SECTION 44
BASIC STATISTICAL METHODS

Edward J. Dudewicz
Department of Mathematics,
Syracuse University, Syracuse, New York

THE STATISTICAL TOOL KIT

Most decision making in quality control, as in most other areas of modern human endeavor (e.g., evaluation of new medical treatments and scanning machines, planning of scientific polling, and marketing and investment strategies, to name a few), rests on a base of *statistics*—defined narrowly as the collection, analysis, and interpretation of data or, more broadly, as "the science of decision making under uncertainty." For the practitioner, *statistics* can be thought of as a kit of tools that helps to solve problems. The statistical tool kit shown in Table 44.1 lists problems to be solved, applicable statistical tools, and where in this handbook the tool is to be found.

Examples of actual practice will be used as extensively as the space allocated to it allows to provide the reader with both a model for solution and a data set with correct analysis that can be used to verify the accuracy of local computer software. Annotated computer program output from such packages as SAS and BMDP will be used in many of these examples.

In addition to the basic statistical methods discussed in this section, four other sections cover specific areas—Section 45, Statistical Process Control; Section 46, Acceptance Sampling; Section 47, Design and Analysis of Experiments; and Section 48, Reliability Data Analysis. Many other sections include additional applications. Also, Appendix III, Selected Quality Standards, Specifications, and Related Documents, includes documents on statistical techniques and procedures.

SOURCES AND SUMMARIZATION OF DATA

The source of a set of data that we desire to analyze to solve a problem is a very important consideration. The two sources we will address and that are the most common are historical data and data from planned experimentation. Investigators using historical data are like blind people probing an elephant, for reasons discussed below under Historical Data, Their Uses, and Caveats, and under Data from Planned Experimentation. Note that all data need careful review, as discussed under Data Screening, below.

Planning for Collection and Analysis of Data. The tools cited in Table 44.1 must be used in an effective manner to yield a return appropriate for the cost of using them. To achieve this return, it is not sufficient to plug numbers into formulas. The full process must include careful planning of data collection, analysis of the data to draw statistical conclusions, and making the transition to answer the original technical problem. A checklist of some of the key steps in achieving this is as follows:

1. Collect sufficient background information to translate the engineering problem statement into a specific statement that can be evaluated by statistical methods.
2. Plan the collection of data.
   - ***a.*** Determine the type of data needed. Variables data (readings on a scale of measurement) may be more expensive than attributes data (go or no-go data), but the information is much more useful.
   - ***b.*** Determine if any past data are available that are applicable to the present problem; however, bear in mind the hazards of historical data sets.
   - ***c.*** If the problem requires an evaluation of several alternative decisions, obtain information on the economic consequences of a wrong decision.
   - ***d.*** If the problem requires the estimation of a parameter, define the precision needed for the estimate.
   - ***e.*** Determine if the error of measurement is large enough to influence the sample size or the method of data analysis; laboratory error often can dwarf experimental variability.
   - ***f.*** Define the assumptions needed to calculate the required sample size.
   - ***g.*** Calculate the required sample size considering the desired precision of the result, statistical risk, variability of the data, measurement error, and other factors.

TABLE 44.1 The Statistical Tool Kit

Problem	Statistical tool	Reference pages or sections
Planning a statistical investigation	Planning and analyzing data for solving specific problems	44.2–44.4
Summarizing data	Frequency distributions, histograms, and indices	44.7–44.17
Predicting future results from a sample	Probability distributions	44.23–44.41
Determining a probability involving several events	Basic theorems of probability	44.17–44.23
Determining the significance of difference between two sets of data or between a set of data and a standard value	Tests of hypotheses	44.58–44.81
Determining the sample size required for testing a hypothesis	Sample size determination for hypothesis testing	44.78–44.79
Determining the ability of a sample result to estimate a true value	Confidence limits	44.41–44.46
Determining the sample size required to estimate a true value	Sample size determination for estimation	44.45–44.46
Determining tolerance limits on single characteristics	Statistical tolerance limits	44.47–44.51
Determining tolerance limits for interacting dimensions	Tolerance limits for interacting dimensions	44.50–44.54
Incorporating past information in predicting future events	Bayes' Theorem	44.21–44.22 44.54–44.58
Incorporating economic consequences in defining decision rules	Statistical decision theory	44.22
Converting data to meet statistical assumptions	Transformations of data	44.81–44.84
Predicting system performance	Monte Carlo sampling methods	44.84–44.85
Determining group membership	Clustering and discrimination	44.86–44.87
Determining which is the best	Selection of the best	44.87–44.88
Evaluating the relationship between two or more variables by determining an equation to estimate one variable from knowledge of the other variables	Regression analysis	44.88–44.108
Controlling process quality by early detection of process changes:		
1. Using measurements data	Variables control charts	Sec. 45
2. Using attributes data	Attributes control charts	Sec. 45
Evaluating quality of lots to a previously defined quality level		
1. Quality measured on an attributes basis	Attributes sampling plans	Sec. 46
2. Quality measured on a variables basis	Variables sampling plans	Sec. 46
3. Sampling to determine reliability	Reliability sampling plans	Sec. 46
4. Bulk product	Bulk sampling plans	Sec. 46
Planning and analyzing experiments:		
1. Investigating the effect of varying one factor	One-factor experiment	Sec. 47
2. Investigating the effect of varying two or more factors	Designs for two or more factors	Sec. 47
3. Investigating the variability of laboratory measurements	Interlaboratory tests	Sec. 47
4. Experimenting under process conditions to determine optimum settings of variables	Evolutionary operation (EVOP)	Sec. 47
5. Determining the optimum set of values of a group of variables that affect a response variable	Response surface methodology (RSM)	Sec. 47
Predicting performance without failure (reliability)	Reliability prediction and analysis	Sec. 48

h. Define any requirements for preserving the order of measurements when time is a key parameter.
i. Determine any requirements for collecting data in groups defined so as to reflect the different conditions that are to be evaluated.
j. Define the method of data analysis and any assumptions required.
k. Define requirements for any computer programs that will be needed.

3. Collect the data.
a. Use methods to ensure that the sample is selected in a random manner.
b. Record the data and also all conditions present at the time of each observation.
c. Examine the sample data to ensure that the process shows sufficient stability to make predictions valid for the future.

4. Analyze the data.
a. Screen the data.
b. Evaluate the assumptions previously stated for determining the sample size and for analyzing the data. Take corrective steps (including additional observations) if required.
c. Apply statistical techniques to evaluate the original problem.
d. Determine if further data and analysis are needed.
e. Conduct sensitivity analyses by varying key sample estimates and other factors in the analysis and noting the effect on final conclusions.

5. Review the conclusions of the data analysis to determine if the original technical problem has been evaluated or if it has been changed to fit the statistical methods.

6. Present the results.
a. Write a report, including an executive summary.
b. State the conclusions in meaningful form by emphasizing results in terms of the original problem rather than the statistical indices used in the analysis.
c. Present the results in graphic form where appropriate. Use simple statistical methods in the body of the report, and place complicated analyses in an appendix.

7. Determine if the conclusions of the specific problem apply to other problems or if the data and calculations could be a useful input to other problems.

Historical Data, Their Uses, and Caveats. *Historical data* are data that we already have and which may seem to be relevant to a question or problem that has arisen. Such data are sometimes also called *existing data sets.* Often data are saved during the production process, for example. If a satisfactory process goes out of control after some years of operation, it is often suggested that it would save both time and expense to statistically analyze the historical data rather than perform a planned experiment to obtain data that could lead to process correction. Thus we have available data that may consist of measurements Y (such as a process yield, e.g., the strength of a material produced) and associated process variables $x_1, x_2, \ldots, x_k$ (such as x_1 = pressure and x_2 = acid concentration, with $k = 2$).

This situation is extremely different from that where experiments have been run at each of a number of settings of $x_1, \ldots, x_k$ that were selected in advance by statistical design criteria, and often little can be learned from such data even with the most thorough statistical analysis. Some of the reasons for this are

The x's may be highly correlated with each other; hence it may not be possible to separate an effect as due to (for example) x_1 or x_2.

The x's may have been manipulated to try to control the output Y of the process (some of them perhaps even in directions that move the output in directions that are not desired), hence giving spurious indications of directions of effects when analyzed.

The x's may cover a very small part of the possible operating range, so small that any indications of changes in Y attributable to changes in the x's may be overwhelmed by the size of the variability of the process (measured by "standard deviation," discussed below).

Other variables that affect the output of the process (e.g., time of day, atmospheric conditions, operator running the process, etc.) may not have been held constant and may in fact be the real

causes of changes observed in the process (while an analysis conducted based only on the x's may erroneously yield a model that has no basis in reality).

For these and other reasons, much more information generally can be obtained from a carefully designed experiment than from extensive analysis of historical data sets collected in uncontrolled circumstances. The best one usually can hope for from such historical data set analysis is an indication of the most important variables to include in the designed experiment.

As an example of one of the failings described above, suppose that a historical data set consists of a yield Y at each of the five (x_1, x_2) pairs given in Table 44.2. We see that there is what appears to be a good spread on the x_1 (acid concentration) values, from 80 to 110, and also a good spread on the x_2 (pressure) values, from 105 to 144. However, a plot of the data points (x_1, x_2), as shown in Figure 44.1, shows that in fact all five of the data points lie on a straight line. Thus there is no hope of any analysis of these data telling us whether any effect we see is due to x_1 or to x_2. The points do not "cover" the space of x_1 between 80 and 110 and x_2 between 105 and 144 well at all. Also, to solve the production problem, it may be desirable to explore outside this space of historical operation, and this is not allowed for in the historical data set.

A problem such as that shown in Figure 44.1 is called *multicollinearity* of the data points, i.e., of the sets of $(x_1 \ldots, x_k)$ that we have available for analysis. With $k = 2$, such a problem is easy to detect by a graph such as that in Figure 44.1, and such a graph should always be made. With $k = 3$ it may be possible to detect such a situation graphically using computer graphics packages. Often, however, k is much larger than 2 or 3, and then statistical analysis is needed to detect multicollinearity. Near multicollinearity (i.e., when the points almost fall on a line when $k = 2$, on a plane when $k = 3$, or on what is called a $k - 1$ dimensional hyperplane when k is larger than 3) is just as much of a problem and is much harder to detect. For an extensive bibliography and comments on computer routines for this problem, see Hoerl and Kennard (1981). For some more recent results, see Huh and Olkin (1995), where numerical illustrations using the classic data of Longley (1967) are included.

