5

Applied Statistical Methods and the Chemical Industry

Stephen Vardeman* and Robert Kasprzyk**

INTRODUCTION

The discipline of statistics is the study of effective methods of data collection, data summarization, and (data based, quantitative) inference making in a framework that explicitly recognizes the reality of nonnegligible variation in real-world processes and measurements.

The ultimate goal of the field is to provide tools for extracting the maximum amount of useful information about a noisy physical process from a given investment of data collection and analysis resources. It is clear that such a goal is relevant to the practice of industrial chemistry. The primary purposes of this chapter are to indicate in concrete terms the nature of some existing methods of applied statistics that are particularly appropriate to industrial chemistry, and to provide an entry into the statistical literature for those readers who find in the discussion here reasons to believe that statistical tools can help them be effective in their work.

This chapter will begin with some simple ideas of modern descriptive statistics, including numerical and graphical data summarization tools, and the notions of fitting equations to data and using theoretical distributions. Next, some tools for routine industrial process monitoring and capability assessment, concentrating primarily on the notion of control charting, will be presented. This will be followed by a more extensive discussion of common statistical data collection strategies and data analysis methods for multifactor experimental situations met in both laboratory and production environments. This section will touch on ideas of partitioning observed variation in a system response to various sources thought to influence the response, factorial and fractional factorial experimental designs, sequential experimental strategy, screening experiments, and response surface fitting and representation. Next come brief discussions of two types of special statistical tools associated specifically with chemical applications, namely, mixture techniques and nonlinear mechanistic

*Iowa State University, Departments of Statistics and of Industrial Engineering and Manufacturing Systems Engineering.
**Dow Chemical Company.

model building. A short exposition of chemical industry implications of relationships between modern business process improvement programs and the discipline of statistics follows. The chapter concludes with a reference section listing sources for further reading.

SIMPLE TOOLS OF DESCRIPTIVE STATISTICS

There are a variety of data summarization or description methods whose purpose is to make evident the main features of a data set. (Their use, of course, may be independent of whether or not the data collection process actually employed was in any sense a "good" one.) To illustrate some of the simplest of these methods, consider the data listed in Table 5.1. These numbers represent aluminum impurity contents (in ppm) of 26 bihourly samples of recycled PET plastic recovered at a Rutgers University recycling pilot plant.

A simple plot of aluminum content against time order, often called a *run chart*, is a natural place to begin looking for any story carried by a data set. Figure 5.1 shows such a plot for the data of Table 5.1, and in this case reveals only one potentially interesting feature of the data. That is, there is perhaps a weak hint of a downward trend in the aluminum contents that might well have been of interest to the original researchers. (If indeed the possible slight decline in aluminum contents is more than "random scatter," knowledge of its physical origin, whether in actual composition of recycled material or in the measurement process, presumably would have been helpful to the effective running of the recycling facility. We will save a discussion of tools for rationally deciding whether there is more than random scatter in a plot like Fig. 5.1 until the next section.)

The run chart is a simple, explicitly dynamic tool of descriptive statistics. In those cases where one decides that there is in fact

TABLE 5.1 Twenty-Six Consecutive Aluminum Contents (ppm)[a]

291,	222,	125,	79,	145,	119,	244,	118,	182,	63,	30,	140,	101
102,	87,	183,	60,	191,	119,	511,	120,	172,	70,	30,	90,	115

[a] Based on data in Albin.[1]

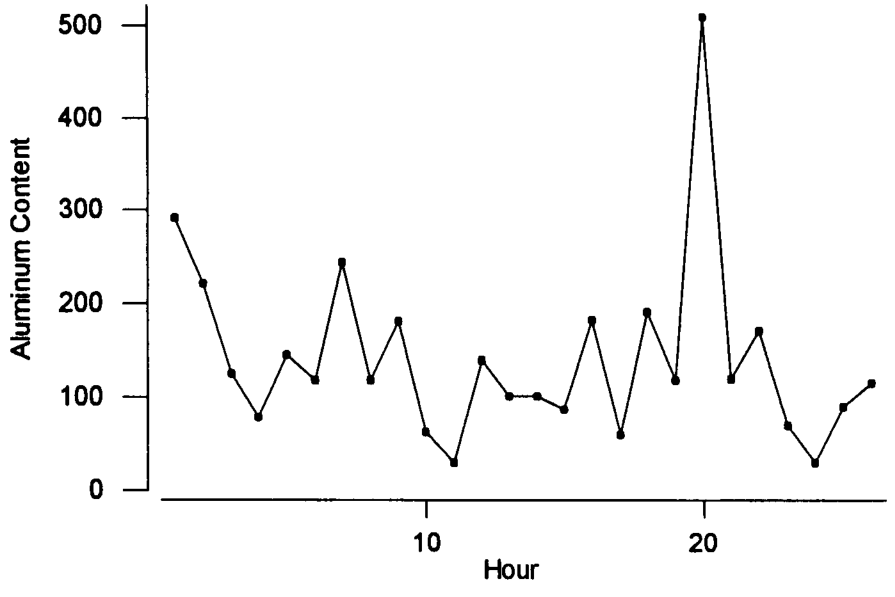

Fig. 5.1. A run chart for 26 consecutive aluminum contents.

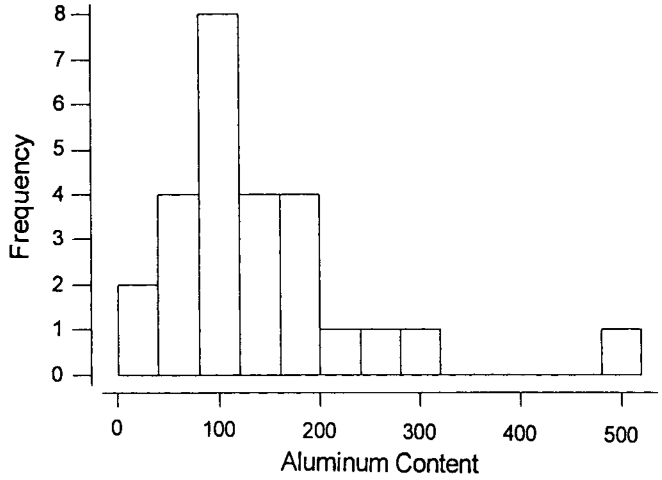

Fig. 5.2. A histogram for 26 aluminum contents.

little information in the time order corresponding to a data set, there are a variety of simple, essentially static, statistical tools that can be used in describing the pattern of variation in a data set. Figures 5.2–5.5 show graphical representations of the data of Table 5.1 in, respectively, *histogram, stem and leaf plot, dot plot*, and *box plot* forms.

The histogram/bar chart idea of Fig. 5.2 is likely familiar to most readers, being readily available, for example, through the use of commercial spreadsheet software. It shows how data are spread out or distributed across the range of values represented, tall bars indicating high frequency or density of data in the interval covered by the base of the bar. Figure 5.2 shows the measured aluminum contents to be somewhat asymmetrically distributed (statistical jargon is that the distribution is "skewed right"), with a "central" value perhaps somewhere in the vicinity of 120 ppm.

Histograms are commonly and effectively used for final data presentation, but as working data analysis tools they suffer from several limitations. In one direction, their appearance is fairly sensitive to the data grouping done to make them, and it is usually not possible to recover from a histogram the exact data values used to produce it, should one wish to try other groupings. In another direction, histograms are somewhat unwieldy, for example, not being particularly suitable to the comparison of, say, 10 or 12 data sets on a single page. The graphical devices of Figs 5.3–5.5 are less common than the histogram, but address some of these shortcomings.

0	30,30
0	60,63,70,79,87,90
1	01,02,15,18,19,19,20,25,40,45
1	72,82,83,91
2	22,44
2	91
3	
3	
4	
4	
5	11

Fig. 5.3. A stem and leaf plot for 26 aluminum contents.

The stem and leaf diagram of Fig. 5.3 and the dot plot of Fig. 5.4 carry shape information about the distribution of aluminum contents in a manner very similar to the histogram of Fig. 5.2. But the stem and leaf and dot diagrams do so without losing the exact identities of the individual data points. The box plot of Fig. 5.5 represents the "middle half" of the data with a box divided at the 50th percentile (or in statistical jargon, the median) of the data, and then uses so-called whiskers to indicate how far the most extreme data points are from the middle half of the data.

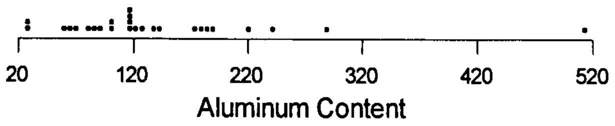

Fig. 5.4. A dot plot for 26 aluminum contents.

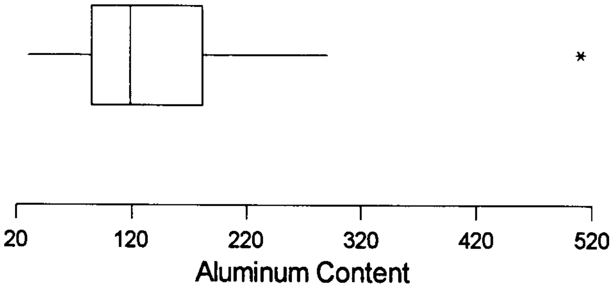

Fig. 5.5. A box plot for aluminum contents.

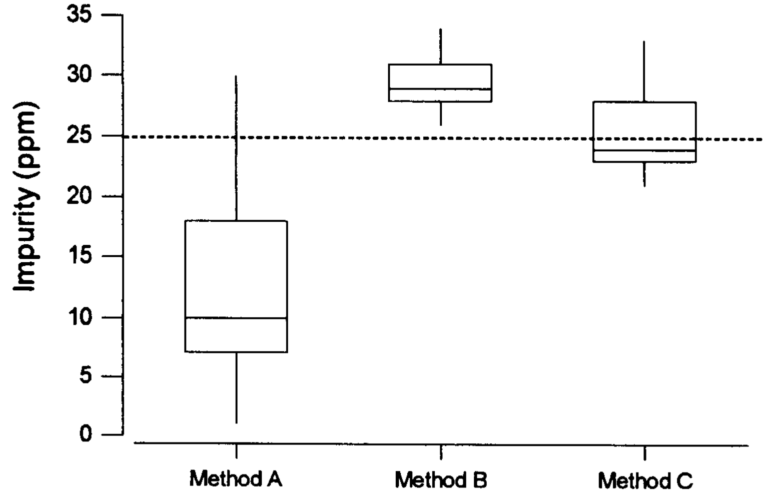

Fig. 5.6. Side-by-side box plots for three laboratory test methods.

Box plots preserve much of the shape information available from the other displays (e.g., portraying lack of symmetry through differing sizes of box "halves" and/or whisker lengths), but do so in a way that is conducive to simultaneous representation and comparison of many data sets on a single graphic, through the placement of box plots side by side. Figure 5.6 illustrates this point with a graphical comparison of three laboratory test methods to a standard.

A total of 90 samples of a stock solution known to contain 25 ppm of an impurity were analyzed by a single lab team using three different test methods (30 of the samples being allocated to each of the three methods), and the box plots in Fig. 5.6 portray the measured impurity levels for the different methods. The figure shows quite effectively that Method A is neither precise nor accurate, Method B is quite precise but not accurate, and Method C is somewhat less precise than B but is accurate. This kind of knowledge can form the basis of an informed choice of method.

Figures 5.2–5.6 give only a hint of the spectrum of tools of statistical graphics that are potentially helpful in data analysis for industrial chemistry. For more details and much additional reading on the subject of modern statistical graphics, the reader is referred to the book by Chambers et al.[2] listed in the references section.

Complementary to graphical data summaries are *numerical summarizations*. For the simple case of data collected under a single set of conditions, the most commonly used measures deal with the location/center of the data set and the variability/spread of the data. The (*arithmetic*) *mean* and the *median* are the most popular measures of location, and the *variance* and its square root, the *standard deviation*, are the most widely used measures of internal variability in a data set.

For n data values $y_1, y_2, \ldots, y_n$ the median is

$$\tilde{y} = \text{the “middle” or } \frac{n+1}{2}\text{th ordered data value} \qquad (5\text{-}1)$$

and the mean is

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (5\text{-}2)$$

The reader is invited to check that upon ordering the $n = 26$ values in Table 5.1, the 13th smallest value is 119 and the 14th smallest value is also 119, so that the only sensible interpretation of (5-1) for the aluminum content data is that

$$\begin{aligned}\tilde{y} &= \text{the 13.5th ordered data value}\\ &= \frac{119+119}{2} = 119\,\text{ppm}\end{aligned}$$

On the other hand, from (5-2) the mean of the aluminum contents is

$$\begin{aligned}\bar{y} &= \frac{1}{26}(291+222+125+\cdots+30+90+115)\\ &\approx 142.7\,\text{ppm}\end{aligned}$$

The median and mean are clearly different measures of location/center. The former is in the middle of the data in the sense that about half of the data are larger and about half are smaller. The latter is a kind of “center of mass,” and for asymmetrical data sets like that of Table 5.1 is usually pulled from the median in the direction of any “skew” present, that is, is pulled in the direction of “extreme” values.

The variance of n data values $y_1, y_2, \ldots, y_n$ is essentially a mean squared deviation of the data points from their mean. In precise terms, the variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (5\text{-}3)$$

and the so-called standard deviation is

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (5\text{-}4)$$

For the example of the aluminum contents, it is elementary to verify that

$$\begin{aligned}s^2 &\approx \frac{1}{26-1}[(291-142.7)^2+(222-142.7)^2\\ &\quad + \cdots + (115-142.7)^2]\\ &\approx 9{,}644\,(\text{ppm})^2\end{aligned}$$

so that

$$s = \sqrt{s^2} \approx 98.2\,\text{ppm}$$

An appropriate interpretation of s is not completely obvious at this point, but it does turn out to measure the spread of a data set, and to be extremely useful in drawing quantitative inferences from data. (In many, but not all, circumstances met in practice, the *range* or largest value in a data set minus the smallest value is on the order of four to six times s.) The variance and standard deviation are time-honored and fundamental quantifications of the variation present in a single group of measurements and, by implication, the data-generating process that produced them.

