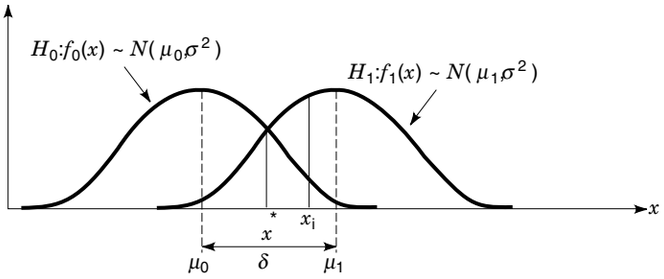

Figure 3. Distributions of x under the null hypothesis H_0, and under the hypothesis H_1 that a positive shift has occurred such that $\mu_1 = \mu_0 + \delta$.

the value x_i. Intuitively, if the value of x_i is "closer" to μ_0 than to μ_1, we will more than likely believe that the value comes from the H_0 distribution than the H_1 distribution. Under a maximum likelihood approach, we can compare the probability density functions:

$$f_0(x_i) \underset{H_1}{\overset{H_0}{\gtrless}} f_1(x_i) \tag{26}$$

that is, if f_1 is greater than f_0 for our observation, we reject the null hypothesis H_0 (that is, we "accept" the alternative hypothesis H_1). If we have prior belief (or other knowledge) affecting the a priori probabilities of H_0 and H_1, these can also be used to scale the distributions f_0 and f_1 to determine *a posteriori* probabilities for H_0 and H_1. Similarly, we can define the "acceptance region" as the set of values of x_i for which we accept each respective hypothesis. In Fig. 3, we have the rule: Accept H_1 if $x_i > x^*$, and accept H_0 if $x_i < x^*$.

In typical use, we select a confidence $1 - \alpha$ with which we must detect a "difference," and we pick a x^* decision point based on that confidence. For two-sided detection with a unit normal distribution, for example, we select $z > z_{\alpha/2}$ and $z < -z_{\alpha/2}$ as the regions for declaring that unusual behavior (i.e., a shift) has occurred.

Alpha and Beta Risk (Type I and Type II Errors)

The hypothesis test gives a clear, unambiguous procedure for making a decision based on the distributions and assumptions outlined above. Unfortunately, there may be a substantial probability of making the *wrong* decision. In the maximum likelihood example of Fig. 3, for the single observation x_i as drawn we accept the alternative hypothesis H_1. However, examining the distribution corresponding to H_0, we see that a nonzero probability exists of x_i belonging to H_1. Two types of errors are of concern:

$$\begin{aligned}\alpha &= \Pr(\text{Type I error}) = \Pr(\text{reject } H_0 | H_0 \text{ is true}) = \int_{x^*}^{\infty} f_0(x)\,dx \\ \beta &= \Pr(\text{Type II error}) = \Pr(\text{accept } H_0 | H_1 \text{ is true}) = \int_{-\infty}^{x^*} f_1(x)\,dx\end{aligned} \tag{27}$$

We note that the Type I error (or probability α of a "false alarm") is based entirely on our decision rule and does not depend on the size of the shift we are seeking to detect. The Type II error (or probability β of "missing" a real shift), on the other hand, depends strongly on the size of the shift. This Type II error can be evaluated for the distributions and decision rules above; for the normal distribution of Fig. 3, we find

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\delta}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\delta}{\sigma}\right) \tag{28}$$

where δ is the shift we wish to detect.

The "power" of a statistical test is defined as

$$\text{Power} \equiv 1 - \beta = \Pr(\text{reject } H_0 | H_0 \text{ is false}) \tag{29}$$

that is, the power of the test is the probability of *correctly* rejecting H_0. Thus the power depends on the shift δ we wish to detect as well as the level of "false alarms" (α or Type I error) we are willing to endure. Power curves are often plotted of β versus the normalized shift to detect $d = \delta/\sigma$, for a given fixed α (e.g., $\alpha = 0.05$), and as a function of sampling size.

Hypothesis Testing and Sampling Plans. In the previous discussion, we described a hypothesis test for detection of a shift in mean of a normal distribution, based on a single observation. In realistic situations, we can often make multiple observations and improve our ability to make a correct decision. If we take n observations, then the sample mean is normally distributed, but with a reduced variance σ^2/n as illustrated by our five-wafer average of oxide thickness. It now becomes possible to pick an α risk associated with a decision on the *sampling distribution* that is acceptable, and then determine a sample size n in order to achieve the desired level of β risk. In the first step, we still select $z_{\alpha/2}$ based on the risk of false alarm α, but now this determines the actual unnormalized decision point using $x^* = \mu + z_{\alpha/2} \cdot \sigma/(\sqrt{n})$. We finally pick n (which determines x^* as just defined) based on the Type II error, which also depends on the size of the normalized shift d = δ/σ to be detected and the sample size n. Graphs and tables are indispensible in selecting the sample size for a given d and α; for example, Fig. 4 shows the Type II error associated with sampling from a unit normal distribution, for a fixed Type I error of 0.05, and as a function of sampling size n and shift to be detected d.

Control Chart Application. The concepts of hypothesis testing, together with issues of Type I and Type II error as well as sample size determination, have one of their most common applications in the design and use of control charts. For example, the $\bar{x}$ control chart can be used to detect when a "significant" change from "normal" operation occurs. The assumption here is that when the underlying process is operating under control, the fundamental process population is distributed normally as $x_i \sim N(\mu, \sigma^2)$. We periodically draw a sample of n observations and calculate the average of those observations ($\bar{x}$). In the control chart, we essentially perform a continuous hypothesis test, where we set the upper and lower control limits (UCLs and LCLs) such that $\bar{x}$ falls outside these control charts with probability α when the process is truly under control (e.g., we usually select 3σ to give a 0.27% chance of false alarm). We would then choose the sample size n so as to have a particular power (that is, a probability of actually detecting a shift of a given size) as previously described. The control

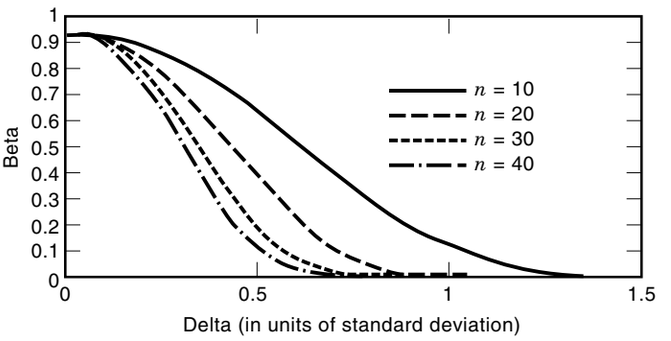

Figure 4. Probability of Type II error (β) for a unit normal process with Type I error fixed at $\alpha = 0.05$, as a function of sample size n and shift delta (equal to the normalized shift δ/σ) to be detected.

limits (CLs) would then be set at

$$\mathrm{CL} = \mu \pm \left(z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right) \tag{30}$$

Note that care must be taken to check key assumptions in setting sample sizes and control limits. In particular, one should verify that each of the n observations to be utilized in forming the sample average (or the sample standard deviation in an s chart) are independent and identically distributed. If, for example, we aggregate or take the average of n successive wafers, we must be sure that there are no systematic within-lot sources of variation. If such additional variation sources do exist, then the estimate of variance formed during process characterization (e.g., in our etch example) will be inflated by accidental inclusion of this systematic contribution, and any control limits one sets will be wider than appropriate to detect real changes in the underlying random distribution. If systematic variation does indeed exist, then a new statistic should be formed that blocks against that variation (e.g., by specifying which wafers should be measured out of each lot), and the control chart based on the distribution of that aggregate statistic.