Data from Planned Experimentation. *Data from planned experimentation* are data gathered in an attempt to study a problem that has arisen or is contemplated. Such data are gathered at various settings of the variables felt to be of importance $(x_1,\ldots, x_k)$, while holding constant (if possible, and recording the values of in any case) all other variables that could conceivably have an effect on the output $(x_{k+1},\ldots, x_m)$—for example, atmospheric pressure may not be able to be controlled in most circumstances but should be monitored and recorded in every experiment if it is felt beforehand that it might have an effect on the output. (If it is irrelevant, it can be disregarded later; if it is relevant but not recorded, we will not be able to detect that relevance.) Details of designs to use, i.e., how to choose the values of $(x_1,\ldots, x_k)$ for the experiments once the variables to be varied have been chosen, are considered in Section 47, Design and Analysis of Experiments, and in Dudewicz and Karian (1985).

Data Screening. Once the data specified in the preceding checklist for collection and analysis of data, step 2, has been collected (step 3), the first step in its analysis (step 4*a*) is to screen the data.

TABLE 44.2 Historical Data Set

Data point	x_1 (Acid concentration)	x_2 (Pressure)
1	100	131.0
2	90	118.0
3	105	137.5
4	110	144.0
5	80	105.0

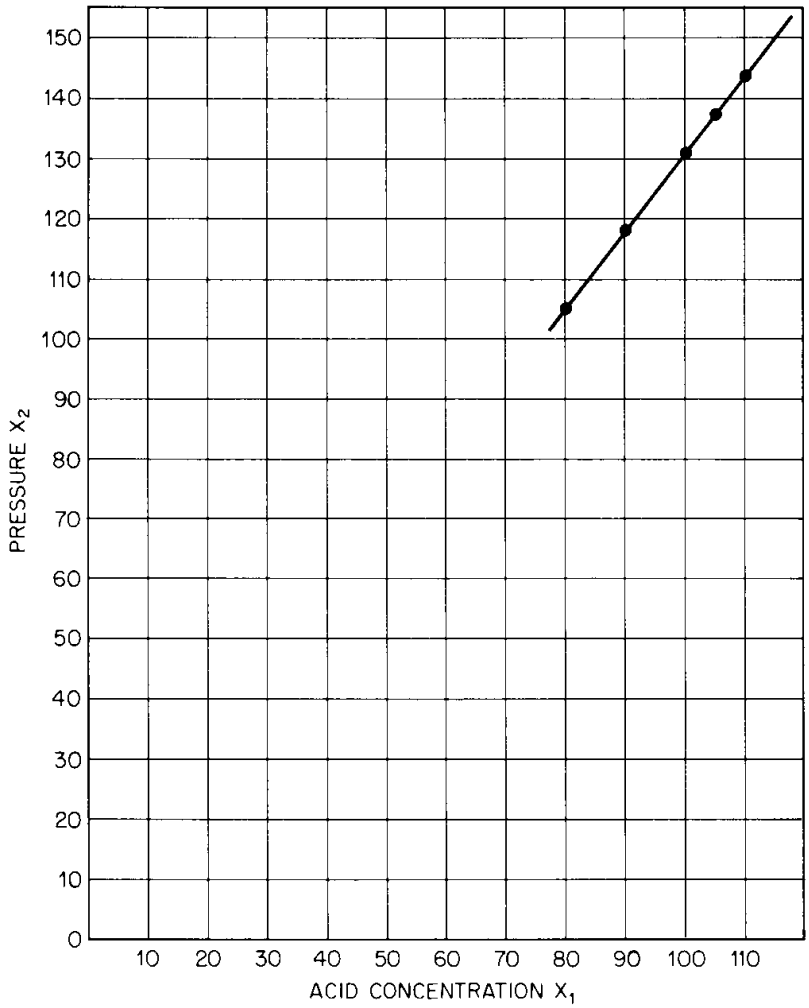


FIGURE 44.1 Plot of (x_1, x_2) in a historical data set. Five data points have been plotted, and a straight line has been drawn through them.

We will now discuss the need for this critical (and often omitted) step, some common methods of screening, and a very commonly used method that has some great dangers.

The Need for Data Screening. A data set that contains no instances of incorrectly transcribed values, contains no values that are technically correct but where the experiment went awry for some reason (such as equipment malfunction), and where the basic model does not change its form over the region of experimentation is called a *clean* data set. Such data sets are commonly expected to be obtained by experimenters who exercise care in their experimental conduct and recording. However, contrary to this expectation of the experimenters, few statisticians have ever seen a clean data set (despite many years of studying many data sets arising in many areas). It follows, then, that all data sets need, as a first stage of analysis, to be examined for values that may cause invalid inferences to be made if those values are left in the data set. Procedures for performing this examination are called *data screening methods.* Among the most powerful such methods are those which are used on the results of a regression analysis (studied later in this section).

There are also a number of methods that can and should be used at the outset, before any regression analyses are performed, with the goal of detecting *outliers,* that is, observations (or groups of observations) that deviate markedly from the other available data. Numerous tests are available for detecting outliers [see Sheesley (1977) for some of these]. One simple rule calls an observation an outlier if it lies 2.5, 3, or 4 standard deviations or further from the mean (Draper and Smith 1981). This is discussed further below.

Methods of Data Screening. One of the most common methods of data screening is to classify observations as outliers if they are outside an interval of L multiples of the standard deviation about the mean. (Standard deviation is discussed below under Sample Characteristics.) The number L is commonly taken to be 2.5, 3, or 4. The larger L is, the less likely it is that outliers will be detected, while the smaller L is, the more good observations one will wrongly detect as potential outliers. For example, from Table B in Appendix II, we see that if $L = 3$, then (100) (.0027) = 0.27 percent of the observations will be further than 3 standard deviations from the mean even if there are no outliers in the data set; this assumes a normal distribution for the observations. Thus, if one uses $L = 3$, one expects to find about 3 possible outliers in a data set of 1000 data points (since 0.27 percent of 1000 is 2.7, i.e., roughly 3). As the data set being considered becomes larger, the more possible outliers one will identify even if there are no problems with the data (which is quite unlikely). For this reason,

> Outliers should be deleted from the analysis only if they can be traced to specific causes (such as recording errors, experimental errors, and the like).
>
> Typically, one takes L to depend on the size of the data set to be screened; with $n = 1000$ points, $L = 3$ is reasonable; with $n = 100$, $L = 2.5$ can be used and only (100) (0.0124) = 1.24 outliers will be expected to be found if the data have no problems.

After bad data are deleted or replaced (this is desirable if the experiment can be rerun under comparable conditions to those specified in the experimental plan), the data should be screened again: With the "worst" points removed/corrected, less extreme cases may come to be identified as possible outliers. Another commonly used method, that of crossplots, is discussed below.

Crossplots and Their Hazards. In *crossplots* (also called *scatter plots*), one simply plots each pair of variables in the data set on a set of axes. For example, in the example of Table 44.2, one would plot Y versus x_1, Y versus x_2, and x_1 versus x_2. The last of these plots was given in Figure 44.1 (and showed some problems with the data that have already been discussed). Note that when using this technique, one uses all variables that were measured (and not just the variables that are thought to be of primary interest; see Nelson 1979, Section 4).

Such plots must be used with great caution. While points that seem odd (e.g., away from the majority of the data points) should be subjected to examination to see if there are problems with them, one should not use such plots to conclude relationships of the yield with x_1 and/or x_2. For example, suppose one has the data set of Table 44.3. Then from crossplots of Y versus x_1 and Y versus x_2 (see Figure 44.2), one would be tempted to conclude that Y is a decreasing function of x_1 and an increasing function of x_2 (and so, in attempting to maximize yield, might set x_1 as low as possible and x_2 as high as possible). However, as can easily be verified, the data in Table 44.3 came exactly from the relationship (model) $Y = 10+x_1+2x_2$. Thus, in fact, Y is an increasing function of x_1 (not a decreasing function as the crossplot had suggested). Daniel (1977) has suggested that the data set of Table 44.4 may give additional insight here. Crossplots for that data set are given in Figure 44.3. Thus one sees that even with a strong true relationship (here $Y = 10 + x_1 + 2x_2$ is used again), the crossplots may yield no insight at all. Even worse, clearly by choosing the points (x_1, x_2) one may give the Y versus x_1 relationship any slope (negative, as with Table 44.3 and Figure 44.2; zero; or positive), or the relationship can be smoothly curved in any direction or degree of complexity. (Note that the regression methods given later in this section would not be fooled by the relationships in the data of Table 44.4. They would give the true relationships.) Thus crossplots are useful for detection of possible outliers; however, they are not a substitute for regression and can easily be misused.

Descriptive Statistics for Summarizing Data. Many of the most practical methods of summarizing data are quite simple in concept. Depending on the goals of the data summarization, sometimes one method will provide a useful and complete summarization. More often, two or more methods will be used to attain the clarity of description that is desired. Several key methods are plots versus time order of data, frequency distributions and histograms, sample characteristics (mean, median, mode, variance, standard deviation, and percentiles), measures of central tendency/location, and measures of dispersion.

TABLE 44.3 Data Set *A* for Crossplots Example

Data point	x_1	x_2	Y
1	−2	1	10
2	−1	−2	5
3	0	−5	0

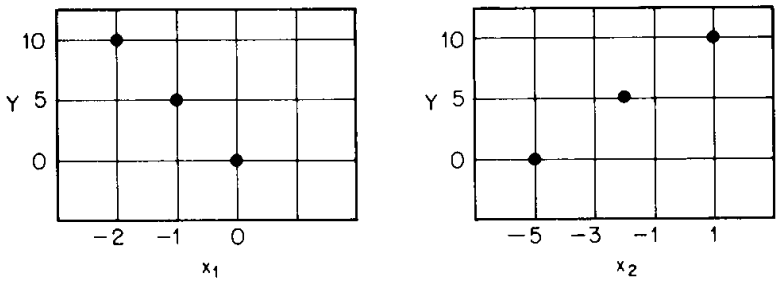


FIGURE 44.2 Crossplots of Y versus x_1 and of Y versus x_2 for data set *A* of Table 44.3.