When data are collected under several different sets of conditions, and those conditions can be expressed in quantitative terms, effective data summarization often takes the form of *fitting an approximate equation* to the data. As the basis of a simple example of this, consider the data in Table 5.2. The variable x, hydrocarbon liquid hourly space velocity, specifies the conditions under which information on the response variable y, a measure of isobutylene conversion, was obtained in a study involving the direct hydration of olefins.

For purposes of economy of expression, and perhaps some cautious interpolation between values of x not included in the original data

TABLE 5.2 Seven Liquid Hourly Space Velocity/Mole % Conversion Data Pairs[a]

Liquid Hourly Space Velocity, x	*Mole % Isobutylene Conversion, y*
1	23.0, 24.5
2	28.0
4	30.9, 32.0, 33.6
6	20.0

[a] Based on a graph in Odioso et al.[3]

set, one might well like to fit a simple equation involving some parameters $\underset{\sim}{b}$, say,

$$y \approx f(x|\underset{\sim}{b}) \qquad (5\text{-}5)$$

to the data of Table 5.2. The simplest possible form for the function $f(x|\underset{\sim}{b})$ that accords with the "up then back down again" nature of the conversion values y in Table 5.2 is the quadratic form

$$f(x|\underset{\sim}{b}) = b_0 + b_1x + b_2x^2 \qquad (5\text{-}6)$$

and a convenient method of fitting such an equation (that is linear in the parameters b) is *the method of least squares.* That is, to fit a parabola through a plot of the seven (x, y) pairs specified in Table 5.2, it is convenient to choose b_0, b_1, and b_2 to minimize the sum of squared differences between the observed conversion values y and the corresponding fitted values y on the parabola. In symbols, the least squares fitting of the approximate relationship specified by (5-5) and (5-6) to the data of Table 5.2 proceeds by minimization of

$$\sum_{i=1}^{7} [y_i - (b_0 + b_1x_i + b_2x_i^2)]^2$$

over choices of the coefficients b. As it turns out, use of standard statistical "regression analysis" software shows that the fitting process for this example produces the approximate relationship

$$y \approx 13.64 + 11.41x - 1.72x^2$$

and Fig. 5.7 shows the fitted (summarizing) parabola sketched on the same set of axes used to plot the seven data points of Table 5.2.

The least squares fitting of approximate functional relationships to data with even multidimensional explanatory variable x typically goes under the (unfortunately obscure) name of *multiple regression* analysis, and is given an introductory treatment in most engineering statistics textbooks, including, for example, the ones by Devore,[4] Vardeman and Jobe,[5] and Vardeman[6] listed in the references. A lucid and rather complete treatment of the subject can also be found in the book by Neter et al.[7]

A final notion that we wish to treat in this section on descriptive statistics is that of representing a distribution of responses and/or the mechanism that produced them (under a single set of physical conditions) by a *theoretical distribution.* That is, there are a number of convenient theoretical distributional shapes, and it is often possible to achieve great economy of expression and thought by seeing in a graphical representation such as Figs 5.2–5.5 the possibility of henceforth describing the phenomenon portrayed via some one of those theoretical distributions. Here we will concentrate on only the most commonly used theoretical distribution, the so-called *Gaussian* or *normal* distribution.

Figure 5.8 is a graph of the function of x

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad (5\text{-}7)$$

where $g(x)$ specifies the archetypical "bell-shaped curve" centered at the number μ, with spread controlled by the number σ (and is in fact usually called the Gaussian probability density with mean μ and standard deviation σ).

Figure 5.8 can be thought of as a kind of idealized histogram. Just as fractional areas enclosed by particular bars of a histogram correspond to fractions of a data set with values in the intervals represented by those bars, areas under the curve specified in (5-7) above particular intervals might be thought of as corresponding to fractions of potential data points having values in those intervals. (It is possible to show that the total area under the curve represented in Fig. 5.8, namely, $\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x$, is 1.) Simple tabular methods presented in every elementary statistics book

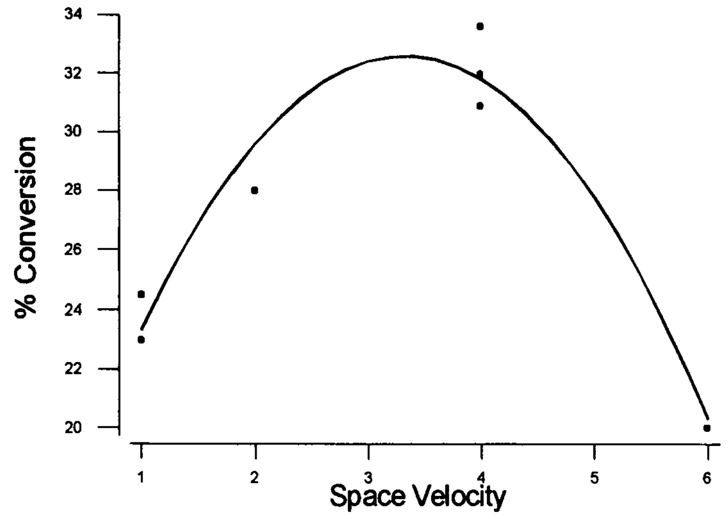

Fig. 5.7. A scatter plot of seven space velocity/mole % conversion data pairs and a fitted parabola.

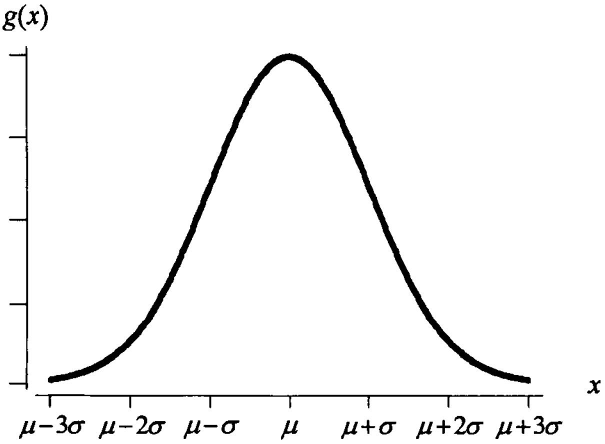

Fig. 5.8. The Gaussian probability density with mean μ and standard deviation σ.

TABLE 5.3 Twenty-Six Logarithms of Aluminum Contents

5.67,	5.40,	4.83,	4.37,	4.98,	4.78,	5.50,	4.77,	5.20,	4.14,	3.40,	4.94,	4.62
4.62,	4.47,	5.21,	4.09,	5.25,	4.78,	6.24,	4.79,	5.15,	4.25,	3.40,	4.50,	4.74

avoid the need to regularly use numerical integration in evaluating such areas. These methods can, for example, be used to show that roughly 68 percent of a Gaussian distribution lies between $\mu - \sigma$ and $\mu + \sigma$, roughly 95 percent lies between $\mu - 2\sigma$ and $\mu + 2\sigma$, and roughly 99.7 percent lies between $\mu - 3\sigma$ and $\mu + 3\sigma$. Part of the convenience provided when one can treat a data-generating process as approximately Gaussian is that, given only a theoretical mean μ and theoretical standard deviation σ, predictions of fractions of future data values likely to fall in intervals of interest are thus easy to obtain.

At this point let us return to the aluminum content data of Table 5.1. The skewed shape that is evident in all of Figs 5.2–5.5 makes a Gaussian distribution inappropriate as a theoretical model for (raw) aluminum content of such PET samples. But as is often the case with right skewed data, considering the *logarithms* of the original measurement creates a scale where a normal distribution is more plausible as a representation of the phenomenon under

3	.40,.40
3	
4	.09,.14,.25,.37,.47
4	.50,.62,.62,.74,.77,.78,.79,.79,.83,.94,.98
5	.15,.20,.21,.25,.40
5	.50,.67
6	.24

Fig. 5.9. A stem and leaf plot for the logarithms of 26 aluminum contents.

study. Thus, Table 5.3 contains the natural logs of the values in Table 5.1, and the corresponding stem and leaf plot in Fig. 5.9 shows the transformed data to be much more symmetrically distributed than the original data. The possibility opened up by this kind of transformation idea is one of using statistical methods based on the normal distribution to reach conclusions about ln$y$ and then simply exponentiating to derive conclusions about the original response y itself. The applicability of statistical methods developed for normal distributions is thereby significantly broadened.

In addition to providing convenient conceptual summarizations of the nature of response distributions, theoretical distributions such as the normal distribution form the mathematical underpinnings of methods of formal quantitative *statistical inference*. It is outside our purposes in this chapter to provide a complete introduction to such methods, but thorough and readable accounts are available in engineering statistics books such as those of Devore[4] and Vardeman and Jobe.[5] Here, we will simply say that, working with a Gaussian description of a response, it is possible to quantify in various ways how much information is carried by data sets of various sizes. For instance, if a normal distribution describes a response variable y, then in a certain well-defined sense, based on $n = 26$ observations producing a mean $\bar{y}$ and a standard deviation s, the interval with end points

$$\bar{y} - 2.060s\sqrt{1+\frac{1}{26}}$$

and (5-8)

$$\bar{y} + 2.060s\sqrt{1+\frac{1}{26}}$$

has a 95 percent chance of predicting the value of an additional observation. For instance, applying formula (5-8) to the log values in Table 5.3, the conclusion is that the interval from 3.45 to 6.10 ln(ppm) has a 95 percent chance of bracketing an additional log aluminum content produced (under the physical conditions of the original study) at the recycling plant. Exponentiating, the corresponding statement about raw aluminum content is that the interval from 31 to 446 ppm has a 95 percent chance of bracketing an additional aluminum content. Methods of statistical inference like that represented in (5-8) are called *prediction interval* methods. The book by Hahn and Meeker[8] provides a thorough discussion of such methods, based not only on the Gaussian distribution but on other theoretical distributional shapes as well.

TOOLS OF ROUTINE INDUSTRIAL PROCESS MONITORING AND CAPABILITY ASSESSMENT

Probably the two most basic generic industrial problems commonly approached using statistical methods are those of (1) monitoring and maintaining the stability/consistency of a process and (2) assessing the capability of a stable process. This section provides a brief introduction to the use of tools of "control" charting in these enterprises.

Working at Bell Labs during the 1920s and 1930s, Walter Shewhart developed the notion of routinely plotting data from an industrial process in a form that allows one to separate observed variability in a response into two kinds of variation. The first is that variation which appears to be inherent, unavoidable, short-term, baseline, and characteristic of the process (at least as currently configured). This variation Shewhart called *random* or *common cause variation*. The second kind of variability is that variation which appears to be avoidable, long-term, and/or due to sources outside of those seen as legitimately impacting process behavior. This variation he called *assignable* or *special cause variation.*

Shewhart reasoned that by plotting summary statistics from periodically collected data sets against time order of collection, one would be able to see interpretable trends or other evidence of assignable variation on the plots, and could intervene to eliminate the physical causes of that variation. The intention was to thereby make process output stable or consistent to within the inherent limits of process precision. As a means of differentiating plotted values that should signal the need for intervention from those that carry no special message of process distress, he suggested drawing so-called control limits on the plots. (The word "control" is something of a misnomer, at least as compared to common modern engineering usage of the word in referring to the active, moment-by-moment steering or regulation of processes. The nonstandard and more passive terminology "monitoring limits" would actually be far more descriptive of the purpose of Shewhart's limits.) These limits were to separate plausible values of the plotted statistic from implausible values when in fact the process was operating optimally, subject only to causes of variation that were part of standard conditions.

By far the most famous implementations of Shewhart's basic logic come where the plotted statistic is either the mean, the range, or, less frequently, the standard deviation. Such charts are commonly known by the names *x-bar charts, R charts*, and *s charts*, respectively. As a basis of discussion of Shewhart charts, consider the data given in Table 5.4. These values represent melt index measurements of specimens of extrusion grade polyethylene, taken four per shift in a plastics plant.

TABLE 5.4 Measured Melt Indices for Ten Groups of Four Specimens[a]

Shift	*Melt Index*	$\bar{y}$	*R*	*s*
1	218, 224, 220, 231	223.25	13	5.74
2	228, 236, 247, 234	236.25	19	7.93
3	280, 228, 228, 221	239.25	59	27.37
4	210, 249, 241, 246	236.50	39	17.97
5	243, 240, 230, 230	235.75	13	6.75
6	225, 250, 258, 244	244.25	33	14.06
7	240, 238, 240, 243	240.25	5	2.06
8	244, 248, 265, 234	247.75	31	12.92
9	238, 233, 252, 243	241.50	19	8.10
10	228, 238, 220, 230	229.00	18	7.39

[a] Based on data from page 207 of Wadsworth, Stephens, and Godfrey.[9]

Figure 5.10 shows plots of the individual melt indices, means, ranges, and standard deviations from Table 5.4 against shift number. The last three of these are the beginnings of so-called Shewhart $\bar{x}$, R, and s control charts.

What remain to be added to the plots in Fig. 5.10 are appropriate control limits. In order to indicate the kind of thinking that stands behind control limits for Shewhart charts, let us concentrate on the issue of limits for the plot of means. The fact is that mathematical theory suggests how the behavior of *means* $\bar{y}$ ought to be related to the distribution of *individual* melt indices y, provided the data-generating process is stable, that is, subject only to random causes. If individual responses y can be described as normal with some mean μ and standard deviation σ, mathematical theory suggests that averages of n such values will behave as if a different normal distribution were generating them, one with a mean $\mu_{\bar{y}}$ that is numerically equal to μ and with a standard deviation $\sigma_{\bar{y}}$ that is numerically equal to $\sigma/\sqrt{n}$. Figure 5.11 illustrates this theoretical relationship between the behavior of individuals and the behavior of means.