EXPERIMENTAL DESIGN AND ANOVA

In this section, we consider application and further development of the basic statistical methods already discussed to the problem of designing experiments to investigate particular effects, and to aid in the construction of models for use in process understanding and optimization. First, we should recognize that the hypothesis testing methods are precisely those needed to determine if a new treatment induces a significant effect in comparison to a process with known distribution. These are known as *one-sample* tests. In this section, we consider *two-sample* tests to compare two treatments in an effort to detect a treatment effect. We will then extend to the analysis of experiments in which many treatments are to be compared (*k-sample* tests), and we present the classic tool for studying the results—ANOVA (6).

Comparison of Treatments: Two-Sample Tests

Consider an example where a new process B is to be compared against the process of record (POR), process A. In the simplest case, we have enough historical information on process A that we assume values for the yield mean μ_A and standard deviation σ. If we gather a sample of 10 wafers for the new process, we can perform a simple one-sample hypothesis test $\mu_B > \mu_A$ using the 10 wafer sampling distribution and methods already discussed, assuming that both process A and B share the same variance.

Consider now the situation where we want to compare two new processes. We will fabricate 10 wafers with process A and another 10 wafers with process B, and then we will measure the yield for each wafer after processing. In order to block against possible time trends, we alternate between process A and process B on a wafer by wafer basis. We are seeking to test the hypothesis that process B is better than process A—that is, H_1: $\mu_B > \mu_A$, as opposed to the null hypothesis H_0: $\mu_B = \mu_A$. Several approaches can be used (5).

In the first approach, we assume that in each process A and B we are random sampling from an underlying normal distribution, with (1) an unknown mean for each process and (2) a constant known value for the population standard deviation of σ. Then $n_A = 10$, $n_B = 10$, $\mathrm{Var}\{\bar{y}_A\} = \sigma^2/n_A$, and $\mathrm{Var}\{\bar{y}_B\} = \sigma^2/n_B$. We can then construct the sampling distribution for the difference in means as

$$\mathrm{Var}\{\bar{y}_B - \bar{y}_A\} = \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B} = \sigma^2 \left(\frac{1}{n_A} + \frac{1}{n_B} \right) \tag{31}$$

$$s_{B-A} = \sigma \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \tag{32}$$

Even if the original process is moderately nonnormal, the distribution of the difference in sample means will be approximately normal by the central limit theorem, so we can normalize as

$$z_0 = \frac{(\bar{y}_B - \bar{y}_A) - (\bar{\mu}_B - \bar{\mu}_A)}{\sigma \sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}} \tag{33}$$

allowing us to examine the probability of observing the mean difference $\bar{y}_B - \bar{y}_A$ based on the unit normal distribution, $\Pr(z > z_0)$.

The disadvantage of the above method is that it depends on knowing the population standard deviation σ. If such information is indeed available, using it will certainly improve the ability to detect a difference. In the second approach, we assume that our 10 wafer samples are again drawn by random sampling on an underlying normal population, but in this case we do not assume that we know *a priori* what the population variance is. In this case, we must also build an internal estimate of the variance. First, we estimate from the individual variances:

$$s_A^2 = \frac{1}{n_A - 1} \sum_{i=1}^{n_A} (y_{A_i} - \bar{y}_A) \tag{34}$$

and similarly for s_B^2. The pooled variance is then

$$s^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} \tag{35}$$

Once we have an estimate for the population variance, we can perform our t test using this pooled estimate:

$$t_0 = \frac{(\bar{y}_B - \bar{y}_A) - (\bar{\mu}_B - \bar{\mu}_A)}{s \sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}} \tag{36}$$

One must be careful in assuming that process A and B share a common variance; in many cases this is not true and more sophisticated analysis is needed (5).

Comparing Several Treatment Means Via ANOVA

In many cases, we are interested in comparing several treatments simultaneously. We can generalize the approach discussed above to examine if the observed differences in treatment means are indeed significant, or could have occurred by chance (through random sampling of the same underlying population).

A picture helps explain what we are seeking to accomplish. As shown in Fig. 5, the population distribution for each treatment is shown; the mean can differ because the treatment is shifted, while we assume that the population variance is fixed. The sampling distribution for each treatment, on the other hand, may also be different if a different number of samples is drawn from each treatment (an "unbalanced" experiment), or because the treatment is in fact shifted. In most of the analyses that follow, we will assume balanced experiments; analysis of the unbalanced case can also be performed (7). It remains important to recognize that particular sample values from the sampling distributions are what we measure. In essence, we must compare the variance between two groups (a measure of the potential "shift" between treatments) with the variance within each group (a measure of the sampling variance). Only if the shift is "large" compared to the sampling variance are we confident that a true effect is in place. An appropriate sample size must therefore be chosen using methods previously described such that the experiment is powerful enough to detect differences between the treatments that are of engineering importance. In the following, we discuss the basic methods used to analyze the results of such an experiment (8).

First, we need an estimate of the within-group variation. Here we again assume that each group is normally distributed and share a common variance σ^2. Then we can form the "sum of squares" deviations within the tth group SS_t as

$$SS_t = \sum_{j=1}^{n_t} (y_{t_j} - \overline{y_t})^2 \tag{37}$$

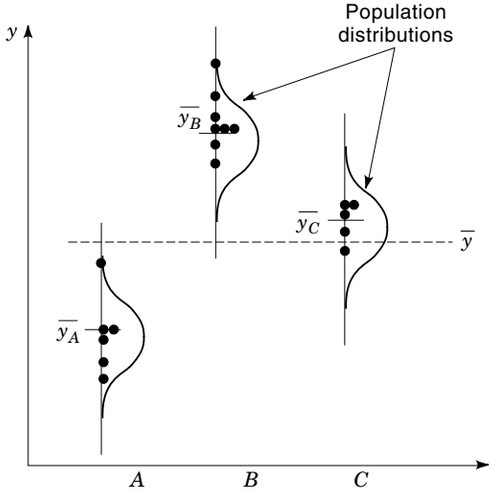

Figure 5. Pictorial representation of multiple treatment experimental analysis. The underlying population for each treatment may be different, while the variance for each treatment population is assumed to be constant. The dark circles indicate the sample values drawn from these populations (note that the number of samples may not be the same in each treatment).