TABLE 44.4 Data Set *B* for Crossplots Example

Data point	x_1	x_2	Y
1	−2	1	10
2	−1	−2	5
3	0	0	10
4	1	2	15
5	2	−1	10

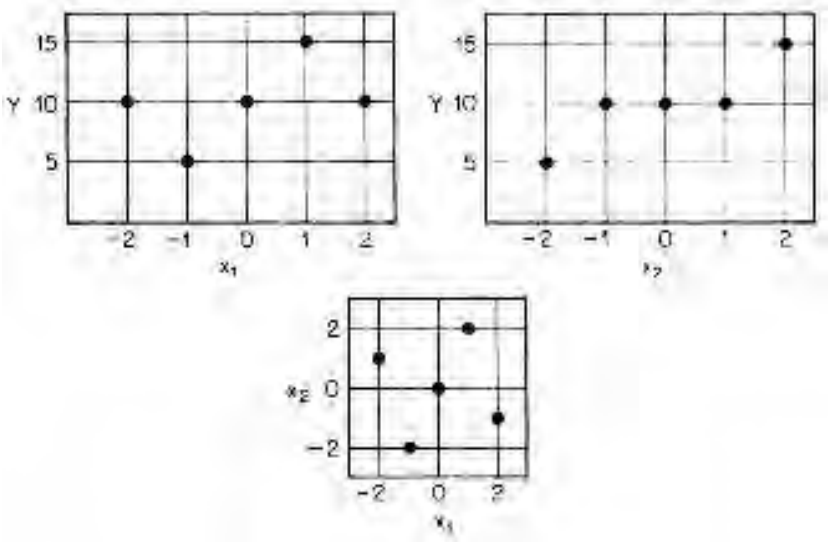


FIGURE 44.3 Crossplots for data set *B* of Table 44.4.

Plots versus Time Order of Data. After a data set has been obtained, it is very instructive in many situations to plot the output Y against the time order in which the experiments were run (which is essentially a crossplot of Y versus time t). Among the possible phenomena that may come to our attention from such a plot are these:

A few observations, often at the start of the experimentation, are far from the others; this often represents a learning curve of the experimenters with the experimental situation, and those experiments should be repeated if possible.

There are trends within each day (or within each week, etc.); this may represent such phenomena as warming of a machine or process, operator fatigue, or similar time-related trends.

Variability decreases (or increases) with time; this may be due to a learning curve or raw material characteristics (as when one lot of material is used up and the next lot has less or greater heterogeneity).

While the preceding trends may be apparent even in a plot of the original observations Y versus time, they are often more easily spotted in plots of the residuals of the observations (difference between the observed and predicted value) after a regression analysis. See later in this section under Regression Analysis.

Frequency Histograms. The *frequency histogram* (or distribution) is a statistical tool for presenting numerous data in a form that makes clearer the central tendency and the dispersion along the scale of measurement, as well as the relative frequency of occurrence of the various values.

Table 44.5 shows "raw data" representing measurements of electrical resistance of 100 coils. A practitioner scanning these 100 numbers has difficulty in grasping their meaning.

Table 44.6 shows the same data after tabulation. Note how the analyst's tallies in the column "Tabulation" make more evident where the central tendency is and what the dispersion is. The column "Frequency" is merely a recorded count of these same tallies. The column "Cumulative frequency" shows the number of coils with resistance equal to or greater than the associated resistance value.

Table 44.6 exhibits a range of values from 3.44 to 3.27, or 17 intervals of 0.01 Ω each. When it is desired to reduce the number of such intervals, the data are grouped into "cells." Table 44.7 shows the same data grouped into a frequency distribution of only six cells, each 0.03 Ω wide. Grouping the data into cells simplifies presentation and study of the distribution but loses some of the detail. (However, one can always go back to the original data if necessary.)

The following are the steps taken to construct a frequency distribution:

1. Decide on the number of cells. Table 44.8 provides guidelines that are adequate for most cases encountered. These guidelines are not rigid and should be adjusted when necessary; their aim is not only to provide a clear data summary but also to reveal any underlying pattern in the data.

TABLE 44.5 Resistance (Ohms) of 100 Coils

3.37	3.34	3.38	3.32	3.33	3.28	3.34	3.31	3.33	3.34
3.29	3.36	3.30	3.31	3.33	3.34	3.34	3.36	3.39	3.34
3.35	3.36	3.30	3.32	3.33	3.35	3.35	3.34	3.32	3.38
3.32	3.37	3.34	3.38	3.36	3.37	3.36	3.31	3.33	3.30
3.35	3.33	3.38	3.37	3.44	3.32	3.36	3.32	3.29	3.35
3.38	3.39	3.34	3.32	3.30	3.39	3.36	3.40	3.32	3.33
3.29	3.41	3.27	3.36	3.41	3.37	3.36	3.37	3.33	3.36
3.31	3.33	3.35	3.34	3.35	3.34	3.31	3.36	3.37	3.35
3.40	3.35	3.37	3.32	3.35	3.36	3.38	3.35	3.31	3.34
3.35	3.36	3.39	3.31	3.31	3.30	3.35	3.33	3.35	3.31

TABLE 44.6 Tally of Resistance Values of 100 Coils

Resistance, ohms	Tabulation	Frequency	Cumulative frequency
3.45			
3.44	|	1	1
3.43			
3.42			
3.41	||	2	3
3.40	||	2	5
3.39	||||	4	9
3.38	𝍸 |	6	15
3.37	𝍸 |||	8	23
3.36	𝍸 𝍸 |||	13	36
3.35	𝍸 𝍸 ||||	14	50
3.34	𝍸 𝍸 ||	12	62
3.33	𝍸 𝍸	10	72
3.32	𝍸 ||||	9	81
3.31	𝍸 ||||	9	90
3.30	𝍸	5	95
3.29	|||	3	98
3.28	|	1	99
3.27	|	1	100
3.26			
Total		100	

TABLE 44.7 Frequency Distribution of Resistance Values

Resistance, ohms		Frequency	Cumulative frequency
Boundaries	Midpoints		
3.415–3.445	3.43	1	1
3.385–3.415	3.40	8	9
3.355–3.385	3.37	27	36
3.325–3.355	3.34	36	72
3.295–3.325	3.31	23	95
3.265–3.295	3.28	5	100
		100	

2. Calculate the approximate cell interval *i*. The cell interval equals the largest observation minus the smallest observation divided by the number of cells. Round this result to some convenient number.
3. Construct the cells by listing cell boundaries. As an aid to later calculation:
   a. The cell boundaries should be to one more decimal place than the actual data and should end in a 5.
   b. The cell interval should be constant throughout the entire frequency distribution.
4. Tally each observation into the appropriate cell and then list the total frequency *f* for each cell.

There are several ways of showing a frequency distribution in graphic form. The most popular is the frequency histogram. Figure 44.4 shows the electrical resistance data of Table 44.7 depicted in histogram form. The diagram is so easy to construct and interpret that it is widely used in elementary analysis of data.

One example of wide, effective use of frequency histograms is comparison of process capabilities with tolerance limits. The histogram of Figure 44.5 shows a process that is inherently capable of holding the tolerances drawn on the same figure. The high degree of defectives being produced is the result of running this process at a setting that does not locate its central tendency near the midpoint of the tolerance range. (See Section 22, under Operations Analysis, for other examples.)

Analyses of histograms to draw conclusions beyond the sample data customarily should be based on at least 50 measurements.

Sample Characteristics: Mean, Median, Mode, Variance, Standard Deviation, Percentiles. Faced with a large data set, *descriptive statistics* furnish a simple method of extracting information from what often seems at first glance to be a mass of numbers without rhyme or reason to it. These characteristics of the data may relate to a "typical (or central) value" (mean, median, mode), a measure of how much variability is present (variance, standard deviation), or a measure of frequency (percentiles). The first two types of characteristics (typical value and variability) will be discussed below, but first we will present the concept of percentiles.

TABLE 44.8 Number of Cells in Frequency Distribution

Number of observations	Recommended number of cells
20–50	6
51–100	7
101–200	8
201–500	9
501–1000	10
Over 1000	11–20

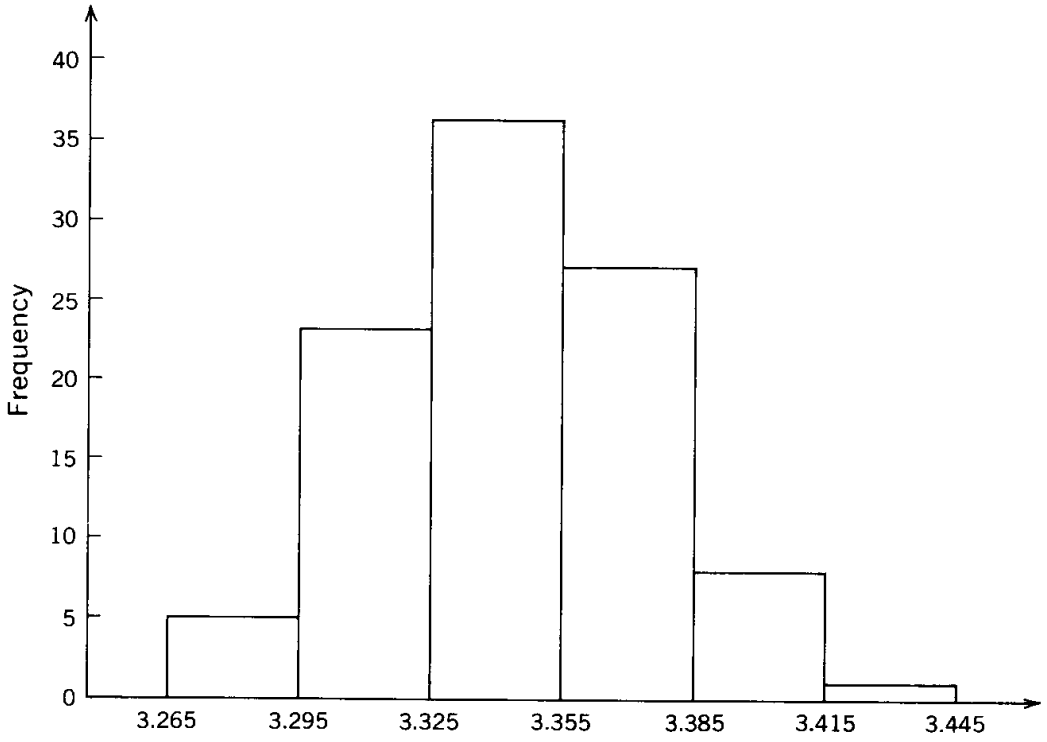


FIGURE 44.4 Histogram of resistance.