The relevance of Fig. 5.11 to the problem of setting control chart limits on means is that if one is furnished with a description of the typical pattern of variation in y, sensible expectations for variation in $\bar{y}$ follow from simple normal distribution calculations. So Shewhart reasoned that since about 99.7 percent (most) of a Gaussian distribution is within three standard deviations of the center of the distribution, means found to be farther than three theoretical standard deviations (of $\bar{y}$) from the theoretical mean (of $\bar{y}$) could be safely attributed to other than chance causes. Hence, furnished with standard values for μ and σ (describing individual observations), sensible control limits for $\bar{y}$ become

$$\text{Upper Control Limit (UCL) for } \bar{y} = \mu_{\bar{y}} + 3\sigma_{\bar{y}}$$

$$= \mu + 3\frac{\sigma}{\sqrt{n}}$$

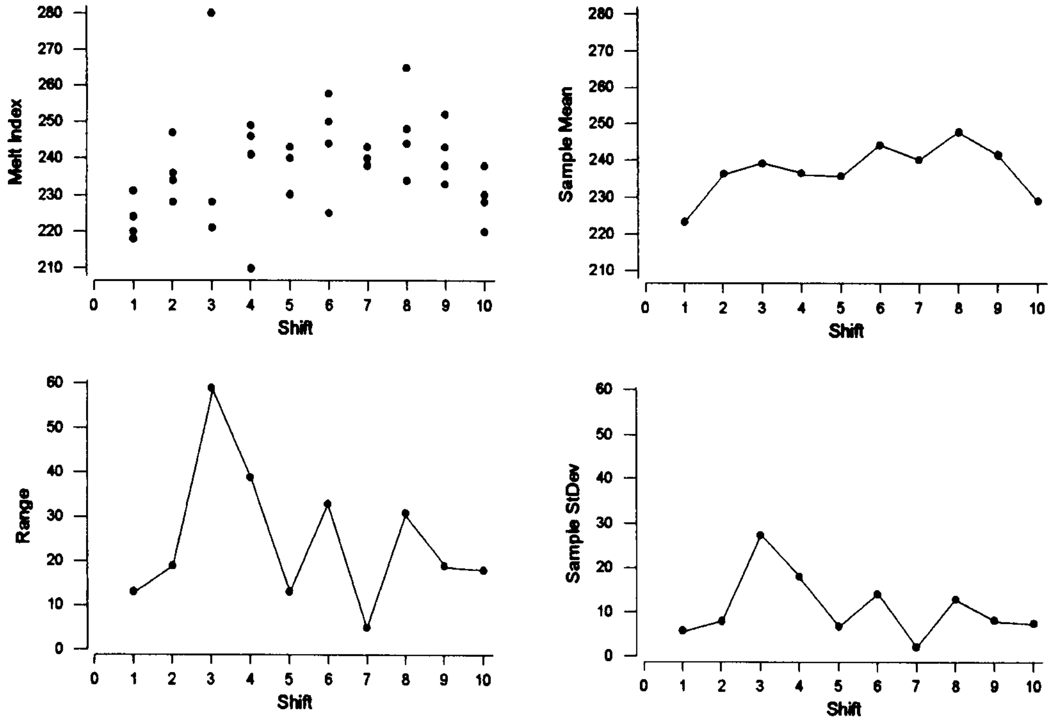

Fig. 5.10. Plots of melt index, $\bar{y}$, R, and s against shift number.

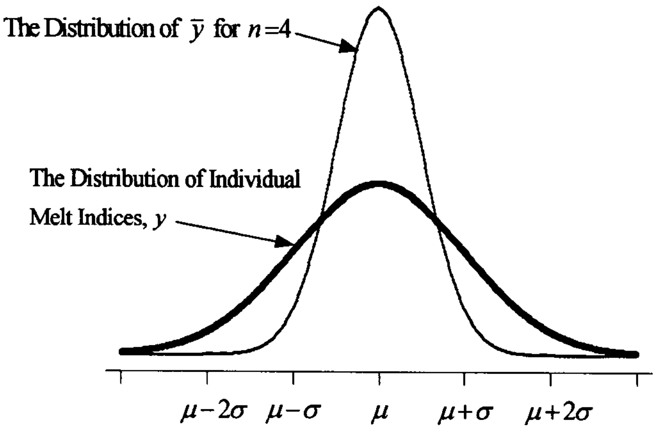

Fig. 5.11. The distribution of individuals, y, and sample means, $\bar{y}$.

and

$$\text{Lower Control Limit (LCL) for } \bar{y} = \mu_{\bar{y}} - 3\sigma_{\bar{y}} = \mu - 3\frac{\sigma}{\sqrt{n}} \qquad (5\text{-}9)$$

Returning to the context of our example represented by the data of Table 5.4, Wadsworth et al.[9] state that the target value for melt index in the original application was in fact 235. So if standard process behavior is "on target" behavior, the value $\mu = 235$ seems appropriate for use in (5-9). No parallel value for σ was provided by the authors. Common practice in such situations is to use the data in hand (the data of Table 5.4) to produce a plausible value for σ to use in (5-9). There are many possible ways to produce such a value, but to understand the general logic behind the standard ones, it is important to understand what σ is supposed to measure. The variable σ is intended as a theoretical measure of baseline, short-term, common cause variation. As such, the safest way to try to approximate it is to somehow use only measures of variation *within* the groups of four values in Table 5.4 *not* influenced by variation *between* groups. (Measures of variation derived from considering all the data simultaneously, e.g., would reflect variation between shifts as well as the shorter-term variation within shifts.) In fact, the most commonly used ways of obtaining from the data in hand a value of σ for use in (5-9) are based on the averages of the (within-group) ranges or standard deviations. For example, the 10 values of R given in Table 5.4 have a mean

$$\overline{R} = \frac{1}{10}(13+19+59+ \cdots +19+18) = 24.9$$

and some standard mathematical theory suggests that because the basic group size here is $n = 4$, an appropriate multiple of $\overline{R}$ for use in estimating σ is

$$\frac{\overline{R}}{2.059} \approx 12.1 \qquad (5\text{-}10)$$

(The divisor above is a tabled factor commonly called d_2, which increases with n.)

Finally, substituting 235 for μ and 12.1 for σ in (5-9) produces numerical control limits for $\bar{y}$:

$$LCL = 235 - 3\frac{(12.1)}{\sqrt{4}} = 216.9$$

and

$$UCL = 235 + 3\frac{(12.1)}{\sqrt{4}} = 253.1$$

Comparison of the $\bar{y}$ values in Table 5.4 to these limits reveals no "out of control" means, that is, no evidence in the means of assignable process variation. Figures 5.12 and 5.13 show control charts for all of $\bar{y}$, R, and s, where control limits for the last two quantities have been derived using standard calculations not shown here.

The R and s charts in Figs 5.12 and 5.13 are related representations (only one is typically made in practice) of the shift-to-shift behavior of melt index *consistency*. It is seen that on both charts, the shift #3 point plots above the upper control limit. The strong suggestion thus is that melt index consistency was detectably worse on that shift than on the others, so that from this point of view the process was in fact *not stable* over the time period represented in Table 5.4. In practice, physical investigation and hopefully correction of the origin of the

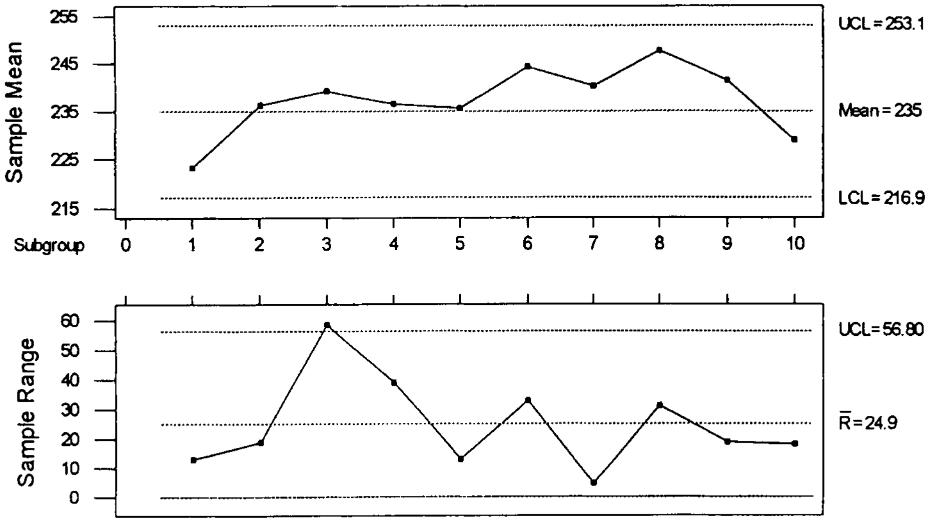

Fig. 5.12. Control charts for $\bar{y}$ and R based on melt indices.

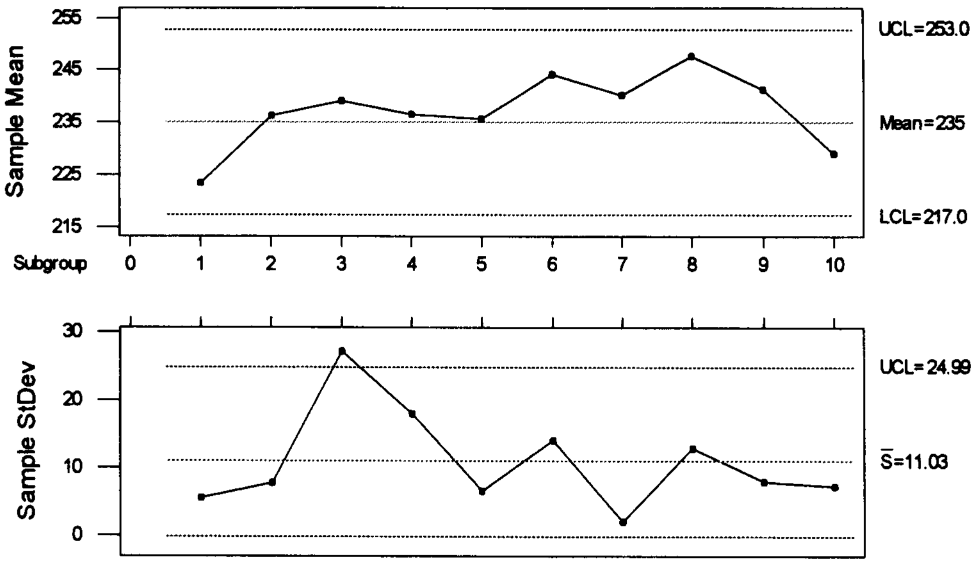

Fig. 5.13. Control charts for $\bar{y}$ and s based on melt indices.

instability typically would follow, as well as some reconsideration of our earlier assessment of 12.1 as a plausible figure to represent the inherent short-term variability of melt index. (If shift #3 could be treated as a special case, explainable as an unfortunate but correctable situation that was not expected to reoccur, there might be reason to revise $\bar{R}$ downward by deletion of shift #3 from the calculation, and thereby to reduce one's view of the size of baseline process variability. Notice that, in general, such a downward revision of $\bar{R}$ might well also have the effect of causing one to need to rethink his or her assessment of the constancy of the melt index *mean*.)

There is a variation on the basic "$\bar{x}$ and R chart" idea that we wish to illustrate here next, because of its frequent use in chemical industry applications. That is the making of a so-called x and MR chart pair. The motivation for this modification of the ideas outlined thus far in this section is that in many chemical process monitoring contexts the natural "group size" is $n = 1$. A mean of $n = 1$ observation(s) is simply that observation itself, and the limits of (5-9) make perfectly good sense for the case of $n = 1$. That is, the analog of an $\bar{x}$ chart for $n = 1$ cases is clear, *at least if one has an externally provided value for* σ. But what, if anything, to do for an $n = 1$ counterpart of the R chart and how to develop an analog of (5-10) in cases where σ is not a priori known are perhaps not so obvious. Table 5.5 contains data representing moisture contents in 0.01 percent of bihourly samples of a polymer, and the question at hand is what besides simply the bihourly y values might be plotted in the style of a Shewhart control chart for such data.

The final column of Table 5.5 gives 19 so-called moving ranges of pairs of successive moisture contents. It is often argued that although these MR values are actually affected not only by variation within a 2-hr production period but by some variation between these periods as well, they come as close to representing purely short-term variation as any measure available from $n = 1$ data. Accordingly, as a kind of $n = 1$ analog of an R chart, moving ranges are often charted in addition to individual values y. Further, the average moving range is used to estimate σ in cases where information on the inherent variability of individuals is a priori lacking, according to the formula

$$\text{estimated}\,\sigma = \frac{\overline{MR}}{1.128}$$

where $\overline{MR}$ is the mean of the moving ranges (and plays the role of $\bar{R}$ in (5-10)), and 1.128

is the $n = 2$ version of the factor d_2 alluded to immediately below (5-10).

In the case of the data of Table 5.5,

$$\overline{MR} = \frac{1}{19}(16 + 4 + 5 + \cdots + 16 + 18) \approx 8.2$$

so that a (possibly somewhat inflated due to between period variation) data-based estimate of within-period variability σ for use, for example in limits (5-9), is

$$\frac{8.2}{1.128} \approx 7.2$$

TABLE 5.5 Moisture Contents for 20 Polymer Samples[a]

Sample	*Moisture, y*	*Moving Range, MR*
1	36	—
2	20	16
3	16	4
4	21	5
5	32	11
6	34	2
7	32	2
8	34	2
9	23	11
10	25	2
11	12	13
12	31	19
13	25	6
14	31	6
15	34	3
16	38	4
17	26	12
18	29	3
19	45	16
20	27	18

[a] Based on data from page 190 of Burr.[10]

Figure 5.14 shows both an x (individuals) chart and an MR (moving range) chart based on these calculations. As no standard value of moisture content was provided in Burr's text,[10] the value $\bar{y} = 28.55$ was used as a substitute for μ in (5-9). The MR chart limits are based on standard $n = 2$ (because ranges of "groups" of two observations are being plotted) R chart control limit formulas. Figure 5.14 shows no evidence of assignable variation in the moisture contents.