where n_t is the number of observations or samples in group t. The estimate of the sample variance for each group, also referred to as the "mean square," is thus

$$s_t^2 = \frac{SS_t}{\upsilon_t} = \frac{SS_t}{n_t - 1} \tag{38}$$

where υ_t is the degrees of freedom in treatment t. We can generalize the pooling procedure used earlier to estimate a common shared variance across k treatments as

$$s_R^2 = \frac{\upsilon_1 s_1^2 + \upsilon_2 s_2^2 + \cdots + \upsilon_k s_k^2}{\upsilon_1 + \upsilon_2 + \cdots + \upsilon_k} = \frac{SS_R}{N - k} = \frac{SS_R}{\upsilon_R} \tag{39}$$

where SS_R/υ_R is defined as the within-treatment or within-group mean square and N is the total of measurements, $N = n_1 + n_2 + \cdots + n_k$, or simply $N = n_k$ is all k treatments consist of the same number of samples, n.

In the second step, we want an estimate of between-group variation. We will ultimately be testing the hypothesis $\mu_1 = \mu_2 = \cdots = \mu_k$. The estimate of the between-group variance is

$$s_T^2 = \sum_{t=1}^{k} n_t (\overline{y_t} - \overline{y})^2 = \frac{SS_T}{\upsilon_T} \tag{40}$$

where $\upsilon_T = k - 1$, SS_T/υ_T is defined as the between-treatment mean square, and $\overline{y}$ is the overall mean.

We are now in position to ask our key question: Are the treatments different? If they are indeed different, then the between group variance will be larger than the within-group variance. If the treatments are the same, then the between-group variance should be the same as the within-group variance. If in fact the treatments are different, we thus find that

$$s_T^2 \text{ estimates } \left(\sigma^2 + \sum_{t=1}^{k} \frac{n_t \tau_t^2}{(k-1)} \right) \tag{41}$$

where $\tau_t = \mu_t - \mu$ is treatment t's effect. That is, s_T^2 is *inflated* by some factor related to the difference between treatments. We can perform a formal statistical test for treatment significance. Specifically, we should consider the evidence to be strong for the treatments being different if the ratio s_T^2/s_R^2 is significantly larger than 1. Under our assumptions, this should be evaluated using the F distribution, since $s_T^2/s_R^2 \sim F_{k-1,N-k}$.

We can also express the total variation (total deviation sum of squares from the grand mean SS_D) observed in the data as

$$\begin{aligned} SS_D &= \sum_{t=1}^{k} \sum_{i=1}^{n_t} (y_{ti} - \overline{y}^2) \\ s_D^2 &= \frac{SS_D}{\upsilon_D} = \frac{SS_D}{N-1} \end{aligned} \tag{42}$$

where S_D^2 is recognized as the variance in the data.

Analysis of variance results are usually expressed in tabular form, such as shown in Table 2. In addition to compactly summarizing the sum of squares due to various components and the degrees of freedom, the appropriate F ratio is shown. The last column of the table usually contains the probability

Table 2. Structure of the Analysis of Variance Table, for Single-Factor (Treatment) Case

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio	Pr(F)
Between treatments	SS_T	$\upsilon_T = k - 1$	s_T^2	s_T^2/s_R^2	p
Within treatments	SS_R	$\upsilon_R = N - k$	s_R^2		
Total about the grand mean	$SS_D = SS_T + SS_R$	$\upsilon_D = \upsilon_T + \upsilon_R = N - 1$	s_D^2		

of observing the stated F ratio under the null hypothesis. The alternative hypothesis—that one or more treatments have a mean different than that of the others—can be accepted with $100(1 - p)$ confidence.

As summarized in Table 2, analysis of variance can also be pictured as a decomposition of the variance observed in the data. That is, we can express the total sum of squared deviations from the grand mean as $SS_D = SS_T + SS_R$, or the between-group sum of squares added to the within-treatment sum of squares. One can further decompose the total sum of squares which includes the sum of squares due to the average: $SS = SS_A + SS_D = SS_A + SS_T + SS_R$, where $SS_A = N\bar{y}^2$.

While often not explicitly stated, the above ANOVA assumes a mathematical model:

$$y_{ti} = \mu_t + \epsilon_{ti} = \mu + \tau_t + \epsilon_{ti} \tag{43}$$

where μ_t are the treatment means, and ϵ_{ti} are the residuals:

$$\epsilon_{ti} = y_{ti} - \hat{y}_{ti} \sim N(0, \sigma^2) \tag{44}$$

where $\hat{y}_{ti} = \hat{y}_t = \mu + \tau_t$ is the estimated treatment mean. It is critical that one check the resulting ANOVA model. First, the residuals ϵ_{ti} should be plotted against the time order in which the experiments were performed in an attempt to distinguish any time trends. While it is possible to randomize against such trends, we lose resolving power if the trend is large. Second, one should examine the distribution of the residuals. This is to check the assumption that the residuals are "random" [that is, independent and identically distributed (IID) and normally distributed with zero mean] and look for gross non-normality. This check should also include an examination of the residuals for each treatment group. Third, one should plot the residuals versus the estimates and be especially alert to dependencies on the size of the estimate (e.g., proportional versus absolute errors). Finally, one should also plot the residuals against any other variables of interest, such as environmental factors that may have been recorded. If unusual behavior is noted in any of these steps, additional measures should be taken to stabilize the variance (e.g., by considering transformations of the variables or by reexamining the experiment for other factors that may need to be either blocked against or otherwise included in the experiment).

Two-Way Analysis of Variance

Suppose we are seeking to determine if various treatments are important in determining an output effect, but we must conduct our experiment in such a way that another variable (which may also impact the output) must also vary. For example, suppose we want to study two treatments A and B but must conduct the experiments on five different process tools (tools 1–5). In this case, we must carefully design the experiment to *block* against the influence of the process tool factor. We now have an assumed model:

$$y_{ti} = \mu + \tau_t + \beta_i + \epsilon_{ti} \tag{45}$$

where β_i are the block effects. The total sum of squares SS can now be decomposed as

$$SS = SS_A + SS_B + SS_T + SS_R \tag{46}$$

with degrees of freedom

$$bk = 1 + (b - 1) + (k - 1) + (b - 1)(k - 1) \tag{47}$$

where b is the number of blocking groups, k is the number of treatment groups, and $SS_B = k \sum_{i=1}^{b} (\bar{y}_i - \bar{y})^2$. As before, if the blocks or treatments do in fact include any mean shifts, then the corresponding mean sum of squares (estimates of the corresponding variances) will again be inflated beyond the population variance (assuming the number of samples at each treatment is equal):

$$\begin{aligned} s_B^2 \text{ estimates } & \left(\sigma^2 + k \sum_{i=1}^{b} \frac{\beta_i^2}{(b-1)} \right) \\ s_T^2 \text{ estimates } & \left(\sigma^2 + \sum_{t=1}^{k} \frac{n_t \tau_t^2}{(k-1)} \right) \end{aligned} \tag{48}$$

So again, we can now test the significance of these potentially "inflated" variances against the pooled estimate of the variance s_R^2 with the appropriate F test as summarized in Table 3.