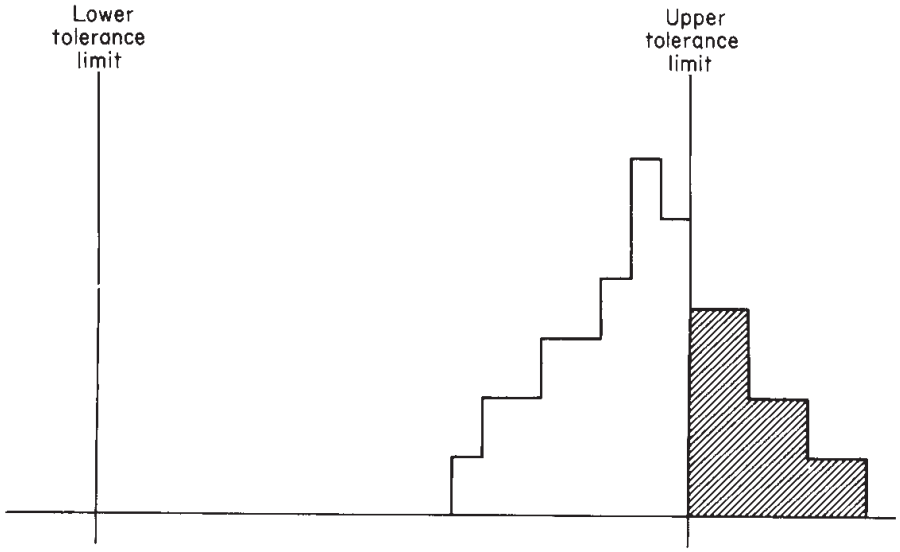


FIGURE 44.5 Histogram of a process.

A *percentile curve* is a plot of the percentile rank of the data against the data values. For example, for the data of resistance of 100 coils given in Table 44.6, 1 percent are at or below resistance 3.27, 2 percent are at or below 3.28, 5 percent are at or below 3.29, and so on, as given in Table 44.9. The percentile rank plot (or percentile curve) for these data is given in Figure 44.6. Note that while the data will result in a "rough" curve (since the percentile curve of the data will jump at each data point and remain constant between actual data points), a smooth curve drawn through the data will be a better representation of reality; this is the dashed curve drawn in Figure 44.6. (Thus while values 3.42 and 3.43 did not occur in our sample, we expect they would in a larger sample; hence the smooth curve is the one we would utilize to assess their chances of occurring.) Note that no data are discarded (or grouped) in making the percentile curve of a data set; in this sense, this is a more precise process than construction of a histogram of the data (since no information is "lost" in the process). Most statistical work uses the percentile curve, under the name *distribution function* of the data. (In early papers this was called the *cumulative* distribution function.) Since the distribution gives the proportion of the data falling at or below each value, the graph and curve are the same; only the scale on the vertical axis is changed to read from 0 to 1 (since 0 to 100, divided by 100, runs from 0 to 1).

Measures of Central Tendency/Location. Most frequency distributions exhibit a *central tendency,* i.e., a shape such that the bulk of the observations pile up in the area between the two extremes. Central tendency is one of the two most fundamental concepts in all statistical analysis.

There are three principal measures of central tendency: arithmetic mean, median, and mode. The *arithmetic mean* (the ordinary "average") is used for symmetric or near symmetric distributions or for distributions that lack a clearly dominant single peak. The arithmetic mean $\overline{X}$ is the most generally used measure in quality work. It is employed so often to report average size, average yield, average percent defective, etc. that control charts have been devised to analyze and keep track of it. Such control charts can give early warning of significant changes in the central value (see Section 45, Statistical Process Control).

The *mean* is calculated by adding the observations and dividing by the number of observations. A short method for calculating the mean is given in a subsequent example under Measures of Dispersion.

The *median* (the middle value when the figures are arranged according to size) is used for reducing the effects of extreme values or for data that can be ranked but are not economically measurable (shades of color, visual appearance, odors) or for special testing situations. If, for example, the average of five parts tested is used to decide whether a life-test requirement has been met, then the life-

time of the third part to fail can sometimes serve to predict the average of all five, and thereby the decision of the test can be made much sooner. As shown on Figure 44.6, the median is simply the horizontal scale value where the percentile curve reaches the height 50 percent.

The *mode* (the value that occurs most often in data) is used for severely skewed distributions, describing an irregular situation where two peaks are found, or for eliminating the effects of extreme values. The statistical "efficiency" of these measures varies. See Dixon and Massey (1969, chap. 9) or Dudewicz (1976, pp. 221–222) for elaboration.

Measures of Dispersion. Data are always scattered around the zone of central tendency, and the extent of this scatter is called *dispersion* or *variation.* A measure of dispersion is the second of the two most fundamental measures in all statistical analysis.

TABLE 44.9 Percentile Rank of Resistance Values in Table 44.6

Resistance (x)	3.27	3.28	3.29	3.30	3.31	3.32	3.33	3.34
Percentile rank (y)	1	2	5	10	19	28	38	50
	3.35	3.36	3.37	3.38	3.39	3.40	3.41	3.42
	64	77	85	91	95	97	99	99
	3.43	3.44						
	99	100						

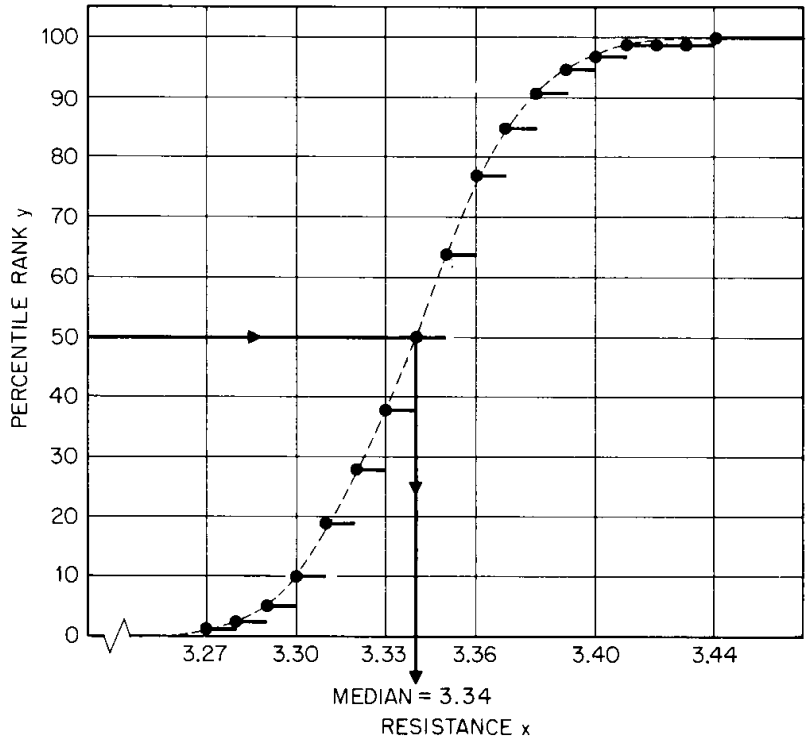


FIGURE 44.6 Percentile curve of a data set.

There are several measures of dispersion. The simplest is the *range,* which is the difference between the maximum and minimum values in the data. Since the range is based on only two values, it is most useful when the number of observations is small (about 10 or fewer).

The most important measure of variation is the *standard deviation.* The definition of the sample standard deviation is

$$s = \sqrt{\frac{\sum(X - \overline{X})^2}{n - 1}}$$

where s = sample standard deviation
$\sum$ = "sum of"
X = observed values
$\overline{X}$ = arithmetic mean
n = number of observations

For calculation purposes, an equivalent formula is

$$s = \sqrt{\frac{n\sum(X^2) - (\sum X)^2}{n(n - 1)}}$$

The square of the standard deviation is called the *variance.* There is also a measure called *covariance,* which gives information on the relationship between pairs of observations on characteristics X and Y. This is defined as

$$s_{XY} = \frac{\sum[(X - \overline{X})(Y - \overline{Y})]}{n - 1}$$

Further discussion of the relationship between two or more variables is given later in this section under Regression and Correlation Analysis.

With data in frequency distribution form, shortcut calculations can simplify finding the average and the standard deviation. This is illustrated in Table 44.10. To start, an arbitrary origin A is assumed as 3.37. A zero is arbitrarily placed on this line in the d' column. The other figures in this column indicate how many cells the entry is above or below the arbitrary zero. Minus signs are attached when the entry is smaller than the assumed value, 3.37. The fd' values in column (4) are found by multiplying together the entries in columns (2) and (3). Similarly $f(d')^2$ is found by multiplying the figures in columns (3) and (4). Note that the totals in the last two columns are identified in the formulas as $\sum fd'$ and $\sum f(d')^2$, respectively, and i is the cell interval. Since the multiplications are small enough to be carried out mentally, the complete table can be made quickly.

$$\overline{X} = A + \left(\frac{\sum fd'}{n}\right) i = 3.37 + \left(\frac{-87}{100}\right) 0.03 = 3.344$$

$$s = i \sqrt{\frac{n\sum f(d')^2 - (\sum fd')^2}{n(n - 1)}}$$

$$s = 0.03 \sqrt{\frac{100(185) - (-87)^2}{100(99)}} = 0.031$$

For sample sizes of about 10 or fewer observations, the standard deviation can be approximated from the range by calculating R/d_2, where d_2 is a factor in Appendix II, Table A. For example, suppose the first column of values in Table 44.5 represents a sample of 10. The range is 3.40−3.29, or 0.11. From Table A in Appendix II, $d_2 = 3.078$. The estimate of the standard deviation is therefore 0.11/3.078 = 0.036. This is much simpler than calculating the standard deviation directly. Subsequent

TABLE 44.10 Calculation of Average and Standard Deviation

Midpoint (1)	Frequency f (2)	d' (3)	fd' (4)	$f(d')^2$ (5)
3.43	1	+2	2	4
3.40	8	+1	8	8
3.37	27	0	0	0
3.34	36	−1	−36	36
3.31	23	−2	−46	92
3.28	5	−3	−15	45
	$\Sigma = 100$		$\Sigma = -87$	$\Sigma = 185$

topics in this section further illustrate this feature of the range. Dixon and Massey (1969, pp. 136–140) furnish procedures and tables for a variety of applications of the range.