Statistical research in the last decade has cast serious doubt on the wisdom of adding the MR chart to the x chart in $n = 1$ situations. The price paid for the addition in terms of "false alarm rate" is not really repaid with an important increase in the ability to detect process change. For a more complete discussion of this issue see Section 4.4 of Vardeman and Jobe.[14]

The use of Shewart control charts is admirably documented in a number of statistical quality control books including those by Vardeman and Jobe,[14] Wadsworth et al.,[9] Duncan,[11] Burr,[10] Grant and Leavenworth,[12] and Ott et al.[13] Our purpose here is not to

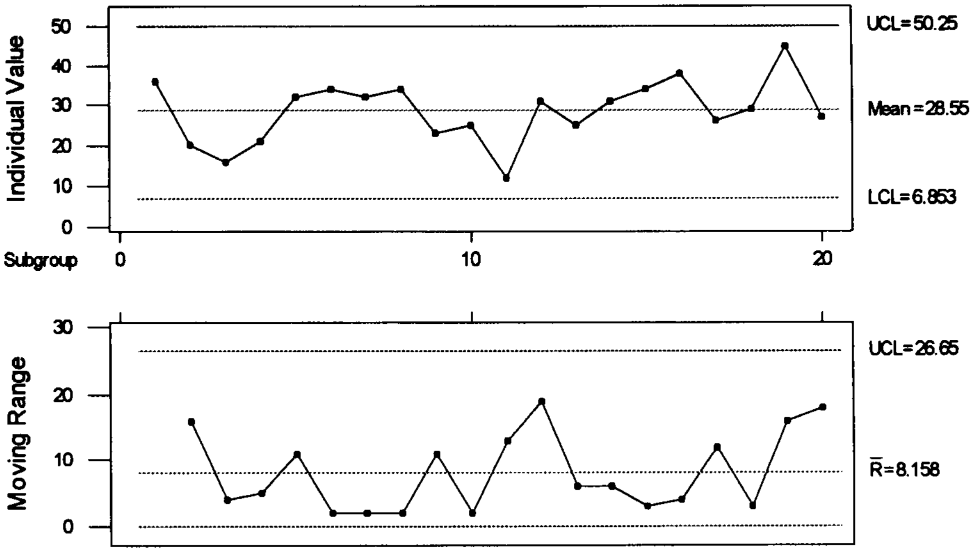

Fig. 5.14. Control charts for y and MR based on moisture contents of 20 polymer samples.

provide all details necessary for their use, but only to give the reader an introduction to the overall function that they serve. It should be said, however, that in recent years other statistical process monitoring tools such as the so-called CUmulative SUM (CUSUM) schemes and Exponentially Weighted Moving Average (EWMA) schemes have been developed as competing methodologies, and can in some circumstances be practically more effective than the original Shewhart charts. Indeed, many computerized controllers for real-time chemical process monitoring and adjustment now employ some form of CUSUM or EWMA logic. For more on these topics, including their integration with model-based process controllers, the reader is referred to Sections 4.1 and 4.2 of Vardeman and Jobe[14] and Vander Wiel et al.[15]

Shewhart's basic conceptualization of common and special cause variation not only leads to control charts as quantitative, rational tools to guide one in knowing when (and when not!) to intervene in an industrial process to correct potential ills, but it also provides a framework for considering the question of what is the best/most consistent performance one can hope for from a particular version of a process. That is, it provides a framework for discussing process capability assessment.

If $\hat{\sigma}$ is some (standard deviation type) estimate of the baseline variation inherent in an industrial process (obtained, e.g., from a calculation such as (5-10) or from data taken from the process after eliminating all physical sources of assignable variation), it essentially specifies what is possible in terms of consistency of process output. There are, however, several common ways of using such an estimate to produce related measures of process capability.

For one thing, remembering again the fact that an interval from $\mu - 3\sigma$ to $\mu + 3\sigma$ (i.e., of length 6σ) will bracket about 99.7 percent of a normal distribution, the figure 6σ is sometimes stated as "the process capability." This usage would say that in the context of the polyethylene melt index example of Table 5.4 the $\hat{\sigma} = 12.1$ figure from (5-10) implies a melt index process capability of approximately $6 \cdot (12.1) \approx 72.6$. If properly monitored, the process appears capable of producing almost all individual melt indices in a 73-point range.

Where there are stated specifications for individual measurements y, σ is sometimes turned into a kind of index comparing it to the difference in upper and lower engineering specifications. For example, one such *process capability index* is

$$C_{\mathrm{p}} = \frac{USL-LSL}{6\sigma}$$

where $USL - LSL$ is the difference in specifications. Fairly obviously, the larger the value of C_{p}, the more comfortably (properly targeted) process output values will fit in an interval from LSL to USL.

Another process capability measure that is frequently used in the industrial chemistry sector is

$$C_{\mathrm{pk}} = \text{minimum} \left\{ C_{\mathrm{pu}} = \frac{USL - \mu}{3\sigma}, \; C_{\mathrm{pl}} = \frac{\mu - LSL}{3\sigma} \right\}$$

where μ is an overall process average for an in-control/stable/predictable process, and σ is as before. This measure is clearly similar to C_{p}, but it takes into account the placement of the process mean in a way that is ignored by C_{p}. A large value of C_{pk} indicates that not only is the process short-term variation small enough for the process output values to potentially fit comfortably between LSL and USL, but that the process is currently so targeted that the potential is being realized.

STATISTICAL METHODS AND INDUSTRIAL EXPERIMENTATION

One of the most important areas of opportunity for the new application of statistical methods in the chemical industry in the twenty-first century is that of increasing the effectiveness of industrial experimentation. That is, it is one thing to bring an existing industrial process to stability (a state of "statistical" control), but it is quite another to determine how to make fundamental changes in that process that will improve its basic behavior. This second activity almost always involves some form of experimentation, whether it be in the laboratory or in a plant. As we indicated in the introduction, efficient methods and strategies of such data collection (and corresponding analysis) are a central concern of applied statistics. In this section, we hope to give the reader some insight into the kinds of statistical tools that are available for use in chemical industry experimentation.

We will here take as our meaning of the term "experimentation" the observation of a (typically noisy) physical process under more than one condition, with the broad goal of understanding and then using knowledge of how the process reacts to the changes in conditions. In most industrial contexts, the "conditions" under which the process is observed can be specified in terms of the settings or so-called levels chosen for several potentially important process or environmental variables, the so-called factors in the experiment. In some cases, the hope is to identify those (often largely unregulated) factors and combinations of factors that seem to most influence an observed response variable, as a means of targeting them for attention intended to keep them constant or otherwise to eliminate their influence, and thereby to improve the consistency of the response. In other situations the hope is to discover patterns in how one or more critical responses depend on the levels of (often tightly controlled) factors, in order to provide a road map for the advantageous guiding of process behavior (e.g., to an increased mean reaction yield) through enlightened changing of those levels.

This section is organized into two subsections. In the first, we will illustrate the notion of variance component estimation through an example of a nested or hierarchical data collection scheme. In the second, we will discuss some general considerations in the planning of experiments to detail the pattern of influence of factors on responses, consider so-called factorial and fractional factorial experimental designs, illustrate response surface fitting and interpretation tools and the data requirements they imply, and, in the process, discuss the integration of a number of statistical tools in a sequential learning strategy.

Identifying Major Contributors to Process Variation

A statistical methodology that is particularly relevant where experimentation is meant to identify important unregulated sources of variation in a response is that of variance component estimation, based on so-called ANalysis Of VAriance (ANOVA) calculations and random effects models. As an example of what is possible, consider the data of Table 5.6 Shown here are copper content measurements for some bronze castings. Two copper content determinations were made on each of two physical specimens cut from each of 11 different castings.

The data of Table 5.6 were *by design* collected to have a "tree type" or so-called hierarchical/nested structure. Figure 5.15 shows a diagram of a generic hierarchical structure for *balanced* cases like the present one, where there are equal numbers of branches leaving all nodes at a given level (there are equal numbers of determinations for each specimen and equal numbers of specimens for each casting).

An important goal in most hierarchical studies is determining the size of the contributions to response variation provided by the different factors, that is, the different levels of the tree structure. (In the present context, the issue is how variation between castings compares to variation between specimens within a casting, and how they both compare to variation between determinations for a given specimen. If the overall variability observed were considered excessive, such analysis could then help guide efforts at variation reduction by identifying the largest contributors to observed variability.) The structure portrayed in Fig. 5.15 turns out to enable an appealing statistical analysis, providing help in that quantification.

If one lets

y_{ijk} = the copper content from the kth determination of the jth specimen from casting i

$\bar{y}_{ij\cdot} = \frac{1}{2}\sum_k y_{ijk}$ = the mean copper content determination from the *j*th specimen from casting *i*

$\bar{y}_{i\cdot\cdot} = \frac{1}{2}\sum_j \bar{y}_{ij\cdot}$ = the mean copper content determination from the *i*th casting

and

$\bar{y}_{\cdot\cdot\cdot} = \frac{1}{11}\sum_i y_{i\cdot\cdot}$ = the overall mean copper determination

it is possible to essentially break down the variance of all 44 copper contents (treated as a single group) into interpretable pieces, identifiable as variation between $\bar{y}_{i\cdot\cdot}$s (casting means), variation between $\bar{y}_{ij\cdot}$s (specimen means) within castings, and variation between $\bar{y}_{ijk}$s (individual measurements) within a specimen. That is, it is an algebraic identity that for 44 numbers y_{ijk} with the same structure as those in Table 5.6

$$(44-1)s^2 = \sum_{i,j,k}(y_{ijk} - \bar{y}_{\cdot\cdot\cdot})^2$$
$$= \sum_{i,j,k}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 + \sum_{i,j,k}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot})^2$$
$$+ \sum_{i,j,k}(y_{ijk} - \bar{y}_{ij\cdot})^2 \qquad (5\text{-}11)$$

The sums indicated in (5-11) are over all data points; so, for example, the first summand on the right is obtained for the copper content data by summing each $(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$ a total of $2 \cdot 2 = 4$ times, one for each determination on a given casting. With the obvious meaning for the $\bar{y}$s and the substitution of the total number of data values for 44, the identity in (5-11) applies to any balanced hierarchical data structure. It is a so-called ANOVA identity, providing an intuitively appealing partitioning of the overall observed variability in the data, an *analyzing of the (observed) variation.*

Some tedious arithmetic "by hand," or use of nearly any commercially available statistical package that includes an ANOVA program, shows that for the copper content data

TABLE 5.6 Forty-four Copper Content Measurements from 11 Bronze Castings[a]

Casting	*Specimen*	*Determination*	*Copper Content, y, (%)*
1	1	1	85.54
1	1	2	85.56
1	2	1	85.51
1	2	2	85.54
2	1	1	85.54
2	1	2	85.60
2	2	1	85.25
2	2	2	85.25
3	1	1	85.72
3	1	2	85.77
3	2	1	84.94
3	2	2	84.95
4	1	1	85.48
4	1	2	85.50
4	2	1	84.98
4	2	2	85.02
5	1	1	85.54
5	1	2	85.57
5	2	1	85.84
5	2	2	85.84
6	1	1	85.72
6	1	2	85.86
6	2	1	85.81
6	2	2	85.91
7	1	1	85.72
7	1	2	85.76
7	2	1	85.81
7	2	2	85.84
8	1	1	86.12
8	1	2	86.12
8	2	1	86.12
8	2	2	86.20
9	1	1	85.47
9	1	2	85.49
9	2	1	85.75
9	2	2	85.77
10	1	1	84.98
10	1	2	85.10
10	2	1	85.90
10	2	2	85.90
11	1	1	85.12
11	1	2	85.17
11	2	1	85.18
11	2	2	85.24

[a] Based on data taken from Wernimont.[16]

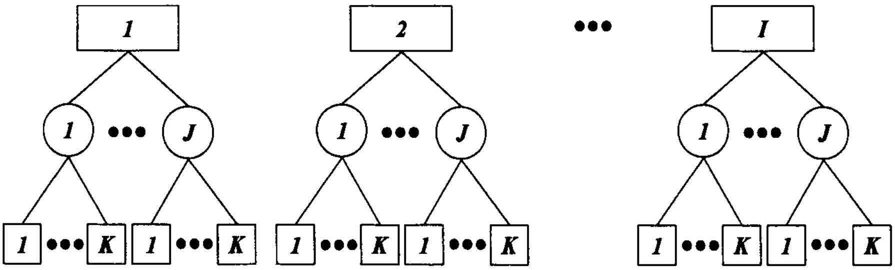

Fig. 5.15. A balanced hierarchical data structure.

of Table 5.6 the numerical version of (5-11) is approximately

$$5.1385 = 3.2031 + 1.9003 + 0.0351 \qquad (5\text{-}12)$$

Although we will not provide any details here, the reader is alerted to the fact that it is common practice to present the elements of an identity such as (5-12) in a tabular form called an "ANOVA table." The use for the elements of (5-12) that we wish to illustrate here is their role in estimating casting, specimen, and determination "variance components."

That is, if one models an observed copper determination as the sum of a *random* casting-dependent *effect* whose distribution is described by a variance σ_c^s, a *random* specimen-dependent *effect* whose distribution is described by a variance σ_s^2, and a *random* determination-dependent *effect* whose distribution is described by a variance σ_d^2, the elements of (5-12) lead to estimates of the *variance components* σ_c^2, σ_s^2, and σ_d^2 in the model. Note that in such a random effects model of the data-generating process, copper measurements from the same casting share the same casting effect, and copper measurements from the same specimen share the both same casting and the same specimen effects. The individual σ^2 values are conceptually the variances that would be seen in copper contents if only the corresponding sources of variation were present. The sum of the σ^2 values is conceptually the variance that would be seen in copper contents if single determinations were made on a number of different castings.

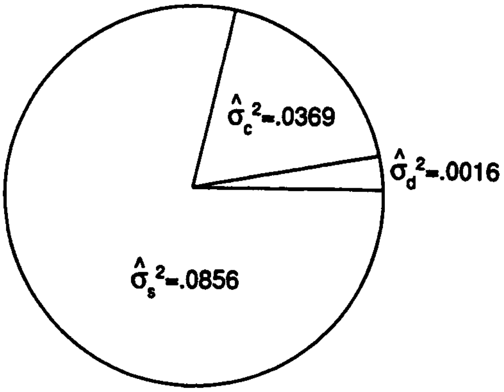

Fig. 5.16. Three estimated variance components for copper contents.