Two-Way Factorial Designs

While the above is expressed with the terminology of the second factor being considered a "blocking" factor, precisely the same analysis pertains if two factors are simultaneously considered in the experiment. In this case, the blocking groups are the different levels of one factor, and the treatment groups are the levels of the other factor. The assumed analysis of variance above is with the simple additive model (that is, assuming that there are no interactions between the blocks and treatments, or between the two factors). In the blocked experiment, the intent of the blocking factor was to isolate a known (or suspected) source of "contamination" in the data, so that the precision of the experiment can be improved.

We can remove two of these assumptions in our experiment if we so desire. First, we can treat both variables as equally legitimate factors whose effects we wish to identify or explore. Second, we can explicitly design the experiment and perform the analysis to investigate *interactions* between the two factors. In this case, the model becomes

$$Y_{tij} = \mu_{ti} + \epsilon_{tij} \tag{49}$$

Table 3. Structure of the Analysis of Variance Table, for the Case of a Treatment with a Blocking Factor

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio	Pr(F)
Average (correction factor)	$SS_A = nk\bar{y}^2$	1			
Between blocks	$SS_B = k\sum_{i=1}^{b}(\bar{y}_i - \bar{y})^2$	$v_B = b - 1$	s_B^2	s_B^2/s_R^2	p_B
Between treatments	$SS_T = b\sum_{t=1}^{k}(\bar{y}_t - \bar{y})^2$	$v_T = k - 1$	s_T^2	s_T^2/s_R^2	p_T
Residuals	SS_R	$v_R = (b-1)(k-1)$	s_R^2		
Total	SS	$v = N = bk$	s_D^2		

where μ_{ti} is the effect that depends on both factors simultaneously. The output can also be expressed as

$$\begin{aligned}\mu_{ti} &= \mu + \tau_t + \beta_i + \omega_{ti} \\ &= \overline{y} + (\overline{y_t} - \overline{y}) + (\overline{y_i} - \overline{y}) + (\overline{y_{ti}} - \overline{y_t} - \overline{y_i} + \overline{y})\end{aligned} \quad (50)$$

where μ is the overall grand mean, τ_t and β_i are the main effects, and ω_{ti} are the interaction effects. In this case, the subscripts are

$t = 1, 2, \ldots, k$
where k is the number of levels of first factor
$i = 1, 2, \ldots, b$
where b is the number of levels of second factor
$j = 1, 2, \ldots, m$
where m is the number of replicates at the t, i factor levels

The resulting ANOVA table will be familiar; the one key addition is explicit consideration of the interaction sum of squares and mean square. The variance captured in this component can be compared to the within-group variance as before, and be used as a measure of significance for the interactions, as shown in Table 4.

Another metric often used to assess "goodness of fit" of a model is the R^2. The fundamental question answered by R^2 is how much better does the model do than simply using the grand average.

$$R^2 = \frac{SS_M}{SS_D} = \frac{(SS_T + SS_B + SS_I)}{SS_D} \quad (51)$$

where an R^2 value near zero indicates that most of the variance is explained by residuals ($SS_R = SS_D - SS_M$) rather than by the model terms (SS_M), while an R^2 value near 1 indicates that the model sum of square terms capture nearly all of the observed variation in the data. It is clear that a more sophisticated model with additional model terms will increase SS_M, and thus an "apparent" improvement in explanatory power may result from adding model terms. An alternative metric is the adjusted R^2, where a penalty is added for the use of degrees of freedom in the model:

$$\begin{aligned}\text{Adjusted}\, R^2 &= 1 - \frac{SS_R/v_R}{SS_D/v_D} \\ &= 1 - \frac{s_R^2}{s_D^2} = 1 - \frac{\text{Mean square of residual}}{\text{Mean square of total}}\end{aligned} \quad (52)$$

which is more easily interpreted as the fraction of the variance that is *not* explained by the residuals (s_R^2/s_D^2). In important issue, however, is that variance may appear to be explained when the model in fact does not "fit" the population. One should formally test for lack of fit (as described in the regression modeling section to follow) before reporting R^2, since the R^2 is only a meaningful measure if there is no lack of fit.

Several mnemonics within the factorial design of experiments methodology facilitate the rapid or manual estimation of main effects, as well as interaction effects (8). Elements of the methodology include (a) assignment of "high" (+) and "low" (−) values for the variables, (b) coding of the experimental combinations in terms of these high and low levels, (c)

Table 4. Structure of the Analysis of Variance Table, for Two-Factor Case with Interaction Between Factors

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio	Pr(F)
Between levels of factor 1	$SS_T = b\sum_{t=1}^{k}(\bar{y}_t - \bar{y})^2$	$v_T = k - 1$	s_T^2	s_T^2/s_E^2	p_T
Between levels of factor 2	$SS_B = k\sum_{i=1}^{b}(\bar{y}_i - \bar{y})^2$	$v_B = b - 1$	s_B^2	s_B^2/s_E^2	p_B
Interaction	SS_I	$v_I = (k-1)(b-1)$	s_I^2	s_I^2/s_E^2	p_I
Within groups (error)	SS_E	$v_E = bk(m-1)$	s_E^2		
Total (mean corrected)	SS_D	$v = bkm - 1$	s_D^2		

Table 5. Full Factorial 2^3 Experimental Design (with Coded Factor Levels)

Experiment Condition Number	Factor A	Factor B	Factor C	Measured Result
1	−	−	−	
2	−	−	+	
3	−	+	−	
4	−	+	+	
5	+	−	−	
6	+	−	+	
7	+	+	−	
8	+	+	+	

randomization of the experimental runs (as always!), and (d) estimation of effects by attention to the experimental design table combinations.

For example, one might perform a 2^3 full factorial design [where the superscript indicates the number of factors (three in this case), and the base indicates the number of levels for each factor (two in this case)], where the factors are labeled A, B, and C. The unique eight combinations of these factors can be summarized as in Table 5, with the resulting measured results added during the course of the experiment.