A final measure of variation is the *coefficient of variation.* This is defined as the standard deviation divided by the mean and is thus a relative measure of variation. It can be helpful in comparing several sets of similar data that differ in mean value but may have some commonality in *relative* variation.

The methods of summarizing data covered in the previous paragraphs can be performed on a computer, as discussed below. (Also see Section 10 in general for additional information on computer programs for quality control.)

Stem-and-Leaf Plots, Boxplots, and Statistical Graphics. Looking at data to find patterns and other characteristics is very important. In fact, it is often said that "the first rule of statistics is: Look at the data." This does not mean to pass the data on to be fed into some computer program without a careful examination of them. It does mean to visually scan the data, to calculate classic measures of important characteristics, to check to see if known measures of validity are satisfied, and to make graphic representations of some data characteristics. There are many methods and approaches to this initial stage of data analysis, some favored by one or another school of statisticians, all of some use, and the total group being of significant size and utility. Some widely valued methods (in addition to those covered in detail earlier) include stem-and-leaf plots and boxplots, for which one can refer to many introductory statistics texts. The area as a whole is part of what is called *statistical graphics,* and computer programs to perform such calculations are widely available. However, to achieve their full potential (and rise above mediocrity or inferiority), these methods need to be used carefully and intelligently (as we saw earlier with crossplots, for example). An excellent reference in this regard (one that is filled with examples, and when the examples are of bad graphics shows how to use a good graphic on the same data) is Schmid (1983).

Accurate Calculation of Descriptive Statistics. Accurate calculation of even simple descriptive statistics is not as easy a task as it might seem at first, as the example in the paragraph on coding of data below shows. In particular,

Calculation by hand is an error-prone process, especially if there are 10 or more data points.

Construction of computer programs to perform the calculations is a task that requires knowledge of both statistics and numerical analysis as well as computer programming; hence it is inadvisable to "roll your own" in most cases.

Many of the computer software routines that come with computers (especially microcomputers) are not of high quality.

In light of the preceding, it is recommended that a high-quality package of computer routines for statistics be obtained and used for all such analyses. In particular, I suggest packages called SAS,

BMDP, STATPRO, and LABONE. These are of high quality and among them cover a wide variety of computers on which they will run (including IBM and Data General mainframe computers, APPLE and IBM PC microcomputers, and a variety of minicomputers). Additional packages that are in wide use and widely valued include MINITAB and SPSS. There are many other packages, some specially designed for particular areas of statistics (such as design of experiments). It is good to bear in mind that quality varies widely; SAS is felt by many to be the "gold standard" of such packages.

For details of SAS, see SAS Institute, Inc. (1982). For BMDP, see Dixon (1983). For STATPRO, see Pinsky (1983). For LABONE, see Levy and Dumoulin (1984). In each case, a telephone call will bring current information regarding which computers the package is available for, as well as licensing information.

Section 10, under Computer Applications and Quality Systems, presents additional information on statistical packages. The graphics available with these packages (such as SAS/GRAPH with SAS) are of such a quality that they would often justify the expenditure for the package to management. The American Society for Quality annually publishes a "Directory of Software for Quality Assurance and Quality Control" (see, for example, ASQC 1996).

Coding of Data. Suppose that we have five observations, $X_1,\ldots, X_5$ as in Table 44.11 and wish to compute the mean and standard deviation. Then on a pocket calculator (such as the Texas Instruments TI-55) or on a mainframe computer, we find

$$\sum_{i=1}^{5} X_1 = 49345 \qquad \sum_{i=1}^{5} X_i^2 = 4.8698581 \times 10^8$$

$$\overline{X} = 9869 \qquad s^2 = 0.985 \qquad s = 0.992$$

The only problem with these answers is that they are *wrong.* For example, 9869 is less than all the observations; hence it cannot be the mean. And the standard deviation is a measure of the spread in the data; that spread from largest to smallest is 0.00008; hence 0.992 is much too large. The problem that has led to this inaccuracy is that the computers used keep only about eight decimal places of accuracy, and this results in discarding digits that are needed for accurate calculation. If it can be so troublesome to calculate the mean and standard deviation of five numbers accurately, clearly, problems of meaningful size require careful analysis of numerical inaccuracy, as discussed above, and for this reason, software should not be trusted without a careful analysis of its capabilities. The four packages suggested are of high quality.

If one must use untested software, it is recommended that one *code* the data; that is, calculate using $Y_1,\ldots, Y_5$, where $Y_i = a(X_i + b)$ for some a and b. For example, if we choose $a = 10^5$ and $b = -9869$, we will have $Y_1 = 13$, $Y_2 = 7$, $Y_3 = 15$, $Y_4 = 8$, and $Y_5 = 9$. For these we calculate (on the same computers)

$$\sum_{i=1}^{5} Y_i = 52 \qquad \sum_{i=1}^{5} Y_i^2 = 588 \qquad \overline{Y} = 10.4 \qquad s_Y^2 = 11.8 \qquad s_Y = 3.4351$$

which are exactly correct. Now it can be shown that the relation between $\overline{X}$, s_X^2 and $\overline{Y}$, s_Y^2 is

$$\overline{X} = \frac{1}{a}\overline{Y} - b \qquad s_X^2 = \frac{1}{a^2} s_Y^2$$

TABLE 44.11 Five Data Points

$X_1 = 9869.00013$
$X_2 = 9869.00007$
$X_3 = 9869.00015$
$X_4 = 9869.00008$
$X_5 = 9869.00009$

Hence we find (exactly again)

$$\overline{X} = 10.4 \times 10^{-5} + 9869 = 9869.000104$$

$$s_X^{\,2} = 11.8 \times 10^{-10} \text{ (hence } s_X = 0.000034)$$

This method of *coding the data* will preserve accuracy in many other, more complex statistical calculations (such as regression) and is in fact done internally by many of the high-quality statistical computer packages. Thus it should be used whenever you are using a package other than one of the high-quality ones listed earlier. In addition, many absurd results from inaccurate software will be detected early if one observes the cardinal rule of statistics, namely, "Look at the data." (It is all too common to have the data gathered and analyzed via computer without a careful look at them with the measures recommended in this section, and this has led to many costly problems for many companies.)

In terms of the choice of a and b for the coding, the best values to use are $a = 1/s_X$ and $b = -\overline{X}$. If these values cannot be calculated in a trustworthy manner by the software, the simple estimates of the next paragraph may be used instead.

Simple Estimates of Location and Dispersion. Simple estimates of the center and the spread of a set of data are often desired (e.g., for use in coding of data, as discussed earlier, and also for rapid analysis of a data set under time pressure). Two simple measures for the center are the median ($\widetilde{X}$) and the midrange, expressed as

$$\frac{\text{Maximum } (X) + \text{minimum } (X)}{2}$$

A simple measure for the variability is

$$\frac{\text{Maximum } (X) - \text{minimum } (X)}{4}$$

For the data of Table 44.11, the true value was $\overline{X} = 9860.000104$. The two simple methods yield 9869.00009 and 9869.000110, respectively. Similarly, the true value of the standard deviation was $s_X = 0.000034$, and the simple estimate yields 0.000020. (If we use Table A in Appendix II and take the range divided by d_2 instead of the range divided by 4, we will have a better estimate. With five data points, $d_2 = 2.326$, so our estimate of s_X is then 0.000080/2.326 = 0.000034.)

PROBABILITY MODELS FOR EXPERIMENTS

A distinction is made between a sample and a population (or *universe*). A *sample* is a limited number of measurements taken from a large source. A *population* is a large source of measurements from which the sample is taken. (Note that a population may physically exist, such as all stereo sets in a certain lot. It also may be conceptual, as all experiments that might be run.)

A *probability distribution* is a mathematical formula that relates the values of the characteristic with their probability of occurrence in the population. Figure 44.7 summarizes some distributions. When the characteristic being measured can take on any value (subject to the fineness of the measuring process), its probability distribution is called a *continuous* probability distribution. For example, the probability distribution for the resistance data of Table 44.7 is an example of a continuous probability distribution because the resistance could have any value, limited only by the fineness of the measuring instrument. Experience has shown that most continuous characteristics either follow one of several common probability distributions, the *normal* distribution, the *exponential* distribution, and the *Weibull* distribution, or can be fitted with an empirical estimate, as discussed later in this section. These distributions find the probabilities associated with occurrences of the actual

values of the characteristic. Other continuous distributions (e.g., *t, F,* and chi square) are important in data analysis but do not provide probabilities of occurrence of actual values.

When the characteristic being measured can take on only certain specific values (e.g., integers 0, 1, 2, 3, etc.), its probability distribution is called a *discrete* probability distribution. For example, the distribution for the number of defectives r in a sample of five items is a discrete probability distribution because r can only be 0, 1, 2, 3, 4, or 5. The common discrete distributions are the Poisson, binomial, negative binomial, and hypergeometric (see Figure 44.7).

The following paragraphs explain how probability distributions can be used with a sample of observations to make predictions about the larger population.

Sample Space. Statistics deals with the *outcomes of experiments.* When an experiment is performed, some outcome results; let us denote a typical outcome by the symbol *e.* Such an outcome is called a *simple event.* If we list all the possible outcomes of the experiment of interest to us, that set is called the *sample space* of the experiment.

As an example, if we perform the experiment of tossing three coins and observing for each coin whether it lands heads (H) or tails (T), the sample space will contain the eight possible outcomes

HHH HHT HTH THH HTT THT TTH TTT

For simplicity of notation, let us denote these simple outcomes, respectively, by e_1, e_2, e_3, e_4, e_5, e_6, e_7, and e_8.