Standard statistical methodology for estimation of the variance components (which we will not detail here but can, e.g., be found in Section 5.5 of Vardeman and Jobe[14] or Chapter 11 of Hicks and Turner[17]) produces

$$\sigma_d^2 = \frac{0.0351}{11\cdot 2\cdot(2-1)} \approx 0.0016\,(\%)^2$$

as an estimate of σ_d^2,

$$\hat{\sigma}_s^2 = \frac{1}{2}\left(\frac{1.9003}{11\cdot(2-1)} - 0.0016\right) \approx 0.0856\,(\%)^2$$

as an estimate of σ_s^2, and

$$\hat{\sigma}_c^2 = \frac{1}{2\cdot 2}\left(\frac{3.2031}{(11-1)} - \frac{1.9003}{11\cdot(2-1)}\right)$$
$$\approx 0.0369\,(\%)^2$$

as an estimate of σ_c^2. Figure 5.16 is a pie chart representation of these three estimated variance components as fractions of their sum (the variance predicted if single determinations were made on single specimens from each casting), and graphically identifies inhomogeneity between specimens cut from a single casting as the biggest contributor to observed variation.

On the standard deviation scale the estimates translate to $\hat{\sigma}_d \approx 0.04\%$, $\hat{\sigma}_s \approx 0.29\%$, $\hat{\sigma}_c \approx 0.19\%$. So, for example, the data of Table 5.6 indicate that even if castings and specimens were all exactly alike, it would still be reasonable to expect measured copper contents to vary according to a standard deviation of about 0.04 percent, presumably due to unavoidable measurement error.

Variance component estimation methodology is not limited to balanced hierarchical experiments, but they do provide an important and straightforward context in which to introduce the technology. More detailed information on the case discussed here and extensions to other kinds of data structures can be found in books by Vardeman,[6] Neter et al.,[7] Mason, Gunst, and Hess,[18] and Hicks and Turner.[17]

Discovering and Exploiting Patterns of Factor Influence on Responses

Having discussed statistical methodology particularly appropriate to studies whose primary purpose is simply to identify factors with the largest influence on a response, we will now consider methods aimed more directly at detailed experimental quantification of the pattern of factor influence on one or more responses. As an example, we will use a "sanitized" account of some statistical aspects of a highly successful and economically important process improvement project. (Data presented here are not the original data, but resemble them in structure. Naturally, details of the project not central to our expository purposes and those of a proprietary nature will be suppressed.) A more complete version of this case study appears as Chapter 11 of Vardeman.[6]

The process monitoring, capability assessment, and variance source identification ideas discussed thus far are almost logical prerequisites for industrial experimentation to detail the nature of dependence of response variables on factors of interest. When an industrial process has been made to operate in a stable manner, its intrinsic variability reduced to the extent practically possible, and that baseline performance quantified and understood, the prospects of success are greatly enhanced for subsequent efforts to understand the effects of potential fundamental process changes.

Preliminary work by various groups left a project team with a batch production process behaving in a stable but unsatisfactory fashion. Obvious sources of variation (both in the process itself and "upstream") had been identified and, to the degree practically possible, eliminated. The result was a process with an average output *purity* of 88 percent and an associated purity standard deviation of around 5 percent, and an average *yield* of 43 percent and an associated yield standard deviation of around 5 percent as well. The project team was charged with finding ways to increase the purity and yield means to, respectively, 95 percent and 59 percent while it is hoped, also further reducing the standard deviations. To accomplish this, the team recognized the need for an improved understanding of how various process variables under their control influenced purity (which we will call y_1) and yield (which we will call y_2). Experimentation to provide this was authorized, and, in particular, attention was focused on four factors consisting of three reactant concentrations and the process run time. We will call the Reactant A mole ratio x_1, the Reactant B mole ratio x_2, the Reactant C mole ratio x_3, and the run time (in hours) x_4.

The choice of experimental factors (what to vary in data collection) is a nontrivial matter of fundamental importance that is best handled by people with firsthand process knowledge. There are a number of popular techniques and tools (such as so-called cause and effect diagrams, discussed for instance in Section 2.1 of Vardeman and Jobe[14]) for helping groups brainstorm and reach a consensus on such matters. Further, in cases where a priori knowledge of a process is scarce, relatively small preliminary screening experiments can help reduce a large list of potential factors to a smaller list apparently worthy of more detailed study. (The fractional factorial plans that will be illustrated shortly often are recommended for this purpose.)

Once a particular set of experimental factors has been identified, questions about exactly how they should be varied must be answered. To begin with, there is the choice of levels for the factors, the matter of how much the experimental factors should be varied. Particular experimental circumstances usually dictate how this is addressed. Widely spaced (substantially different) levels will in general lead to bigger changes in responses, and therefore clearer indications of how the responses depend upon the experimental factors, than will closely spaced (marginally different) levels. But they may do so at the expense of potentially creating unacceptable or even disastrous process conditions or output. Thus, what may be an acceptable strategy in a laboratory study might be completely unacceptable in a production environment and vice versa.

Given a set or range of levels for each of the individual experimental factors, there is still the question of exactly what combinations of levels actually will be used to produce experimental data. For example, in the process improvement study, standard process operating conditions were $x_1 = 1.5$, $x_2 = 1.15$, $x_3 = 1.75$, and $x_4 = 3.5$, and the project team decided on the ranges

$$1.0 \leqslant x_1 \leqslant 2.5, 1.0 \leqslant x_2 \leqslant 1.8, 1.0 \leqslant x_3 \leqslant 2.5,$$

$$\text{and} \qquad (5\text{-}13)$$

$$2.0 \leqslant x_4 \leqslant 5.0$$

as defining the initial limits of experimentation. But the question remained as to exactly what sets of mole ratios and corresponding run times were appropriate for data collection.

A natural (but largely discredited) strategy of data collection is the one-variable-at-a-time experimental strategy of picking some base of experimental operations (such as standard operating conditions) and varying the level of only one of the factors away from that base at a time. The problem with such a strategy is that sometimes two or more factors act on responses jointly, doing things in concert that neither will do alone. For example, in the process improvement study, it might well have been that an increase in either x_1 or x_2 alone would have affected yield very little, whereas a simultaneous increase in both would have caused an important increase. Modern strategies of industrial experimentation are conceived with such possibilities in mind, and attempt to spread out observations in a way that gives one some ability to identify the nature of the response structure no matter how simple or complicated it turns out to be.

There are several issues to consider when planning the combinations of levels to include in an experiment. We have already said that it is important to "vary several factors simultaneously." It also is important to provide for some replication of at least a combination or two in the experiment, as a means of getting a handle on the size of the experimental error or baseline variation that one is facing. The replication both verifies the reproducibility of values obtained in the study and identifies the limits of that reproducibility. Also, one must balance the urge to "cover the waterfront" with a wide variety of combinations of factor levels against resource constraints and a very real law of diminishing practical returns as one goes beyond what is really needed in the way of data to characterize response behavior. In addition, the fact that real-world learning is almost always of a sequential rather than a "one shot" nature suggests that it is in general wise to spend only part of an experimental budget on early study phases, leaving resources adequate to follow up directions suggested by what is learned in those stages.

It is obvious that a minimum of two different levels of an experimental factor must appear in a set of experimental combinations if any information is to be gained on the effects of that factor. So one logical place to begin thinking about a candidate design for an industrial experiment is with the set of all possible combinations of two levels of each of the experimental factors. If there are p experimental factors, statistical jargon for such an arrangement is to call it a (complete) $2 \times 2 \times 2 \times \cdots \times 2$ or 2^p factorial plan. For example, in the process improvement situation, an experiment consisting of the running of all 16 possible combinations of

$$\begin{aligned} x_1 &= 1.0 \quad \text{or} \quad x_1 = 2.5 \\ x_2 &= 1.0 \quad \text{or} \quad x_2 = 1.8 \\ x_3 &= 1.0 \quad \text{or} \quad x_3 = 2.5 \end{aligned}$$

and

$$x_4 = 2.0 \quad \text{or} \quad x_4 = 5.0$$

would be called a complete 2 × 2 × 2 × 2 or 2^4 factorial experiment. Notice that in geometric terms, the (x_1, x_2, x_3, x_4) points making up this 2^4 structure amount to the 16 "corners" in four-dimensional space of the initial experimental region defined in (5-13).

A complete factorial experimental plan is just that, in some sense "complete." It provides enough information to allow one to assess (for the particular levels used) not only individual but also joint or interaction effects of the factors on the response or responses. But when in fact (unbeknownst to the investigator) a system under study is a relatively simple one, principally driven by only a few individual or low-order joint effects of the factors, fewer data actually are needed to characterize those effects adequately. So what is often done in modern practice is initially to run only a carefully chosen part of a full 2^p factorial, a so-called fractional factorial plan, and to decide based on the initial data whether data from the rest of the full factorial appear to be needed in order adequately to characterize and understand response behavior. We will not discuss here the details of how so-called 2^{p-q} fractional factorials are intelligently chosen, but there is accessible reading material on the subject in books by Box, Hunter, and Hunter,[19] and by Vardeman and Jobe.[5]

In the process improvement study, what was actually done in the first stage of data collection was to gather information from one-half of a full 2^4 factorial (a 2^{4-1} fractional factorial) augmented by four observations at the "center" of the experimental region (thereby providing both some coverage of the interior of the region, in addition to a view of some of its corners, and important replication as well).

TABLE 5.7 Data from an Initial Phase of a Process Improvement Study

x_1	x_2	x_3	x_4	*Purity,* $y_1(\%)$	*Yield,* $y_2(\%)$
1.00	1.0	1.00	2.0	62.1	35.1
2.50	1.0	1.00	5.0	92.2	45.9
1.00	1.8	1.00	5.0	7.0	4.0
2.50	1.8	1.00	2.0	84.0	46.0
1.00	1.0	2.50	5.0	61.1	41.4
2.50	1.0	2.50	2.0	91.6	51.2
1.00	1.8	2.50	2.0	9.0	10.0
2.50	1.8	2.50	5.0	83.7	52.8
1.75	1.4	1.75	3.5	87.7	54.7
1.75	1.4	1.75	3.5	89.8	52.8
1.75	1.4	1.75	3.5	86.5	53.3
1.75	1.4	1.75	3.5	87.3	52.0

The data in Table 5.7 are representative of what the group obtained.

The order in which the data are listed is simply a convenient systematic one, not to be confused with the order in which experimental runs were actually made. The table order is far too regular for it to constitute a wise choice itself. For example, the fact that all $x_3 = 1.0$ combinations precede the $x_3 = 2.5$ ones might have the unfortunate effect of allowing the impact of unnoticed environmental changes over the study period to end up being confused with the impact of x_3 changes. The order in which the 12 experimental runs were actually made was chosen in a "completely randomized" fashion. For a readable short discussion of the role of randomization in industrial experimentation, the reader is referred to Box.[20]

For purposes of this discussion, attention is focused on the yield response variable, y_2. Notice first that the four y_2 values from the center point of the experimental region have $\bar{y} = 53.2$ and $s = 1.13$ (which incidentally already appear to be an improvement over typical process behavior). As a partial indication of the logic that can be used to investigate whether the dependence of yield on the experimental factors is simple enough to be described adequately by the data of Table 5.7, one can compute some estimated "main effects" from the first eight data points. That is, considering first the impact of the variable x_1 (alone) on yield, the quantity

$$\begin{aligned}\bar{y}_{\text{high}\,x_1} - \bar{y}_{\text{low}\,x_1} &= \tfrac{1}{4}(45.9+46.0+51.2+52.8)\\ &\quad - \tfrac{1}{4}(35.1+4.0+41.4+10.0)\\ &= 26.35\end{aligned}$$

is perhaps a sensible measure of how a change in x_1 from 1.00 to 2.50 is reflected in yield. Similar measures for the other variables turn out to be

$$\bar{y}_{\text{high}\,x_2} - \bar{y}_{\text{low}\,x_2} = -\ 15.20$$

$$\bar{y}_{\text{high}\,x_3} - \bar{y}_{\text{low}\,x_3} = 6.10$$

and

$$\bar{y}_{\text{high}\,x_4} - \bar{y}_{\text{low}\,x_4} = 0.45$$

These measures provide some crude insight into the directions and magnitudes of influence of the experimental variables on y_2. (Clearly, by these measures $x_1 = 2.50$ seems preferable to $x_1 = 1.00$, and the run time variable x_4 seems to have little impact on yield.) But they also provide strong evidence that the nature of the dependence of yield on the experimental factors is too complicated to be described by the action of the factors individually. For example, if it *were* the case that the separate actions of the experimental factors were adequate to describe system behavior, then standard statistical theory and the data indicate that the mean response for the $x_1 = 1.00$, $x_2 = 1.0$, $x_3 = 1.00$, and $x_4 = 2.0$ set of conditions would be around

$$\begin{aligned}\hat{y} = \bar{y}_{\text{corners}} - \tfrac{1}{2}(-26.35) - \tfrac{1}{2}(-15.20)\\ - \tfrac{1}{2}(6.10) - \tfrac{1}{2}(0.45) = 27.45\end{aligned}$$

(where $\bar{y}_{\text{corners}}$ is standing for the mean of the first eight yields in Table 5.7). But the observed yield of 35.1 is clearly incompatible with such a mean and the standard deviation value (of $s = 1.13$) derived from the repeated center point. Also, other simple evidence that (at least linear and) separate action of the four factors is not enough to describe yield adequately is given by the large difference between $\bar{y}_{\text{corners}} = 35.8$ and the observed mean from the center point $\bar{y} = 53.2$. (As it turns out, calculations that we will not show here indicate the *possibility* that individual action of the factors *plus joint action* of the Reactant A and Reactant B mole ratios is sufficient to describe yield. But in any case, the point is that the data of Table 5.7 provide evidence that the pattern of dependence of yield on the experimental variables is not simple, and thus that completion of the 2^4 factorial is in order.)