The main effect of a factor can be estimated by taking the difference between the average of the + level for that factor and the average of the − levels for that factor—for example, $\text{Effect}_A = \bar{y}_{A+} - \bar{y}_{A-}$ and similarly for the other main effects. Two-level interactions can be found in an analogous fashion; $\text{Interaction}_{AB} = \frac{1}{2}(\bar{y}_{AB+} - \bar{y}_{AB-})$, where one takes the difference between one factor averages at the high and low values of the second factor. Simple methods are also available in the full factorial case for estimation of factor effect sampling variances, when replicate runs have been performed. In the simple case above where only a single run is performed at each experimental replicate, there are no simple estimates of the underlying process or measurement variance, and so assessment of significance is not possible. If, however, one performs m_i replicates at the ith experimental condition, one can pool the individual estimates of variance s_i^2 at each of the experimental conditions to gain an overall variance estimate (8):

$$s^2 = \frac{v_1 s_1^2 + v_2 s_2^2 + \cdots + v_g s_g^2}{v_1 + v_2 + \cdots + v_g} \tag{53}$$

where $v_i = m_i - 1$ are the degrees of freedom at condition i and g is the total number of experimental conditions examined.

The sampling variance for an effect estimate can then be calculated; in our previous example we might perform two runs at each of the eight experimental points, so that $v_i = 1$ and

$$\text{Var}\{\text{Effect}_A\} = \text{Var}\{\bar{y}_{A+}\} + \text{Var}\{\bar{y}_{A-}\} = \frac{s^2}{8} + \frac{s^2}{8} = \frac{s^2}{4} \tag{54}$$

These methods can be helpful for rapid estimation of experimental results and for building intuition about contrasts in experimental designs; however, statistical packages provide the added benefit of assisting not only in quantifying factor effects and interactions, but also in examination of the significance of these effects and creation of confidence intervals on estimation of factor effects and interactions.

Nested Variance Structures

While blocking factors may seem somewhat esoteric or indicative of an imprecise experiment, it is important to realize that blocking factors do in fact arise extremely frequently in semiconductor manufacturing. Indeed, most experiments or sets of data will actually be taken under situations where great care must be taken in the analysis of variance. In effect, such blocking factors arise due to *nested variance structures* in typical semiconductor manufacturing which restrict the full randomization of an experimental design (9). For example, if one samples die from multiple wafers, it must be recognized that those die reside *within* different wafers; thus the wafer is itself a blocking factor and must be accounted for.

For example, consider multiple measurements of oxide film thickness across a wafer following oxide deposition. One might expect that measurements from the same wafer would be more similar to each other than those across multiple wafers; this would correspond to the case where the within-wafer uniformity is better than the wafer to wafer uniformity. On the other hand, one might also find that the measurements from the corresponding sites on each wafer (e.g., near the lower left edge of the wafer) are more similar to each other than are the different sites across the same wafer; this would correspond to the case where wafer-to-wafer repeatability is very good, but within-wafer uniformity may be poor. In order to model the important aspects of the process and take the correct improvement actions, it will be important to be able to distinguish between such cases and clearly identify where the components of variation are coming from.

In this section, we consider the situation where we believe that multiple site measurements "within" the wafer can be treated as independent and identical samples. This is almost never the case in reality, and the values of "within wafer" variance that result are not true measures of wafer variance, but rather only of the variation across those (typically fixed or preprogrammed) sites measured. In our analysis, we are most concerned that the wafer is acting as a blocking factor, as shown in Fig. 6. That is, we first consider the case where we find that the five measurements we take on the wafer are relatively similar, but the wafer-to-wafer average of these values varies dramatically.

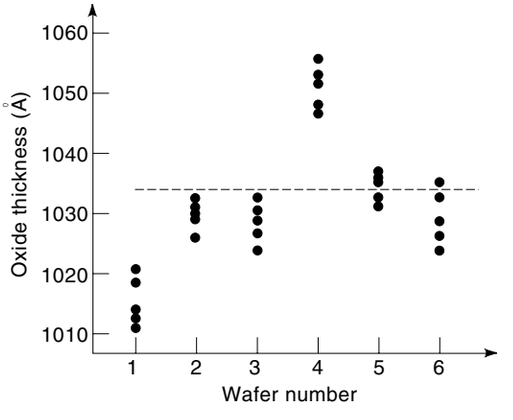

Figure 6. Nested variance structure. Oxide thickness variation consists of both within-wafer and wafer-to-wafer components.

This two-level variance structure can be described as (9)

$$\begin{aligned} y_{ij} &= \mu + W_i + M_{j(i)} \\ W_i &\sim \mathrm{N}(0, \sigma_W^2) \qquad \text{for} \qquad i = 1, \ldots, n_W \\ M_{j(i)} &\sim \mathrm{N}(0, \sigma_M^2) \qquad \text{for} \qquad j = 1, \ldots, n_M \end{aligned} \tag{55}$$

where n_M is the number of measurements taken on each of n_W wafers, and W_i and $M_{j(i)}$ are independent normal random variables drawn from the distributions of wafer-to-wafer variations and measurements taken within the i^{th} wafer, respectively. In this case, the total variation in oxide thickness is composed of both within-wafer variance σ_M^2 and wafer-to-wafer variance σ_W^2:

$$\sigma_T^2 = \sigma_W^2 + \sigma_M^2 \tag{56}$$

Note, however, that in many cases (e.g., for control charting), one is not interested in the total range of all individual measurements, but rather, one may desire to understand how the set of *wafer means* itself varies. That is, one is seeking to estimate

$$\sigma_{\overline{W}}^2 = \sigma_W^2 + \frac{\sigma_M^2}{n_M} \tag{57}$$

where $\overline{W}$ indicates averages over the measurements within any one wafer.

Substantial care must be taken in estimating these variances; in particular, one can directly estimate the measurement variance σ_M^2 and the wafer average variance $\sigma_{\overline{W}}$, but must infer the wafer-level variance σ_W^2 using

$$\sigma_W^2 = \sigma_{\overline{W}}^2 - \frac{\sigma_M^2}{n_M} \tag{58}$$

The within-wafer variance is most clearly understood as the average over the available n_W wafers of the variance s_i^2 within each of those i wafers:

$$s_M^2 = \frac{1}{n_W} \sum_{i=1}^{n_W} s_i^2 = \frac{1}{n_W} \sum_{i=1}^{n_W} \left(\sum_{j=1}^{n_M} \frac{(Y_{ij} - Y_{i\cdot})^2}{n_M - 1} \right) \tag{59}$$

where $Y_{i\cdot}$ indicates an average over the j index (i.e., a within-wafer average):

$$Y_{i\cdot} = \frac{1}{n_M} \sum_{j=1}^{n_M} Y_{ij} \tag{60}$$

The overall variance in wafer averages can be estimated simply as

$$s_{\overline{W}}^2 = \frac{1}{n_W - 1} \sum_{i=1}^{n_W} (Y_{i\cdot} - Y_{\cdot\cdot})^2 \tag{61}$$

where $Y_{\cdot\cdot}$ is the grand mean over all measurements:

$$Y_{\cdot\cdot} = \frac{1}{n_W n_M} \sum_{i=1}^{n_W} \sum_{j=1}^{n_W} Y_{ij} \tag{62}$$

Thus, the wafer-level variance can finally be estimated as

$$s_W^2 = s_{\overline{W}}^2 - \frac{s_M^2}{n_M} \tag{63}$$

The same approach can be used for more deeply nested variance structures (9,10). For example, a common structure occurring in semiconductor manufacturing is measurements within wafers, and wafers within lots. Confidence limits can also be established for these estimates of variance (11). The computation of such estimates becomes substantially complicated, however (especially if the data are unbalanced and have different numbers of measurements per samples at each nested level), and statistical software packages are the best option.