We associate a number called *probability* with each of the simple events. We think of this number as representing the proportion of times each simple event would occur in a very large number of experiments of this type. For example, the probability of HHH in our experiment of tossing three coins is usually taken to be $1/8 = 0.125$ because it typically occurs in about one-eighth of a large number of experiments where three coins are tossed. We denote the probability of a simple event e by $P(e)$; thus we usually have $P(\text{HHH}) = 1/8$.

Since some outcome occurs in each experiment, when we add up the proportion of times that each e in the sample space occurred, we must obtain a sum of 1. Since probabilities represent what those proportions would be in a large number of experiments, we also must have *probabilities* that *sum to 1 when all outcomes are accounted for.* For example, in our example with three coins,

$$P(e_1) + P(e_2) + P(e_3) + P(e_4) + P(e_5) + P(e_6) + P(e_7) + P(e_8) = 1$$

Events. Very often we are interested not in a simple event but in a combination of them, called a *composite event.* For example, the event "more heads than tails" occurs if and only if one of the simple events e_1, e_2, e_3, e_4 (i.e., the simple events HHH, HHT, HTH, THH) occurs in our example of tossing three coins. The frequency with which we find "more heads than tails" will be the sum of the relative frequencies of e_1, e_2, e_3, and e_4. Thus we say the probability of the event "more heads than tails" is the sum of the probabilities of the simple events that comprise the event "more heads than tails":

$$P(\text{more heads than tails}) = P(e_1) + P(e_2) + P(e_3) + P(e_4)$$

To make this simpler to write, we often denote the event of interest by a symbol, such as A for the event "more heads than tails." Then

$$P(A) = P(e_1) + P(e_2) + P(e_3) + P(e_4)$$

Thus *the probability of a composite event is the sum of the probabilities of all the simple events that comprise it.* (Note that simple events are always mutually exclusive.) Since in the example with three coins we have $P(e_1) = P(e_2) = \ldots = P(e_8) = 1/8$, we find

$$P(A) = 1/8 + 1/8 + 1/8 + 1/8 = 1/2$$

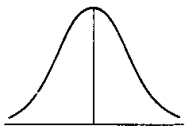
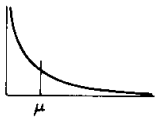
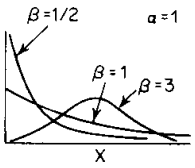
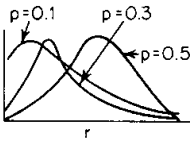
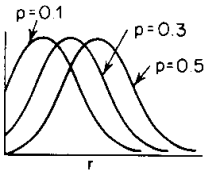
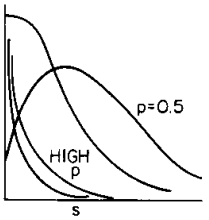
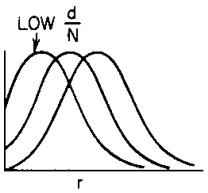
DISTRIBUTION	FORM	PROBABILITY FUNCTION	COMMENTS ON APPLICATION
NORMAL		$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$ μ = Mean σ = Standard deviation	Applicable when there is a concentration of observations about the average and it is equally likely that observations will occur above and below the average. Variation in observations is usually the result of many small causes.
EXPONENTIAL		$y = \frac{1}{\mu} e^{-\frac{x}{\mu}}$	Applicable when it is likely that more observations will occur below the average than above.
WEIBULL		$y = \alpha\beta(X-\gamma)^{\beta-1} e^{-\alpha(X-\gamma)^{\beta}}$ α = Scale parameter β = Shape parameter γ = Location parameter	Applicable in describing a wide variety of patterns of variation, including departures from the normal and exponential.
POISSON*		$y = \frac{(np)^r e^{-np}}{r!}$ n = Number of trials r = Number of occurrences p = Probability of occurrence	Same as binomial but particularly applicable when there are many opportunities for occurrence of an event, but a low probability (less than 0.10) on each trial.
BINOMIAL*		$y = \frac{n!}{r!(n-r)!} p^r q^{n-r}$ n Number of trials r = Number of occurrences p = Probability of occurrence q = 1−p	Applicable in defining the probability of r occurrences in n trials of an event which has a probability of occurrence of p on each trial.
NEGATIVE BINOMIAL*		$y = \frac{(r+s-1)!}{(r-1)!(s!)} p^r q^s$ r = Number of occurrences s = Difference between number of trials and number of occurrences p = probability of occurrence q = 1-p	Applicable in defining the probability that r occurrences will require a total of r + s trials of an event which has a probability of occurrence of p on each trial. (Note that the total number of trials n is r + s.)
HYPERGEOMETRIC*		$y = \frac{\binom{d}{r}\binom{N-d}{n-r}}{\binom{N}{n}}$	Applicable in defining the probability of r occurrences in n trials of an event when there are a total of d occurrences in a population of N.

FIGURE 44.7 Summary of common univariate probability distributions. (Asterisks indicate that these are discrete distributions, but the curves are shown as continuous for ease of comparison with the continuous distributions.)

i.e., we expect to find more heads than tails (when three coins are tossed) in about 50 percent of such experiments.

In the example with three coins, we have *equally likely simple events,* i.e., $P(e_i) = P(e_j)$ for all *i, j.* When this is true, it follows that for any composite event A we have

$$P(A) = \frac{\text{number of simple events in } A}{\text{number of points in the sample space}}$$

In the case of the three coins, this yields the same answer obtained before, namely, $P(A) = {}^4/_8 = {}^1/_2$.

We say two composite events A_1 and A_2 are *mutually exclusive* if no e_i is in both A_1 and A_2. For example, if A_1 is the event "2 heads" and A_2 is the event "more tails than heads," then A_1 and A_2 are mutually exclusive because $A_1 = \{e_2, e_3, e_4\}$ and $A_2 = \{e_5, e_6, e_7, e_8\}$ have no point in common. We often express the fact that A_1 and A_2 are mutually exclusive in shorthand by writing

$$A_1 A_2 = \phi$$

If A_1 and A_2 are mutually exclusive, then for the event "A_1 or A_2" (which occurs if and only if at least one of A_1, A_2 occurs), *we have*

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2)$$

This follows because $P(A_1$ or $A_2)$ equals the number of e_i in either A_1 or A_2 divided by the total number of simple events; since there are no points in both A_1 and A_2, this is the same as taking the number of points in A_1 and adding to it the number in A_2 and then dividing by the total number of simple events. In our example, $P(A_1 \text{ or } A_2) = P(A_1) + P(A_2) = {}^3/_8 + {}^4/_8 = {}^7/_8$.

In our example so far, we have discussed the events

A: "more heads than tails"

A_1: "2 heads"

A_2: "more tails than heads"

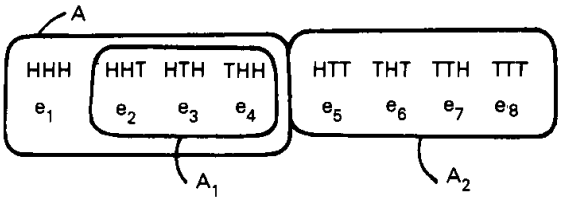


FIGURE 44.8 Events and sample space, experiment of tossing 3 coins.

These are shown on the sample space in Figure 44.8. Here A and A_1 are not mutually exclusive, so the simple addition of the probabilities does not hold for them, since $P(A) + P(A_1) = {}^4/_8 + {}^3/_8 = {}^7/_8$ counts the points e_2, e_3, e_4 twice as to their probabilities. Thus we see that a correct equation for $P(A \text{ or } A_1)$ will need to subtract this overcounting part, which is $P(A \text{ and } A_1)$, i.e., the probability that we have an experimental outcome where both A *and* A_1 occur, which is $P(A \text{ and } A_1) = {}^3/_8$ in this example. Thus we have reasoned to the fact that *for any events A and A_1 the addition rule is*

$$P(A \text{ or } A_1) = P(A) + P(A_1) - P(A \text{ and } A_1)$$

For mutually exclusive events, the $P(A \text{ and } A_1)$ would be 0 so that this would reduce to simple addition of $P(A)$ and $P(A_1)$ in that case.

Rules of Probability, Combinatorics. Probability theory underlies all decisions that are based on sampling. As we have seen, probability is expressed as a number that lies between 1.0 (certainty that an event will occur) and 0.0 (impossibility of occurrence), and the most intuitive definition of probability is one based on a frequency interpretation. In the simple case when an event A can occur in s cases out of a total of n possible and equally probable cases, then the probability that the event will occur is

$$P(A) = \frac{s}{n} = \frac{\text{number of successful cases}}{\text{total number of possible cases}}$$

Counting s and n can be complex, in which case it is called a problem of *combinatorics.*

Example: A lot consists of 100 parts. A single part is selected *at random,* which means that each of the 100 parts has an equal chance of being selected. Suppose a lot contains a total of 8 nonconforming parts. Then the probability of drawing a single part that is nonconforming is 8/100, or 0.08.

The following theorems are useful in solving problems involving probability:

Theorem 1: If $P(A)$ is the probability that an event A will occur, then the probability that A will not occur is $1 - P(A)$.

Theorem 2: If A and B are two events, then the probability that either A or B occurs is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

A special case of this theorem occurs when A and B cannot occur simultaneously (i.e., A and B are "mutually exclusive"). Then the probability that either A or B occurs is

$$P(A \text{ or } B) = P(A) + P(B)$$

Example: The probabilities of r defectives in a sample of six units from a 5 percent defective lot are found below by the binomial. The probability of zero defectives is 0.7351; the probability of one defective is 0.2321. The probability of zero or one defective is then $0.7351 + 0.2321$, or 0.9672.

Theorem 3: If A and B are two events, then the probability of the joint occurrence of both A and B is

$$P(A \text{ and } B) = P(A) \times P(B|A)$$

where $P(B|A)$ = probability that B will occur assuming A has already occurred.

A special case of this theorem occurs when the two events are *independent*; i.e., when the occurrence of one event has no influence on the probability of the other event. If A and B are independent, then the probability of both A and B occurring is

$$P(A \text{ and } B) = P(A) \times P(B)$$

Example: A complex system consists of two major independent subsystems. The probability of successful performance of the first subsystem is 0.95; the corresponding probability for the second subsystem is 0.90. Both subsystems must operate successfully to achieve total system success. The probability of the successful operation of the total system is therefore $0.95 \times 0.90 = 0.855$.

The preceding theorems have been stated in terms of two events but can be expanded for any number of events.