TABLE 5.8 Data from a Second Phase of a Process Improvement Study

x_1	x_2	x_3	x_4	*Purity,* $y_1(\%)$	*Yield,* $y_2(\%)$
1.00	1.0	1.00	5.0	64.0	35.3
2.50	1.0	1.00	2.0	91.9	47.2
1.00	1.8	1.00	2.0	6.5	3.9
2.50	1.8	1.00	5.0	86.4	45.9
1.00	1.0	2.50	2.0	63.9	39.5
2.50	1.0	2.50	5.0	93.1	51.6
1.00	1.8	2.50	5.0	6.8	9.2
2.50	1.8	2.50	2.0	84.6	54.3

After a complete analysis of the first round of experimental data, the project team "ran the second half fraction" of the 2^4 factorial, and data similar to those in Table 5.8 were obtained. (Again, no significance should be attached to the order in which the observations in Table 5.8 are listed. It is not the order in which the experimental runs were made.)

The data from the second phase of experimentation served to complete the project team's 2^4 factorial picture of yield and confirm the tentative understanding drawn first from the initial half fraction. It is seen that the combinations listed in Table 5.8 are in the same order as the first eight in Table 5.7 as regards levels of experimental variables x_1, x_2, and x_3, and that the corresponding responses are very similar. (This, by the way, has the happy practical implication that run time seems to have little effect on final purity or yield, opening the possibility of reducing or at least not increasing the standard run time.) Thorough data analysis of a type not shown here left the project team with a clear (and quantified version of the) understanding that

Reactant A and B mole ratios have important individual and joint effects on the responses, and that, acting independently of the other two reactants, Reactant C also has an important effect on the responses. However, it did *not* yet provide a solution to the team's basic problem, which was to reach a 59 percent mean yield goal.

The data of Tables 5.7 and 5.8 do hold out hope that conditions producing the desired purity and yield can be found. That is, though none of the 16 corners of the experimental region nor the center point appeared to meet the team's yield goal, the data do show that there is substantial *curvature* in the yield response. (The joint effect of x_1 and x_2 amounts to a kind of curvature, and the non-linearity of response indicated by a large difference between $\bar{y}_{\text{corners}} \approx 35.8$ and $\bar{y} = 53.2$ at the center of the experimental region also is a kind of curvature.) If one could "map" the nature of the curvature, there is at least the possibility of finding favorable future operating conditions in the interior of the initial experimental region defined in (5-13).

It ought to be at least plausible to the reader that 2^4 factorial data (even supplemented with center points) are not really sufficient to interpolate the nature of a curved response over the experimental region. More data are needed, and a standard way of augmenting a 2^p design with center points to one sufficient to do the job is through the addition of so-called star points to produce a central composite design. Star points are points outside the original experimental region whose levels of all but one of the p experimental factors match those of the center point. Figure 5.17 shows graphical representations of central composite designs in $p = 2$ and $p = 3$ factors.

The project team conducted a third phase of experimentation by adding eight star points to their study and obtained data similar to those in Table 5.9.

The data in Tables 5.7–5.9 taken together turn out to provide enough information to enable one to rather thoroughly quantify the "curved" nature of the dependence of y_2 on x_1, x_2, x_3, and x_4. A convenient and often successful method of accomplishing this quantification is through the least squares fitting of a general *quadratic response surface*. That is, central composite data are sufficient to allow one to fit an equation to a response that involves a constant term, linear terms in all the experimental variables, quadratic terms in all of the experimental variables, and cross-product terms in all pairs of the experimental variables. Appropriate use of a multiple regression program with the project

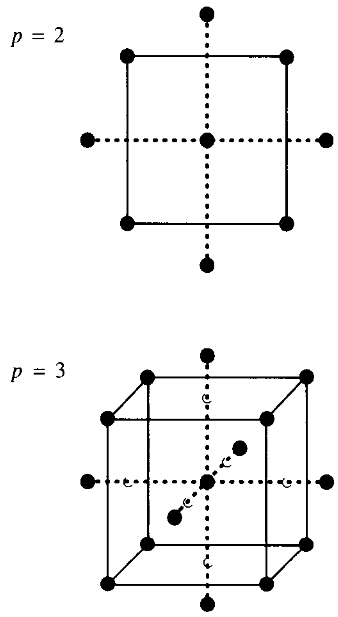

Fig. 5.17. $p = 2$ and $p = 3$ central composite designs.

TABLE 5.9 Data from a Third Phase of a Process Improvement Study

				Purity,	*Yield,*
x_1	x_2	x_3	x_4	$y_1(\%)$	$y_2(\%)$
0.6895	1.4	1.75	3.5	20.8	13.0
2.8105	1.4	1.75	3.5	95.9	54.3
1.75	0.8344	1.75	3.5	99.9	62.4
1.75	1.9656	1.75	3.5	65.9	41.2
1.75	1.4	0.6895	3.5	64.4	32.7
1.75	1.4	2.8105	3.5	64.8	40.3
1.75	1.4	1.75	1.379	88.1	52.7
1.75	1.4	1.75	5.621	88.9	50.5

data represented here produces the fitted equation

$$\begin{aligned} y_2 \approx\ & 15.4 + 37.9x_1 - 66.2x_2 + 48.8x_3 \\ & + 0.97x_4 - 16.1x_1^2 - 0.03x_2^2 \\ & - 13.6x_3^2 - 0.046x_4^2 + 26.5x_1x_2 \\ & + 0.344x_1x_3 - 0.217x_1x_4 \\ & + 1.31x_2x_3 - 0.365x_2x_4 + 0.061x_3x_4 \end{aligned}$$

This may not seem to the reader to be a particularly helpful data summary, but standard multiple regression tools can be used to deduce that an essentially equivalent, far less cluttered, and more clearly interpretable representation of the relationship is:

$$\begin{aligned} y_2 \approx\ & 13.8 + 37.8x_1 - 65.3x_2 + 51.6x_3 \\ & - 16.2x_1^2 - 13.6x_3^2 + 26.5x_1x_2 \qquad (5\text{-}14) \end{aligned}$$

Equation (5-14) provides an admirable fit to the data in Tables 5.7–5.9, is in perfect agreement with all that has been said thus far about the pattern of dependence of yield on the experimental factors, *and* allows one to do some intelligent interpolation in the initial experimental region. Use of an equation like (5-14) ultimately allowed the project team to determine that an increase of x_1 only would, with minimal change in the existing process, allow them to meet their yield goal. (In fact, the single change in x_1 proved to be adequate to allow them to meet all of their yield *and* purity goals!)

Graphical representations similar to those in Figs 5.18 and 5.19 for (5-14) with $x_3 = 1.75$ (the standard operating value for x_3) were instrumental in helping the team understand the message carried by their data and how yield could be improved. Figure 5.18 is a so-called contour plot (essentially a topographic map) of the fitted equation, and Fig. 5.19 is a more three-dimensional-looking representation of the same surface. Both types of display are commonly used tools of modern statistical experiment design and analysis. The contour plot idea is particularly helpful where several responses are involved, and by overlaying several such plots one can simultaneously picture the various implications of a contemplated choice of process conditions.

SPECIAL STATISTICAL TOOLS FOR CHEMICAL APPLICATIONS

The statistical methods discussed thus far are of a quite general nature, routinely finding application beyond the bounds of the chemical industry. In this section, we will briefly highlight two statistical methodologies whose most important applications are to chemical

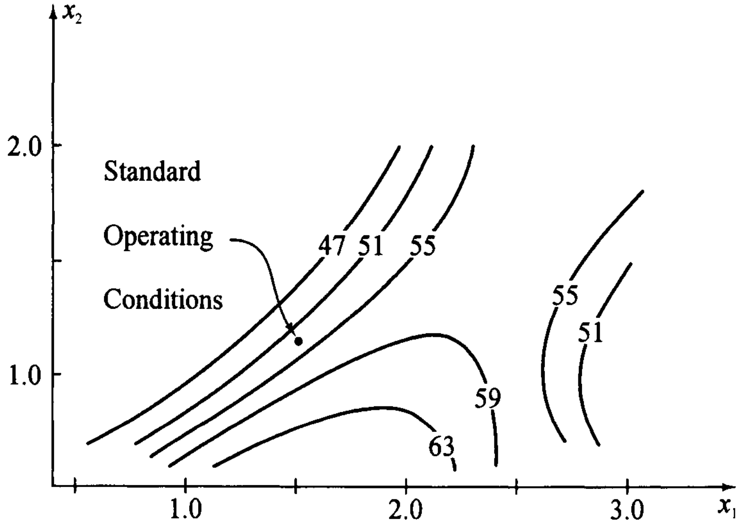

Fig. 5.18. A contour plot of fitted yield when $x_3 = 1.75$. (From *Statistics for Engineering Problem Solving (1st Ed.)* by S. B. Vardeman © 1994. Reprinted with permission of Brooks/Cole, a Division of Thomson Learning; www.thomsonlearning.com. FAX 800-730-2215.)

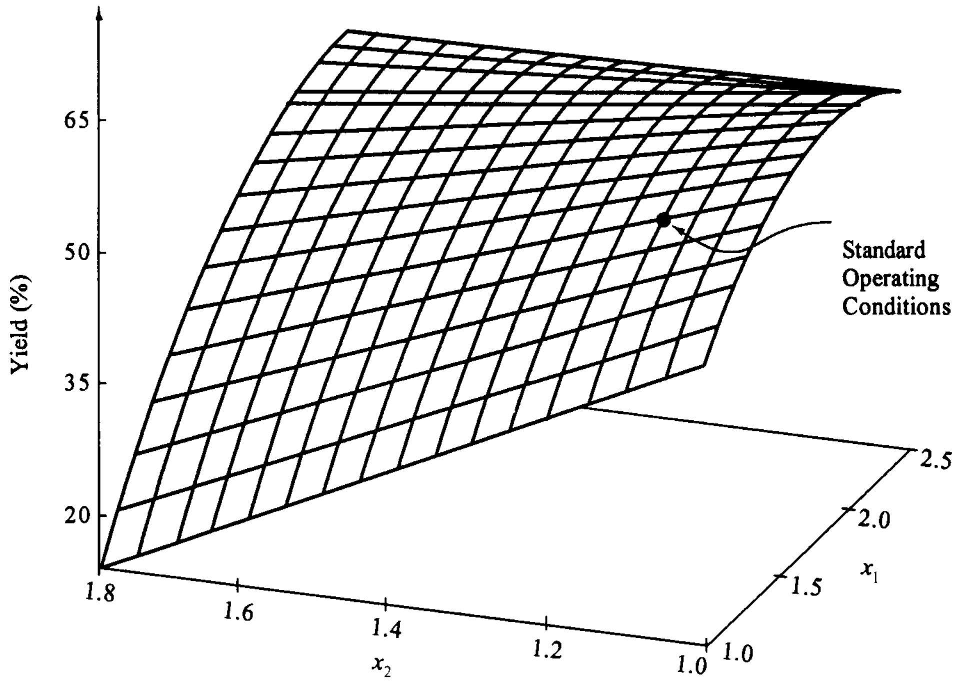

Fig. 5.19. A perspective graph of fitted yield when $x_3 = 1.75$. (From *Statistics for Engineering Problem Solving* by S. B. Vardeman © 1994. Reprinted with permission of Brooks/Cole, a Division of Thomson Learning; www.thomsonlearning.com. Fax 800-730-2215.)

problems. That is, we will touch on some of the ideas of mixture experiments and the role of statistics in mechanistic modeling.

Mixture Experiments

In many situations in industrial chemistry, some important measured property of a product is a function of the proportions in which a set of p ingredients or components are represented in a mixture leading to the product. For example, case studies in the literature have covered subjects ranging from octanes of gasoline blends, discussed by Snee;[21] to strengths of different formulations of ABS pipe compound, treated in Koons and Wilt;[22] to aftertaste intensities of different blends of artificial sweeteners used in an athletic sport drink, discussed by Cornell;[23] to moduli of elasticity of different rocket propellant formulations, considered by Kurotori.[24] For experimenting in such contexts, special statistical techniques are needed. These tools have been discussed at length by Cornell,[25,26] and our purpose here is not to attempt a complete exposition, but only to whet the reader's appetite for further reading in this area.

The goal of mixture experimentation is to quantify how proportions $x_1, x_2, x_3, \ldots, x_p$ of ingredients 1 through p affect a response y. Usually, the hope is to fit some kind of approximate equation involving some parameters $\underset{\sim}{b}$, say

$$y \approx f(x_1, x_2, \ldots, x_p | \underset{\sim}{b})$$

to a set of n data points $(x_1, x_2, \ldots, x_p, y)$, for the purpose of using the fitted equation to guide optimization of y, that is, to find the "best" blend. The logic of data collection and equation fitting is complicated in the mixture scenario by the fact that

$$x_1 + x_2 + \cdots + x_p = 1 \qquad (5\text{-}15)$$

The linear constraint (5-15) means that (p way) factorial experimentation is impossible, and that special measures must be employed in order to use standard regression analysis software to do least squares equation fitting. We will briefly describe in turn some approaches to experimental design, equation fitting, and presentation of results for the

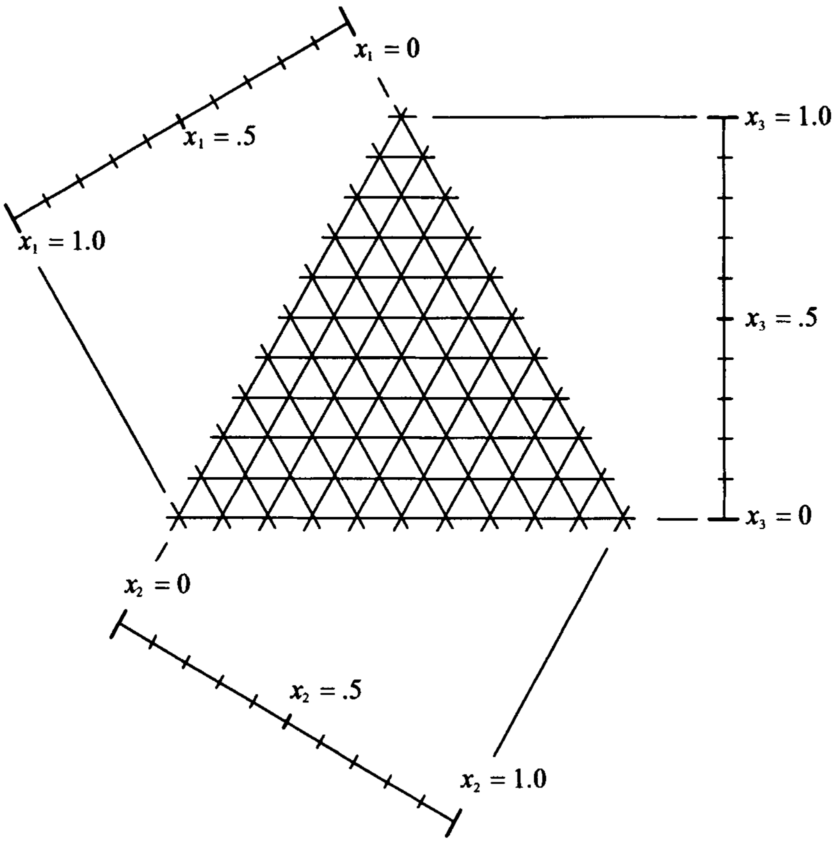

Fig. 5.20. The set of points with $x_1 + x_2 + x_3 = 1$ and a simplex coordinate system. (From *Statistics for Engineering Problem Solving* by S. B. Vardeman © 1994. Reprinted with permission of Brooks/Cole, a Division ofThomson Learning; www.thomsonlearning.com. Fax 800-730-2215.)

mixture problem under its fundamental constraint (5-15).