Several assumptions are made in the analysis of variance components for the nested structures above. Perhaps the most important is an assumption of random sampling within each level of nesting. For example, we assume that each measurement (within each wafer) is IID and a random sample from within the wafer. If the same measurement points are taken on each wafer, however, one is not in fact truly estimating the within-wafer variation, but rather the fixed-effect variance between these measurement points. For example, it is common practice to use a spatially consistent five-point (or 21-point or 49-point) sampling scheme when making measurements within a wafer. An option which adds complexity but also adds precision is to model each of these sites separately (e.g., maintain left, right, top, bottom, and center points) and consider how these compare with other points within the wafer, as well as from wafer-to-wafer. Great care is required in such site modeling approaches, however, because one must account for the respective variances at multiple levels appropriately in order to avoid biased estimates (12–14).

Experimental designs that include nested variance sampling plans are also sometimes referred to as split-plot designs, in which a factorial design in fact has restrictions on randomization (7,15). Among the most common restrictions are those due to spatial factors, and spatial modeling likewise requires great care (16). Other constraints of the "real" world, such as hardware factors, may make complete randomization infeasible due to the time and cost of installing/removing hardware (e.g., in studying alternative gas distribution plates in a plasma reactor). Methods exist for handling such constraints (split-plot analysis), but the analysis cannot be done if the experimental sampling plan does not follow an appropriate split-plot design. Using split-plot analyses, however, we can resolve components of variation due (in the case of sites within wafers) into residual, site, wafer, and wafer–site interactions, as well as the effects of the treatment under consideration. For these reasons, it can be expected that nested variance structures or split-plot designs will receive even greater future attention and application in semiconductor manufacturing.

Progression of Experimental Designs

It is worth considering when and how various experimental design approaches might best be used. When confronted with a new problem which lacks thorough preexisting knowledge, the first step should be screening experiments which seek to identify what the important variables are. At this stage, only crude predictions of experimental effects as discussed above are needed, but often a large number of candidate factors (often six or more) are of potential interest. By sacrificing accuracy and certainty in interpretation of the results (primarily by allowing interactions to confound with other interactions or even with main effects), one can often gain a great deal of

initial knowledge with reasonable cost. In these cases, fractional factorial, Plackett–Burmann, and other designs may be used.

Once a smaller number of effects have been identified, full factorial or fractional factorial designs are often utilized, together with simplified linear model construction and analysis of variance. In such cases, the sampling plan must again be carefully considered in order to ensure that sufficient data are taken to draw valid conclusions. It is often possible to test for model lack of fit, which may indicate that more thorough experiments are needed or that additional experimental design points should be added to the existing experimental data (e.g., to complete half fractions). The third phase is then undertaken, which involves experimental design with small numbers of factors (e.g., two to six) to support linear effects, interactions, and second-order (quadratic) model terms. These regression models will be considered in the next section.

A variety of sophisticated experimental design methods are available and applicable to particular problems. In addition to factorial and "optimal" design methods (8,10), robust design approaches (as popularized by Taguchi) are helpful, particularly when the goal is to aid in the optimization of the process (17,18).

RESPONSE SURFACE METHODS

In the previous section, we considered the analysis of variance, first in the case of single treatments and then in the case when blocking factors must also be considered. These were generalized to consideration of two factor experiments, where the interaction between these factors can also be considered. In all of this discussion, the factor *levels* were treated as either nominal or continuous parameters. An important issue is the estimation of the *effect* of a particular factor, and determination of the *significance* of any observed effect. Such results are often pictured graphically in a succinct fashion, as illustrated in Fig. 7 for a two-factor experiment, where two levels for each of factor A and factor B are examined.

In the full factorial case, interactions can also be explored, and the effects plots modified to show the effect of each factor on the output parameter of concern (yield in this case), but at different levels of the other factor. Various cases may result; as shown in Fig. 8 no interaction may be observed, or a synergistic (or anti-synergistic) interaction may be present. These analyses are applicable, for example, when the factor levels are discrete or nominal decisions to be made; perhaps level (+) for factor A is to perform a clean step while (−) is to omit

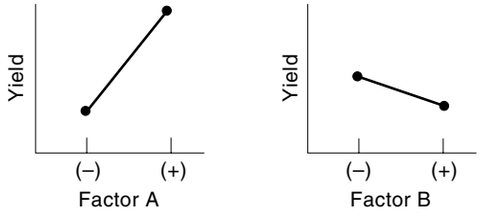

Figure 7. Main effects plot for two-factor, two-level experiment. The influence of Factor A on yield is larger than that of Factor B. Analysis of variance is required in order to determine if the observed results are significant (and not the result of chance variation).

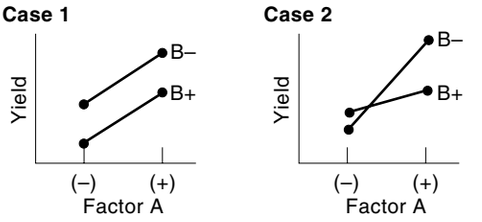

Figure 8. Interaction plot. In case 1, no clear interaction is observed: Factor B does not appear to change the effect of factor A on the output. Rather, the effects from factor A and factor B appear to be additive. In case 2, the level of factor B does influence strongly the response of yield to the high level of factor A.

the step, and level (+) for factor B is to use one chemical in the step while level (−) is to use a different chemical.

If the factors can take on continuous values, the above analysis is still applicable. However, in these cases, it is often more convenient and useful to consider or seek to model (within the range of factor levels considered) an entire *response surface* for the parameter of interest. Specifically, we wish to move from our factorial experiments with an assumed model of the form

$$\hat{y}_{ij} = \hat{\mu} + A_i + B_j + \epsilon_{ij} \tag{64}$$

where we can only predict results at discrete prescribed i, j levels of factors A and B, toward a new model of the process of the form

$$\begin{aligned} \hat{y} &= \hat{\mu} + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \epsilon \\ x^{(j)} &\in [x^{(j)}_{\min}, x^{(j)}_{\max}] \in \Re \end{aligned} \tag{65}$$

where each $x^{(j)}$ is a particular factor of interest.

In this section, we briefly summarize the methods for estimation of the factor response coefficients β_j, as well as for analysis of the significance of such effects based on experimental design data. We begin with a simple one-parameter model, and we build complexity and capability from there.