Conditional Probability; Bayes' Theorem. In *conditional probability,* we seek an answer to such questions as, "If I know that A_2 has occurred, then on those trials of the experiment where A_2 has occurred, how often does A_1 occur?" We use a special shorthand symbol for this conditional probability:

$$P(A_1|A_2)$$

which is read as "the probability of A_1, given that A_2 is known to have occurred" and is calculated from the formula

$$P(A_1|A_2) = \frac{P(A_1 \text{ and } A_2)}{P(A_2)}$$

If $P(A_1|A_2) > P(A_1)$, we say A_2 carries "positive information" about A_1.

If $P(A_1|A_2) < P(A_1)$, we say A_2 carries "negative information" about A_1.

If $P(A_1|A_2) = P(A_1)$, we say A_2 carries no information about A_1 or that A_1 and A_2 are "independent events." In this last case, knowing that A_2 has (or has not) occurred does not change the chances of A_1 occurring.

The powerful *conditional probability reversal formula,* namely,

$$P(A_1|A_2)P(A_2) = P(A_1 \text{ and } A_2) = P(A_2 \text{ and } A_1) = P(A_2|A_1)P(A_1)$$

is the basis of Bayes' theorem. In the simplest setting, this states that

$$P(A_2|A_1) = \frac{P(A_1|A_2)P(A_2)}{P(A_1|A_2)P(A_2) + P(A_1|\text{not } A_2)P(\text{not } A_2)}$$

(Here, "not A_2" is the event that A_2 does not occur.) For worked solutions of problems in probability, see Dudewicz (1993).

The techniques presented under Tests of Hypotheses in this section consist of analyzing a sample of observations and reaching a conclusion (with defined sampling risks) to accept or reject a hypothesis. The experimenter considers the consequences of drawing incorrect conclusions and, to a lesser degree, the likelihood that extreme values of the population parameter will occur. However, this is usually done on a qualitative basis and involves judgment. In practice, a sample size is limited by economics, and the experimenter defines the type I error (usually 0.05 or 0.01) in numerical terms and then must accept the type II error that results with the sample size fixed by economics. There is a methodical way of defining the consequences of the type I and type II errors and the likelihood of extreme values. The approach involves Bayes' theorem and statistical decision theory.

Statistical Decision Theory. This concept requires two items of information not formally used in classic analysis:

1. The economic consequences of making type I or type II errors.
2. The probabilities that different values of the population parameter will occur. (The classical approach has no assumption concerning different values of the population parameter.)

Statistical tables and sampling plans based on Bayes' theorem or statistical decision concepts are not common, but the concepts can have a significant effect, and therefore development work seems imminent and worthwhile. Oliver and Springer (1972) give an example of Bayesian acceptance sampling tables. Hadley (1967) provides background material including examples on tests of hypotheses, confidence limits, and acceptance sampling plans. Lenz and Rendtel (1984) compare the performance of MIL-STD-105D, Skip Lot, and Bayesian sampling plans. See also the discussion in Section 46, under Bayesian Procedures.

Simpson's Paradox. What is called *Simpson's paradox* was discovered in 1951 by E. H. Simpson. Since then, it has been found to have many important implications in industry, medicine, and many other fields; a discussion is given in Dudewicz and Mishra (1988, pp. 55–57) and examples are given in Wardrop (1995) and its references. I will give an example to illustrate the problem and discuss how to avoid it.

Suppose that a company has two plants in different parts of the world. Both make a state-of-the-art product that is not yet fully understood (and hence much rework or scrap results). A new process modification is to be tried in both plants, head to head with the present method. The results of the trial come back and are summarized for management as follows: 46 percent successes (on 11,000

units made) with the new method but only 11 percent successes (on 10,100 units made) with the present method. Do you have enough information, and if so, what decision do you make on selection of a method to use in the future?

It is tempting to decide to use the new process; after all, it had over four times the success rate of the present process with a substantial number of trials. However, suppose we ask for more than the executive summary report, namely, for results in both the plants. In fact, it may be the case that the present method was substantially superior in *both* the plants, as illustrated in Table 44.12. There we see that the new method decreased the success rate in Plant 1 from 10 to 5 percent and in Plant 2 from 95 percent with the present method to 50 percent with the new method. Thus the present method was substantially superior in both plants (both Plant 1, which operates in adverse conditions and has a correspondingly lower success rate, and Plant 2, which has a higher success rate and thus received 10,000 of the new materials to test with).

This is paradoxical because the present method does better in both plants, but when one sums over plants, it does worse. In fact, people often think that this is impossible before studying a numerical example such as we have provided. To avoid erroneous decisions (such as selection of the inferior method), there are a number of actions one can take. First, in the design of trials that will take place in different facilities, assign the same number of test cases to each facility. (This is not always possible, since facilities may vary in size, prior commitments, etc.) Second, when assessing results that come from several (two or more) locations, machines, technicians, states, etc., insist that more than the overall summary of successes by method be provided (namely, one also needs to see the successes by location, machine, technician, state, etc.). This should allow one to avoid errors due to Simpson's paradox (and in addition will allow one to detect possibly important nonuniformity in success rates at different locations, machines, etc. that might otherwise go undetected).

DISCRETE PROBABILITY DISTRIBUTIONS

Discrete probability distributions are used to model situations where the outcome of interest can take on only a few discrete values (such as 0 or 1 for failure or success or 0, 1, 2, 3,…as a number of occurrences of some event of interest). Below I give the model leading to the most commonly occurring such distributions and consider where one can obtain numerical values of their probabilities, and their uses in quality control.

The Discrete Uniform Distribution. If each of the values $x_1,\ldots, x_n$ is equally likely to occur as the result of an experiment, then we say the value obtained has the uniform distribution on the set of values $x_1,\ldots, x_n$. In this case the probability of x_i is $1/n$. Since the probabilities are so simple, no special tables are needed.

Model Leading to a Uniform Distribution. The model leading to a uniform distribution is random selection from a finite population in which each value occurs the same number of times. (This makes values equally likely to occur in the sample.)

Uses of Random Choices. Random choices are often used in sampling inspection. For example, suppose that a lot of 1000 items is sequentially numbered from 500 through 1499. Then the chance

TABLE 44.12 Simpson's Paradox with Two Methods in Two Locations

		Plant 1	Plant 2
Present method:	Successes	1000	95
	Failures	9000	5
New method:	Successes	50	5000
	Failures	950	5000

that an item selected at random from the lot will have number i (for any i between 500 and 1499) is 1/1000 = 0.001. The probability that such an item will have a serial number at least 1400 is 100/1000 = 0.10.

The Binomial Distribution. If the probability of occurrence p of an event is constant on each of n independent trials of the event, then the probability of r occurrences in n trials is

$$\frac{n!}{r!(n-r)!}\,p^r q^{n-r}$$

where $q = 1 - p$.

In practice, the assumption of a constant probability of occurrence is considered reasonable when the population size is at least 10 times the sample size (under this circumstance, the change in probability from one trial to the next due to depletion of the population is negligible).

Table F in Appendix II provides partial tables for the binomial and gives references for more complete tables. King (1971, chaps. 20 through 22) discusses binomial probability.

Model Leading to a Binomial Distribution. When n independent trials of an experiment each have a constant probability p of occurrence of an event of interest (commonly termed a *success*), then the number of occurrences follows a binomial distribution. The name comes from the fact that the factor

$$\frac{n!}{r!(n-r)!}$$

in the probabilities is called a *binomial coefficient* in mathematics.

Binomial Probabilities and Uses. A lot of 100 units of product is submitted by a vendor whose past quality has been about 5 percent nonconforming. A random sample of six units is selected from the lot. The probabilities of various sample results are given in Table 44.13.

In using the formula, note that 0! = 1. Table F in Appendix II lists binomial probabilities in cumulative form, i.e., the probability of r or fewer occurrences in n trials. For the preceding example, the probability of 1 or fewer nonconforming items in a sample of 6 can be read from the table as 0.9672. Note that this is the sum of the probabilities for $r = 0$ and $r = 1$, i.e., 0.7351 + 0.2321 = 0.9672.

The Hypergeometric Distribution. Occasionally, the assumptions of the Poisson (see below) or binomial cannot be met even approximately. Subject only to the assumption of a random

TABLE 44.13 Table of Binomial Probabilities

r	P (exactly r defectives in 6) = $[6!/r!\,(6-r)!](0.05)^r(0.95)^{6-r}$
0	0.7351
1	0.2321
2	0.0306
3	0.0021
4	0.0001
5	0.0000
6	0.0000

sample, the hypergeometric gives the probability of exactly r occurrences in n trials from a lot of N items having d defectives as

$$\frac{\binom{d}{r}\binom{N-d}{n-r}}{\binom{N}{n}}$$

where $\binom{N}{n}$ is the "combinations" of N items taken n at a time and is equal to $N!/[n!(N-n)!]$, where $N! = [N(N-1)(N-2)\cdots 1]$ and $0! = 1$. The calculations can be avoided by using tables such as those prepared by Lieberman and Owen (1961). Duncan (1974) compares the results of Poisson, binomial, and hypergeometric distributions.

Model Leading to a Hypergeometric Distribution. The hypergeometric distribution is appropriate when independent trials are conducted, but the probability of occurrence of the event of interest changes from trial to trial because of depletion of a finite population. Because of their simpler form, in this situation one often uses the binomial or Poisson distributions if their assumptions are approximately met.

Hypergeometric Probabilities and Uses. A lot of 100 units is submitted by a vendor whose past quality has been about 5 percent nonconforming. A random sample of 20 units is selected from the lot. To calculate the probability of 0 nonconforming in 20, note that the lot has 5 nonconforming items and 95 conforming. Then

$$P(0 \text{ in } 20) = \frac{\binom{5}{0}\binom{95}{20}}{\binom{100}{20}} = \frac{\left[\dfrac{5!}{0!(5-0)!}\right]\left[\dfrac{95!}{20!(95-20)!}\right]}{\dfrac{100!}{20!(100-20)!}} = 0.319$$

Repeat substitutions into the formula are made to find $P(r \text{ in } 20)$, where r in this example is 0, 1, 2, 3, 4, and 5.