In the case of $p = 3$ (a three-component mixture problem), the set of all possible combinations of values for x_1, x_2, and x_3 satisfying (5-15) can be conveniently represented as an equilateral triangular region. Figure 5.20 shows such a region and the so-called simplex coordinate system on the region. The corners on the plot stand for cases where the "mixture" involved is actually a single pure component. Points on the line segments bounding the figure represent two-component mixtures, and interior points represent genuine three-component mixtures. For example, the center of the simplex corresponds to a set of conditions where each component makes up exactly one-third of the mixture.

TABLE 5.10 (x_1,x_2,x_3) Points in a Particular $p = 3$ Simplex Lattice Design

x_1	x_2	x_3
1	0	0
0	1	0
0	0	1
$\frac{1}{3}$	$\frac{2}{3}$	0
$\frac{2}{3}$	$\frac{1}{3}$	0
$\frac{1}{3}$	0	$\frac{2}{3}$
$\frac{2}{3}$	0	$\frac{1}{3}$
0	$\frac{1}{3}$	$\frac{2}{3}$
0	$\frac{2}{3}$	$\frac{1}{3}$
$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

One standard mixture (experimental) design strategy is to collect data at the extremes (corners) of the experimental region along with collecting data on a regular grid in the experimental region. Figure 5.21 shows a $p = 3$ example of such a so-called simplex lattice design, and Table 5.10 lists the (x_1,x_2,x_3)

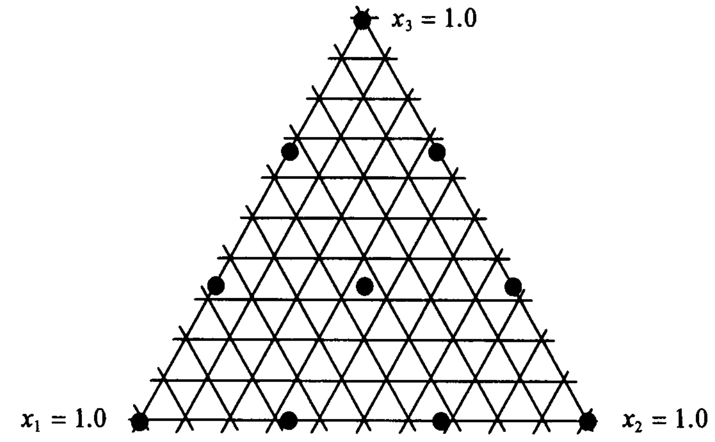

Fig. 5.21. A $p = 3$ simple lattice design.

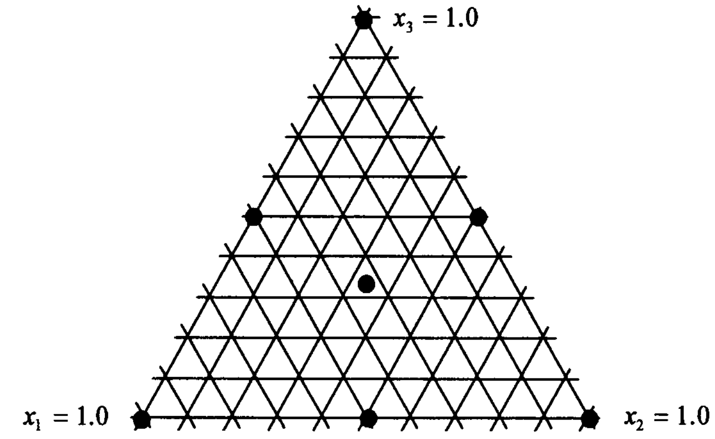

Fig. 5.22. A $p = 3$ simplex centroid design.

points involved. (As in the cases of the data in Tables 5.7–5.9, the order used in the listing in Table 5.10 is not one that would be used in sequencing data collection runs. Instead, a randomly chosen order often is employed.)

Another standard mixture experiment strategy is the so-called simplex centroid design, where data are collected at the extremes of the experimental region and for every equal-parts two-component mixture, every equal-parts three-component mixture, and so on. Figure 5.22 identifies the blends included in a $p = 3$ simplex centroid design.

Often, the space of practically feasible mixtures is smaller than the entire set of x_1, $x_2, \dots, x_p$ satisfying (5-15). For example, in many contexts, "pure" mixtures do not produce viable product. Concrete made using only water and no sand or cement obviously is a useless building product. One common type of constraint on the proportions $x_1, x_2, \dots, x_p$ that produces quite simple experimental regions is that of lower bounds on one or more of the individual proportions. Cornell,[25] for example, discusses a situation where the effectiveness in grease stain removal of a $p = 3$ bleach mixture was studied. Past experience with the product indicated that the proportions by weight of bromine, x_1, of powder, x_2, and of HCl, x_3, needed to satisfy the constraints:

$$x_1 \geqslant 0.30, x_2 \geqslant 0.25, \text{ and } x_3 \geqslant 0.02 \qquad (5\text{-}16)$$

for effective action of the product (i.e., the mixture needed to be at least 30% bromine, at least 25% powder, and at least 2% HCl by weight.)

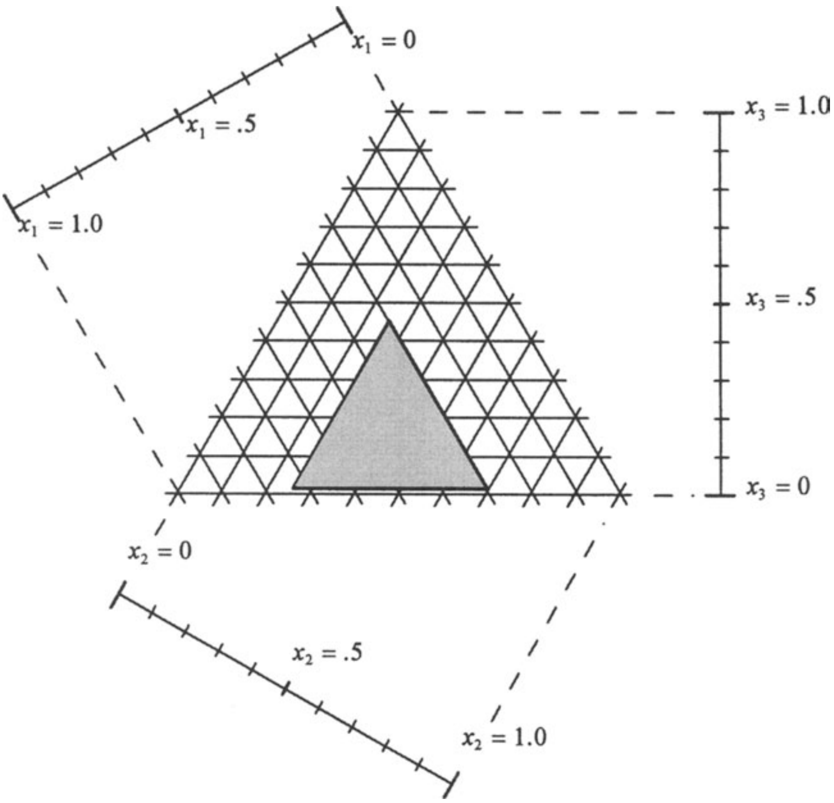

Fig. 5.23. The $p = 3$ simplex and a set of feasible bleach mixtures.

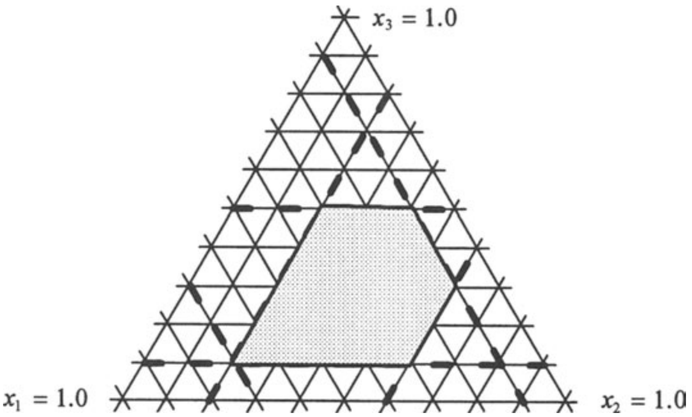

Fig. 5.24. An irregularly shaped experimental region in a $p = 3$ mixture study.

The effect of adding the lower bound constraints (5-16) to the basic mixture constraint (5-15) can be pictured as in Fig. 5.23. There, a triangular subregion of the basic $p = 3$ simplex depicts the feasible (x_1,x_2,x_3) points. The choice of experimental mixtures for such an experimental region can be made by direct analogy to or rescaling of designs such as the simplex lattice and simplex centroid designs illustrated above to cover the entire simplex. (It is common to refer to the rescaling process as the use of pseudo-components.)

Constraint systems more complicated than simple lower bounds produce irregularly shaped experimental regions and less obvious methods of choosing $(x_1, x_2, \ldots, x_p)$ points to cover the experimental region. When $p = 3$, it is possible to sketch the region of feasible points on a simplex plot and use it to help guide the choice of mixture experiment strategy. Figure 5.24 illustrates the kind of region that can arise with other than exclusively lower bound constraints.

When more than three components are involved in a mixture study, such plots are, of course, no longer possible, and other more analytic methods of identifying candidate experimental mixtures have been developed. For example, McLean and Anderson[27] presented an algorithm for locating the vertices of an experimental region defined by the basic constraint (5-15) and any combination of upper and or lower bound constraints

$$0 \leqslant a_i \leqslant x_i \leqslant b_i \leqslant 1$$

on the proportions x_i. Cornell[25,26] discusses a variety of algorithms for choosing good mixture experiment designs under constraints, and many of the existing algorithms for the problem have been implemented in the MIXSOFT software package developed by Piepel.[28]

Empirical polynomial descriptions of (approximately) how a response y depends upon proportions x_1, x_2, ..., x_p are popular mixture analysis tools. The process of fitting polynomials to mixture experiment data in principle uses the same least squares notion illustrated in the fitting of a parabola to the data of Table 5.2. However, the mechanics of using standard multiple regression analysis software in the mixture context is complicated somewhat by the basic constraint (5-15). For example, in view of (5-15) the basic ($p + 1$ parameter) linear relationship

$$y \approx b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p \quad (5\text{-}17)$$

is in some sense "overparameterized" in the mixture context, in that it is equivalent to the (p parameter) relationship

$$y \approx b_1x_1 + b_2x_2 + \cdots + b_px_p \quad (5\text{-}18)$$

if one identifies the coefficients in (5-18) with the sums of the corresponding coefficients in (5-17) and the coefficient b_0. As a result, it is the "no intercept" relationship (5-18) that is typically fit to mixture data when a linear relationship is used. In a similar way, when a second-order or (multivariable) quadratic relationship between the individual proportions and the response variable is used, it has no intercept term and no pure quadratic terms. For example, in the $p = 3$ component mixture case, the general quadratic relationship typically fit to mixture data is

$$\begin{aligned} y \approx b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1x_2 \\ + b_5x_1x_3 + b_6x_2x_3 \end{aligned} \quad (5\text{-}19)$$

(Any apparently more general relationship involving an intercept term and pure quadratic terms can by use of (5-15) be shown to be equivalent to (5-19) in the mixture context.) Relationships of the type of (5-19) are often called Scheffé models, after the first author to treat them in the statistical literature. Other more complicated equation forms are also useful in some applications, but we will not present them in this chapter. The interested reader is again referred to Cornell[25,26] for more information on forms that have been found to be tractable and effective.

We should point out that the ability to fit equations of the form (5-18) or like (5-19), or of an even more complicated form, is predicated on having data from enough different mixtures to allow unambiguous identification of the parameters $\underset{\sim}{b}$. This requires proper data collection strategy. Much of the existing statistical research on the topic of mixture experiment design has to do with the question of wise allocation of experimental resources under the assumption that a particular type of equation is to be fit.