Single-Variable Least-Squares Regression

The standard approach used here is *least-squares regression* to estimate the coefficients in regression models. In the simple one-parameter case considered here, our actual response is modeled as

$$y_i = \beta x_i + \epsilon_i \tag{66}$$

where y_i indicates the ith measurement, taken at a value x_i for the explanatory variable x. The estimate for the output is thus simply $\hat{y}_i = \hat{\beta} x_i = b x_i$ where we expect some residual error ϵ_i. Least-squares regression finds the best fit of our model to the data, where "best" is that b which minimizes the sum of squared errors between the prediction and observed n data values:

$$SS_{\min} = SS_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{67}$$

It can easily be shown that for linear problems, a direct solution for b which gives $SS_{\min}$ is possible and will occur when

the vector of residuals ϵ_i is normal to the vector of x_i values:

$$b = \frac{\sum xy}{\sum x^2} \tag{68}$$

An important issue is the estimate of the experimental error. If we assume that the model structure is adequate, we can form an estimate s^2 of σ^2 simply as

$$S^2 = \frac{SS_R}{n-1} \tag{69}$$

One may also be interested in the precision of the estimate b—that is, the variance in b:

$$\text{Var}\{b\} = \frac{s^2}{\sum_i x_i^2} \tag{70}$$

assuming that the residuals are independent and identically normally distributed with mean zero. More commonly, one refers to the standard error s.e.$(b) = \sqrt{\text{Var}\{b\}}$ and writes $b \pm$ s.e.(b). In a similar fashion, a confidence interval for our estimate of β can be defined by noting that the standardized value for b should be t-distributed:

$$t = \frac{b - \beta'}{\text{s.e.}(b)} \tag{71}$$

where β' is the true value for β, so that

$$\beta = b \pm [t_{\alpha/2} \cdot \text{s.e.}(b)] \tag{72}$$

Regression results should also be framed in an analysis of variance framework. In the simple one factor case, a simple ANOVA table might be as shown in Table 6. In this case, SS_M is the sum of squared values of the estimates, and s_M^2 is an estimate of the variance "explained" by the model, where our model is purely linear (no intercept term) as given in Eq. 66. In order to test significance, we must compare the ratio of this value to the residual variance s_R^2 using the appropriate F test. In the case of a single variable, we note that the F test degenerates into the t test: $F_{1,n} = t_n^2$, and a t test can be used to evaluate the significance of the model coefficient.

In the analysis above, we have assumed that the values for x_i have been selected at random and are thus unlikely to be replicated. In many cases, it may be possible to repeat the experiment at particular values, and doing so gives us the opportunity to decompose the residual error into two contributions. The residual sum of squares SS_R can be broken into a component SS_L due to "lack of fit" and a component SS_E due to "pure error" or "replication error":

$$SS_R = SS_L + SS_E \tag{73}$$

This enables a further test for "lack of fit" in our model by comparing the ratio of the corresponding variances; that is, we compare s_L^2/s_E^2 with $F_{\upsilon_L,\upsilon_E}$, where $\upsilon_E = m - 1$ is the degrees of freedom corresponding to the pure error for m replicates, and $\upsilon_L = \upsilon_R - \upsilon_E$ is the degrees of freedom corresponding to the lack-of-fit variance estimate. It is highly recommended that at least some points (if not the entire experiment) be replicated, so that the lack of fit and pure error can be assessed; otherwise, some question will remain as to the validity of the model.

These tests can also be summarized as part of the ANOVA table, as shown in Table 7. In this case, we assume a true response of the form $y = \beta_0 + \beta_1 x$ which we estimate or fit with a two-parameter model $y = b_0 + b_1 x + \epsilon$ or $\hat{y} = b_0 + b_1 x$ to also capture the mean (or intercept β). We assume that we have made n total measurements, of which m are replicates. In this table, one should first check for lack of fit. If no evidence of lack of fit exists, then there is no reason to reject the assumed model structure, and one can assess the significance of the overall model or individual model coefficients. Note that the test of significance for s_M^2 compared to s_E^2, and the probability p_M of observing the corresponding ratio, is then equivalent to testing if the adjusted $R^2 = 0$. If evidence of lack of fit does indeed exist, however, then one must seek alternative model forms, either through transformations of the data or by seeking a higher-order (e.g., polynomial) model structure. As always, one should also examine the residuals.

Just as one can assess significance and formulate confidence intervals for the single model coefficient case, so too can one find interval estimates for the model coefficients. Typically, statistical packages can be utilized to assist in the formulation of such estimates, but care must be taken to understand the above framework in order to correctly interpret the output of such packages.

Table 6. Structure of the Analysis of Variance Table, for Single-Factor Response Surface Regression[a]

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio	Pr(F)
Model	SS_M	$\upsilon_M = 1$ (number of model coefficients)	s_M^2	s_M^2/s_R^2	α
Residual	SS_R	$\upsilon_R = n - \upsilon_M$	s_R^2		
Total	SS	$\upsilon = n$	s_T^2		

[a] The degrees of freedom in the model are shown for the case when only one model coefficient is used (strictly linear response).

Response Surface Modeling—Experimental Designs

The first part of this section focused on regression modeling and analysis for single-factor experiments, concluding with polynomial models of the response. In many cases, one is interested in modeling the response as a function of multiple factors, with linear or quadratic models. Here we briefly review aspects of popular experimental designs, and we interpret the results of analysis of variance in this multiple factor context. While a number of experimental designs with different properties (and indeed an entire arena of design methods that are "optimal" in various senses) exist, two of the more popular designs will be summarized here. The central composite design, as pictured in the two factor case in Fig. 9, is especially useful as a complement or addition to existing factorial design data. In this case, the addition of center point and axial points completes the central composite, and it supports quadratic modeling of the responses (if found to be necessary).

A second popular option is the Box–Bhenken design, as illustrated for the two factor case in Fig. 9. In this case, the center point is complemented by experimental points at the

Table 7. Structure of the Analysis of Variance Table for a Single-Factor Response Surface Regression, in the Case of Replication of Experimental Design Points

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio	Pr(F)
Model	SS_M	$\upsilon_M = 2$	$s_M^2 = \frac{SS_M}{\upsilon_M}$	s_M^2/s_E^2	p_M
b_0	SS_0	$\upsilon_0 = 1$	$s_0^2 = SS_0$	s_0^2/s_E^2	p_0
b_1	SS_1	$\upsilon_1 = 1$	$s_1^2 = SS_1$	s_1^2/s_E^2	p_1
Residual	SS_R	$\upsilon_R = n - \upsilon_M$	s_R^2		
lack-of-fit	SS_L	$\upsilon_L = \upsilon_R - \upsilon_E$	s_L^2	s_L^2/s_E^2	Pr (lack of fit)
pure error	SS_E	$\upsilon_E = m$	s_E^2		
Total	SS	$\upsilon = n$	s^2		

midpoint of each segment of the n-dimensional "bounding box" around that center point. Alternatively, this can be viewed as the center point augmented by the aggregate of n full factorial designs in $n - 1$ experimental factors while holding each remaining factor at its center value. The Box–Bhenken design is generally used when the expense or time of the experiment is influenced by the number of levels, because the Box–Bhenken only requires three different levels for each factor, while the central composite design requires five. In both of these designs, it should be again emphasized that replicates at one or more experimental points (typically the center point) are highly recommended so that lack of fit can be assessed, and so a measure of pure or experimental error can be established.