One's understanding of fitted polynomial (and other) relationships often is enhanced through the use of contour plots made on coordinate systems such as that in Fig. 5.25. (This is even true for $p \geqslant 3$ component mixture scenarios, but the use of the idea is most transparent in the three-component case.) A plot like Fig. 5.25 can be a powerful tool to aid one in understanding the nature of a fitted equation

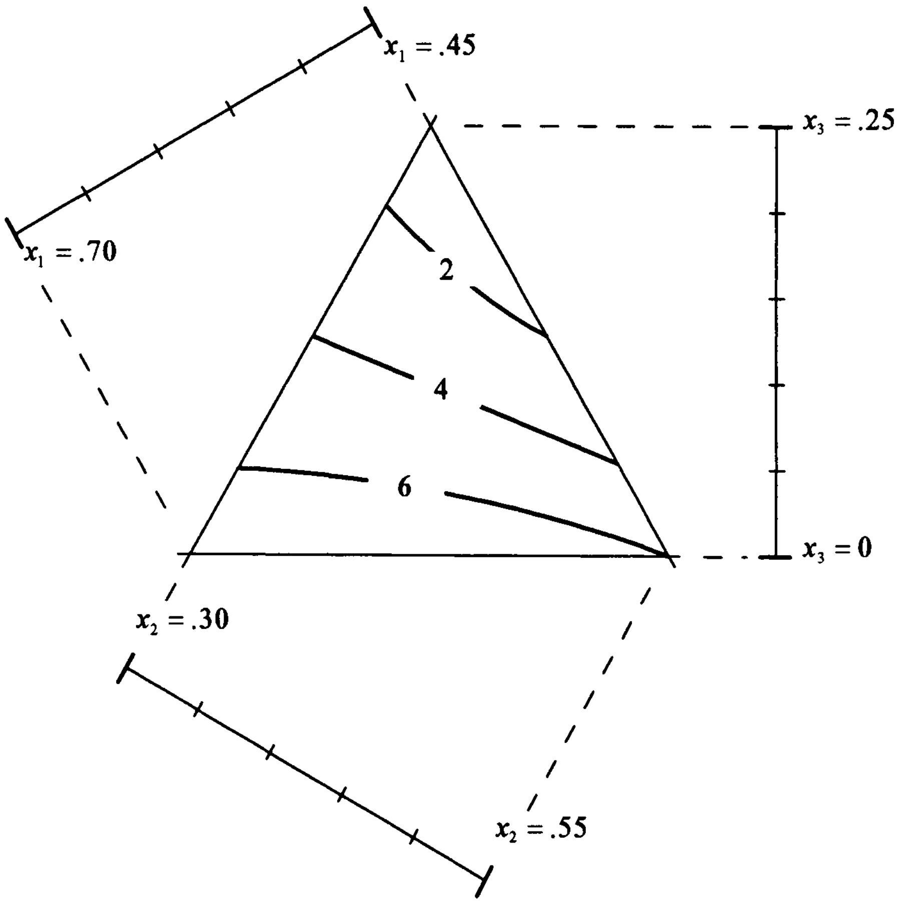

Fig. 5.25. A contour plot made on the $p = 3$ simplex. (From *Statistics for Engineering Problem Solving* by S. B. Vardeman © 1994. Reprinted with permission of Brooks/Cole, a Division of Thomson Learning: www.thomsonlearning.com. Fax 800-730-2215.)

and finding regions of optimum fitted response.

The mixture experiment counterpart to conventional screening/fractional factorial experimentation also is possible. So-called axial designs have been developed for the purpose of providing screening-type mixture data for use in rough evaluation of the relative effects of a large number of mixture components on a response variable. The same kind of sequential experimental strategy illustrated in the process improvement example is applicable in mixture contexts as well as contexts free of a constraint such as (5-15).

Mechanistic Model Building

The kinds of equations most easily fit to multi-factor data using standard (least squares) regression analysis techniques are polynomial equations such as (5-6), (5-14), (5-18), and (5-19). These are particularly convenient because they are linear in their parameters, b. But they are probably best thought of as empirical "mathematical French curve" descriptions of the relation of a response, y, to the explanatory variables, x. Polynomial equations function as extremely useful summaries of observed patterns in one's data, but they do not typically provide direct insight into chemical mechanisms that produce those patterns, and the fitted parameters, b, do not often have direct physical meanings. Their use is particularly appropriate where there is little a priori knowledge of mechanisms involved in a process that might aid in its description, and/or no such knowledge is really essential to achieving one's goals.

Sometimes, however, it is desirable (on the basis of possible reaction kinetics or for other reasons) to posit theoretical descriptions of process outputs in terms of explanatory variables. That is, physicochemical principles often lead (through differential or integral equation descriptions of a system) to equation forms for a response that, like

$$y = \frac{K_1 K_A K_B P_A P_B}{(1 + K_A P_A + K_B P_B)^2}$$

$$y = C_0 \exp(-Kt)$$

and

$$y = \frac{K_1 x}{1 + K_2 x}$$

are nonlinear in the parameters. Although such equations or models may be less tractable than empirical polynomial equations, the parameters involved more often than not *do have* direct physical interpretations. Further, when such a model can be verified as being an adequate description of a process (thereby confirming scientific understanding) and the parameters involved are estimated from process data, such mechanistic models can provide much safer extrapolations beyond an experimental region than the cruder empirical polynomial models.

The process of research in chemical systems is one of developing and testing different models for process behavior. Whether empirical or mechanistic models are involved, the discipline of statistics provides data-based tools for discrimination between competing possible models, parameter estimation, and model verification for use in this enterprise. In the case where empirical models are used, techniques associated with "linear" regression (linear least squares) are used, whereas in mechanistic modeling contexts "nonlinear" regression (nonlinear least squares) techniques most often are needed. In either case, the statistical tools are applied most fruitfully in iterative strategies.

Reilly and Blau[29] and Chapter 16 of Box et al.[19] provide introductions to the general philosophy of using statistical methods in mechanistic modeling contexts, as well as a number of useful references for further reading.

Fairly sophisticated and specialized statistical software is needed in the practical application of nonlinear regression methods to mechanistic modeling for industrial chemistry applications. The techniques implemented in such software are discussed in Seber and Wild,[32] Bates and Watts,[30] Bard,[31] and Riley and Blau.[29]

MODERN BUSINESS PROCESS IMPROVEMENT AND THE DISCIPLINE OF STATISTICS

The modern global business environment is fiercely competitive in all sectors, including the chemical sector. It is by now widely recognized that corporate survival in this environment depends upon constant improvement in *all* business processes, from billing to production. Companies have adopted a variety of programs and focuses aimed at facilitating that improvement. A decade ago, efforts organized around a *Total Quality Management* banner (with liberal references to emphases of consultants like W. E. Deming, J. M. Juran, and A. Feigenbaum) were popular. More recently, programs keyed to ISO 9000[33] certification criteria and Malcolm Baldridge Award[34] criteria have become prominent. And currently probably the most visible programs are the so-called *Six Sigma* programs.

In one sense there is nothing new under the sun, and all successful business process improvement programs (including those in the chemical sector) must in the end reduce to organized problem-solving disciplines. So it is not surprising that programs quite different in name are often very alike in fundamental content. And as they must necessarily make use of empirical information (data), they must have significant statistical components. To make this connection to statistics slightly more explicit, we proceed to provide a few additional details on the Six Sigma movement. (Further material on the subject is easy to find using an Internet search engine, as there are many consultants eager to sell their advice and Six Sigma training. The American

Society for Quality at www.asq.org offers many entries into the subject. And a search at amazon.com for "Six Sigma" books already produced 6666 hits in May 2004. Fashions change quickly enough that it seems pointless to provide more detailed recommendations for follow up on the subject.)

The phrase "Six Sigma" originated at Motorola in the late 1980s. Programs there and at General Electric in the mid-1990s are widely touted as important contributors to company profits and growth in stock values. The name is now commonly used in at least three different ways. "Six Sigma" refers to

- a goal for business process performance
- a discipline for improvement to achieve that performance
- a corporate program of organization, training, and recognition conceived to support the process improvement discipline

As a goal for business process improvement, "Six Sigma" is equivalent to "$C_{\mathrm{pk}} = 2$." What is perhaps confusing to the uninitiated is that this goal has connections (through normal distribution tail area calculations) to small ("parts per million") fractions defective relative to two-sided specifications on y. Six Sigma proponents often move between the "small process variation" and "parts per million" understandings with little warning.

Six Sigma process improvement disciplines are typically organized around the acronym "**MAIC**." The first step in a MAIC cycle is a **M**easure step, wherein one finds appropriate process responses to observe, identifies and validates measurement systems and collects baseline process performance (process monitoring) data. The second step is an **A**nalyze step. This involves summarizing the initial process data and drawing appropriate inferences about current process performance. The third step in a MAIC cycle is an **I**mprove step, where process knowledge, experimentation, and more data analysis are employed to find a way to make things work better. Finally, the four-step cycle culminates in a **C**ontrol (process monitoring) effort. The object here is to see that the newly improved performance is maintained after a project team moves on to other problems.

Six Sigma corporate organization, training, and recognition programs borrow from the jargon and culture of the martial arts. People expert in the process improvement paradigm are designated "black belts," "master black belts," and so on. These people lead project teams and help train new initiates ("green belts") in the basics of the program and its tools (including statistical tools). The emphasis throughout is on completing projects with verifiable large dollar impact.

Having made the point that improvement in all business activities is of necessity data-driven, it is hopefully obvious that the emphases and methods of the subject of statistics are useful beyond the lab and even production. Of course, for broad implementation, it is the most elementary of statistical methods that are relevant.

CONCLUSION

We have tried in this chapter to give readers the flavor of modern applied statistical methods and to illustrate their usefulness in the chemical industry. Details of their implementation have of necessity been reserved for further more specialized reading, for which the interested reader is encouraged to consult the references given in this chapter.

REFERENCES

1. Albin, S., "The Lognormal Distribution for Modeling Quality Data When the Mean is Near Zero," *J. Qual. Technol.*, **22**, 105–110 (1990).
2. Chambers, J., Cleveland, W., Kleiner, B., and Tukey, P., *Graphical Methods of Data Analysis*, Duxbury, Boston, MA, 1983.
3. Odioso, R., Henke, A., Stauffer, H., and Frech, K., "Direct Hydration of Olefins," *Ind. Eng. Chem.*, **53**(3), 209–211 (1961).
4. Devore, J., *Probability and Statistics for Engineering and the Sciences* (3rd ed.), Brooks/Cole, Pacific Grove, CA, 1991.

5. Vardeman, S. B., and Jobe, J. M., *Basic Engineering Data Collection andAnalysis*, Duxbury/Thomson Learning, Pacific Grove, CA, 2001.
6. Vardeman, S. B., *Statistics for Engineering Problem Solving*, PWS-Kent, Boston, MA, 1994.
7. Neter, J., Kutner, M., Nachsteim, C., and Wasserman, W., *Applied Linear Statistical Models* (4th ed.), McGraw-Hill, New York, 1996.
8. Hahn, G., and Meeker, W., *Statistical Intervals: A Guide for Practitioners*, Wiley, New York, 1991.
9. Wadsworth, H., Stephens, K., and Godfrey, B., *Modern Statistical Methods for Quality Control and Improvement*, Wiley, New York, 1986.
10. Burr, I., *Statistical Quality Control Methods*, Dekker, New York, 1976.
11. Duncan, A., *Quality Control and Industrial Statistics* (5th ed.), Irwin, Homewood, IL, 1986.
12. Grant, E., and Leavenworth, R., *Statistical Quality Control* (7th ed.), McGraw-Hill, New York, 1996.
13. Ott, E., and Schilling, E., *Process Quality Control*, McGraw-Hill, NY, 1990.
14. Vardeman, S., and Jobe, J. M., *Statistical Quality Assurance Methods for Engineers*, Wiley, New York, 1999.
15. Vander Wiel, S., Tucker, W., Faltin, F., and Doganaksoy, N., "Algorithmic Statistical Process Control: Concepts and an Application," *Technometrics*, **34**(3), 286–297 (1992).
16. Wernimont, G., "Statistical Quality Control in the Chemical Laboratory," *Qual. Eng.*, **2**, 59–72 (1989).
17. Hicks, C., and Turner, K., *Fundamental Concepts in the Design of Experiments* (5th ed.), Oxford University Press, New York, 1999.
18. Mason, R., Gunst, R., and Hess, J., *Statistical Design and Analysis of Experiments*, Wiley, New York, 1989.
19. Box, G., Hunter, W., and Hunter, J. S., *Statistics for Experimenters*, Wiley, New York, 1978.
20. Box, G., "George's Column," *Qual. Eng.*, **2**, 497–502 (1990).
21. Snee, R., "Developing Blending Models for Gasoline and Other Mixtures," *Technometrics*, **23**, 119–130 (1981).
22. Koons, G., and Wilt, M., "Design and Analysis of an ABS Pipe Compound Experiment," in *Experiments in Industry: Design, Analysis and Interpretation of Results*, R. Snee, L. Hare, and J. R. Trout (eds.), pp. 111–117, ASQC Quality Press, Milwaukee, WI, 1985.
23. Cornell, J., "A Comparison between Two Ten-Point Designs for Studying Three-Component Mixture Systems," *J. Qual. Technol.*, **18**, 1–15 (1986).
24. Kurotori, I., "Experiments with Mixtures of Components Having Lower Bounds," *Ind. Qual. Control*, **22**, 592–596 (1966).
25. Cornell, J., *How to Run Mixture Experiments for Product Quality*, Vol. 5 in the ASCQ "Basic References in Quality Control" series, American Society for Quality Control, Milwaukee, WI, 1983.
26. Cornell, J., *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data* (2nd ed.), Wiley, New York, 1990.
27. MacLean, R., and Anderson, V., "Extreme Vertices Design of Mixture Experiments," *Technometrics*, **8**, 447–454 (1966).
28. Piepel, G., *Mixsoft and Misoft User's Guide, Version 1.0*, Mixsoft-Mixture Experiment Software, Richland, WA, 1989.
29. Reilly, P., and Blau, G., "The Use of Statistical Methods to Build Mathematical Models of Chemical Reacting Systems," *Can. J. Chem. Eng.*, **52**, 289–299 (1974).
30. Bates, D., and Watts, D., *Nonlinear Regression Analysis and Its Applications*, Wiley, New York, 1988.
31. Bard, Yonathan, *Nonlinear Parameter Estimation*, Academic Press, New York, 1974.
32. Seber, G., and Wild, C., *Nonlinear Regression*, Wiley, New York, 1989.
33. International Organization for Standardization (www.iso.ch).
34. National Institute of Standards and Technology (www.quality.nist.gov).