The response surface models for each case are found using a least-squares fit to a specified model structure (typically quadratic or polynomial) as previously discussed. An analysis-of-variance examination is required to check for model lack of fit, examine factor and model coefficient significance, and establish confidence intervals on model coefficients. Careful examination of residuals is crucial to ensure the validity of the modeling assumptions—namely, that the residuals are IID and normally distributed.

In the case of multiple model coefficients, one often desires the most parsimonious or simple model possible. Analysis of variance can indicate those coefficients which appear to be insignificant. In step-wise regression, model coefficients are dropped or added one at a time, and the reduction (or improvement) in the model is evaluated until some stopping criteria are met.

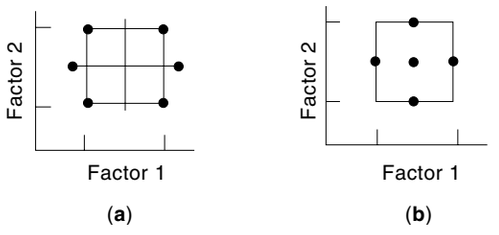

Figure 9. Experimental designs often used in response surface modeling. (a) The factorial design (augmented with center points) can be extended into (b) the central composite design by adding axial design points. (c) The Box–Bhenken design features design points at midpoints of bounding box segments around the center point.

CATEGORICAL MODELING

Sometimes the measurement of the result is discrete. One example is an engineer's evaluation of a photograph, usually from a scanning electron microscope (SEM). The photograph might be a top-down image of a metal or polysilicon patterned line. The engineer decides whether the line is too rough, a little rough, or smooth. While many of these evaluations are being replaced with automatic defect detection and classification tools which provide continuous numbers, especially in production, early development still relies heavily on manual inspection and evaluation. The engineer would like to perform an experiment whereby he or she can determine what optimal value of the bake temperature will result in the best chance of smooth lines. In addition, he or she would like to predict how often the process will produce rough lines. Another example is the profile or sidewall slope of a line. Except perhaps Atomic Force Microscopy (AFM), no measurement method exists to obtain a line profile quickly and easily. However, a cross-sectional SEM can be used to roughly estimate the profile—that is, to note if the slope in degrees is >88, 85–88, or <88. No definite constant scale exists, but a relative scale does (i.e., >88 is bigger than 85–88, which is bigger than <88). The engineer would like to know if performing a clean (and the length of the clean step) will result in a sharper profile (>88). Categorical methods are statistical methods aimed at use for these questions (19). While the mathematics are too complicated to introduce here, many statistical packages (such as SAS, JMP, and Statgraphics) provide these methods and can be applied to practical problems. The methods can be shown to be similar in nature to fuzzy logic (20).

SUMMARY

In this article, we have focused on the fundamental issues in modeling important statistical elements of semiconductor manufacturing. In many cases, we have only begun to touch on the issues of statistical distribution modeling, hypothesis testing, experimental design and analysis of variance, and response surface modeling. The intent here has been to assist in the proper interpretation of results that are now readily available by way of statistical software packages; further consultation with the statistical modeling literature and statisticians is highly recommended for those seeking to get the most value out of experimental resources and data. An excellent

source for further reading are the case studies of statistical methods applied to semiconductor manufacturing contained in Ref. 21.

BIBLIOGRAPHY

1. C. J. Spanos, Statistical process control in semiconductor manufacturing, *Proc. IEEE,* **80**: 819–830, 1992.
2. J. B. Keats and D. C. Montgomery (eds.), *Statistical Applications in Process Control,* New York: Dekker, 1996.
3. A. Madansky, *Prescriptions for Working Statisticians,* Berlin: Springer-Verlag, 1988.
4. D. M. H. Walker, *Yield Simulation for Integrated Circuits,* Norwell, MA: Kluwer, 1987.
5. D. C. Montgomery, *Introduction to Statistical Quality Control,* New York: Wiley, 1985.
6. R. G. Miller, Jr., *Beyond ANOVA—Basics of Applied Statistics,* New York: Chapman & Hall, 1997.
7. G. A. Milliken and D. E. Johnson, *Analysis of Messy Data,* Vol. I: Designed Experiments, New York: Chapman & Hall, 1992.
8. G. E. P. Box, W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters,* New York: Wiley, 1978.
9. D. Drain, *Statistical Methods for Industrial Process Control,* New York: Chapman & Hall, 1997.
10. D. Drain, *Handbook of Experimental Methods for Process Improvement,* New York: Chapman & Hall, 1997.
11. R. K. Burdick and F. A. Graybill, *Confidence Intervals on Variance Components,* New York: Dekker, 1992.
12. T. H. Smith et al., Bias and variance in single and multiple response surface modeling, *3rd Int. Workshop Stat. Metrol.,* Honolulu, HI, 1998.
13. R. Guo and E. Sachs, Modeling, optimization, and control of spatial uniformity in manufacturing processes, *IEEE Trans. Semicond. Manuf.,* **6**: 41–57, 1993.
14. P. K. Mozumder and L. M. Lowenstein, Method for semiconductor process optimization using functional representations of spatial variations and selectivity, *IEEE Trans. Comp. Hybrids Manuf. Tech.,* **15**: 311–316, 1992.
15. R. L. Mason, R. F. Gunst, and J. L. Hess, *Statistical Design and Analysis of Experiments with Applications to Engineering and Science,* New York: Wiley, 1989.
16. B. D. Ripley, *Spatial Statistics,* New York: Wiley, 1981.
17. P. J. Ross, *Taguchi Techniques for Quality Engineering,* 2nd ed., New York: McGraw-Hill, 1996.
18. M. S. Phadke, *Quality Engineering Using Robust Design,* Englewood Cliffs, NJ: Prentice-Hall, 1989.
19. A. Agresti, *Categorical Data Analysis,* New York: Wiley, 1990.
20. C. Spanos and R. Chen, Using qualitative observations for process tuning and control, *IEEE Trans. Semicond. Manuf.,* **10**: 307–316, 1997.
21. V. Czitrom and P. D. Spagon (eds.), *Statistical Case Studies for Industrial Process Improvement,* Philadelphia: ASA-SIAM, 1997.

DUANE S. BONING
Massachusetts Institute of Technology

JERRY STEFANI
STEPHANIE W. BUTLER
Texas Instruments