The Poisson Distribution. In practice, the most important discrete distribution is the Poisson. It is an approximation to more exact distributions (the hypergeometric and the binomial) and applies when the sample size is at least 16, the population size is at least 10 times the sample size, and the probability of occurrence p on each trial is less than 0.1. (These conditions are often met.)

Figure 44.7 states the Poisson probability function, but the real work is done by cumulative probability tables.

Model Leading to a Poisson Distribution. As well as being an approximation to more exact distributions, the Poisson is the exact distribution when certain assumptions are met. These assumptions are that events occur at random (in time, or in space, or in location, for example) with a probability of occurrence roughly proportional to the length of time (or volume of space, or area) and that there is no "clumping." [For details, see Dudewicz (1976, Section 3.2).] For example, if a target 0.1 mi^2 in size is known to be contained in an area 10 mi^2 in size, and this area is shelled at random (one shell at a time so there will be no clumping), the probability of a hit will be $0.1/10 = 0.01$. The number of hits will follow a Poisson distribution, and the number of shells fired may be set so that the probability of eight or more hits on the target will be at least 0.95; this is often done in practice when it is known that eight or more hits will effectively destroy the target and a 95 percent kill probability is desired.

Poisson Probabilities and Uses. A lot of 300 units of product is submitted by a supplier whose past quality has been about 2 percent nonconforming. A random sample of 40 units is selected from the

lot. Table E in Appendix II provides the probability of r or fewer defectives in a sample of n units. (The application of these probabilities is explained in Section 46, under Operating Characteristic, OC, Curve.) Entering the table with a value of np equal to 40(0.02), or 0.8, gives Table 44.14. Individual probabilities can be found by subtracting cumulative probabilities. Thus the probability of exactly two defectives is 0.953−0.809, or 0.144.

The Negative Binomial Distribution. The negative binomial distribution is one of the most commonly occurring distributions in situations where the sample size is not set in advance but rather is determined as the experiment proceeds.

Model Leading to a Negative Binomial Distribution. If the probability of occurrence of an event is constant from trial to trial and we make trials until we find m occurrences, then the probability that r trials will be needed is

$$\frac{(r-1)!}{(m-1)!(r-m)!}\,p^m(1-p)^{r-m}$$

where r can be m, $m+1$, $m+2$,…. This equation is equivalent to the more complex one listed in Figure 44.7. Other situations leading to a negative binomial distribution are discussed in Chapter 5 of Johnson and Kotz (1969). Tables are available in Williamson and Bretherton (1963) in case direct calculation is burdensome; Johnson and Kotz (1969) give references to additional tables.

Negative Binomial Probabilities and Uses. A large lot is inspected until the first defective ($m = 1$) is found; if this occurs in the first five trials, the lot is rejected. Hence the lot will be accepted if no defective is found in the first five trials (and thus trials 6, 7,… need not be performed—we will never inspect more than five items with this scheme). If the lot is 10 percent nonconforming, then from Table

TABLE 44.14 Table of Poisson Probabilities

r	Probability of r or fewer in sample
0	0.449
1	0.809
2	0.953
3	0.991
4	0.999
5	1.000

TABLE 44.15 Table of Negative Binomial Probabilities ($m=1$, $p=0.10$)

r	Probability r trials are needed to find $m = 1$ defectives
1	0.10
2	0.09
3	0.081
4	0.0729
5	0.06561
6	0.059049
7	0.053144
8	0.047830

44.15 we see that the probability the lot is accepted is $1 - (0.10 + 0.09 + 0.081 + 0.0729 + 0.06561) = 0.59049$.

The Multinomial Distribution. The discrete probability distributions discussed up to this point (the uniform, binomial, hypergeometric, Poisson, and negative binomial) all relate to situations that are *univariate,* that is, where the outcome of interest relates to one variable's value (such as the number of defectives in a sample of size n in the binomial case). However, there are important cases where the outcome is *multivariate.* One such example is that where in a sample of size n one observes both the number needing rework and the number to be scrapped, since there are two quantities, this is called a *bivariate* situation. The multinomial distribution can be used when there are any number of categories into which the items may be classified.

Model Leading to a Multinomial Distribution. If exactly one of the events $E_1,\ldots, E_k$ occurs on each of n independent trials and the probabilities of occurrence of the events are respectively $p_1,\ldots, p_k$ (with $p_1 + \ldots + p_k = 1$ so that one of them must occur), then the probability that E_1 occurs x_1 times and E_2 occurs x_2 times,..., and E_k occurs x_k times (where $x_1 + x_2 + \ldots + x_k = n$ since there are n trials and one of $E_1,\ldots, E_k$ must occur on each trial) is

$$\frac{n!}{x_1!x_2!\ldots x_k!} p_1^{x_1} p_2^{x_2}\ldots p_k^{x_k}$$

Tables are not widely available, and calculations are usually done directly from the probability formula, or (if n is large) the multivariate normal distribution (a continuous multivariate distribution discussed later in this section) is used as an approximation.

Multinomial Probabilities and Uses. Suppose that $n = 5$ large assemblies are manufactured and inspected each day. The results of each inspection are either pass, rework, or scrap. Past results have shown that 80 percent pass, 15 percent need rework, and 5 percent need to be scrapped. What are the probabilities of the various possible outcomes of one day's output?

Figure 44.9 shows the probabilities of the outcomes, calculated directly from the basic formula. Note that once the number to be reworked (e.g., 1) and the number to be scrapped (e.g., 0) are specified, the number passed is determined (e.g., $5 - 1 - 0 = 4$). We have had to use the multinomial probability distribution because the categories are not independent (an item cannot be both passed and scrapped).

Selecting a Discrete Distribution. Selection of which discrete distribution to use is usually made either by knowledge of the underlying situation or by fitting a model from relative frequency probability. In either case, a test of the model selected is desirable to check its validity.

Selection from a Model of Reality. In many cases one will know (or assume) that the model that leads to one of the distributions we have discussed underlies the practical situation. For example, if one draws 50 items at random from a large lot with $100p$ percent defective, one will assume the binomial model with $n = 50$ and probability p of a defective on each trial.

(Chi Square) Test of Model Validity. To test the validity of an assumed discrete model, where cell i should occur with probability p_i, one compares the observed cell totals with those predicted by the model using the chi square test discussed later (Test 12*b*). Such a confirmatory test can be omitted only if one is willing to run the risk of assuming a model that in fact may have little basis in reality. [See Section 9.12 of Dudewicz and Mishra (1988) and especially Problem 9.12.2 on p. 532.]

Empirical Models via Relative Frequency Probability. If there is little or no reason to lead to the adoption of one of the specific models discussed, a model can be fitted to the data using the relative frequencies observed in the past. For example, if one has observed that in 100 items produced in the

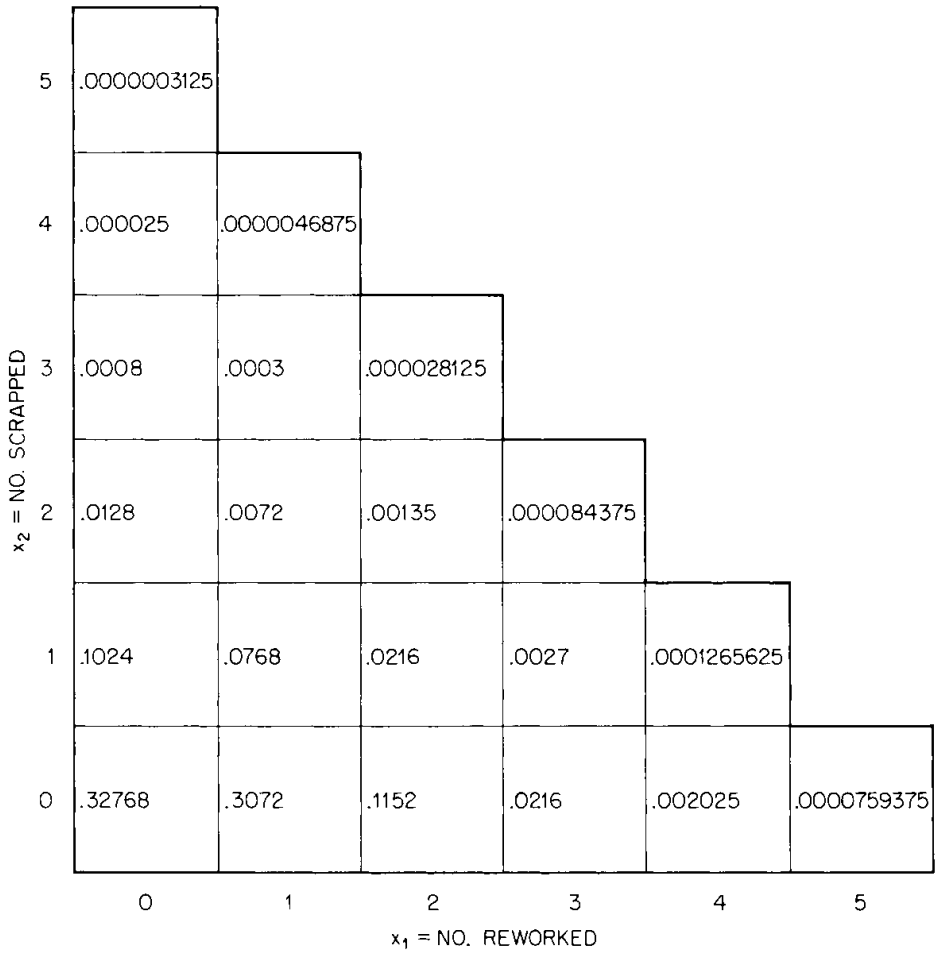


FIGURE 44.9 Table of multinomial probabilities, $k = 3$ Categories, $n = 5$, $p_1 = P$(Rework) $= 0.15$, $p_2 = P$(Scrap) $= 0.05$, $p_3 = P$(Pass) $= 0.80$.

past there have been blemishes present in the numbers given in Table 44.16, one would estimate the probability of r blemishes as its relative frequency in the past work, i.e., as in Table 44.17. Such a model should be tested by taking a new sample and performing a chi square test as previously discussed. See also Bootstrap Methods later in this section.

CONTINUOUS PROBABILITY DISTRIBUTIONS

Continuous probability distributions are used to model situations where the outcome of interest can take on values in a continuous range (such as all values greater than zero for the failure time of a motor that is run continuously). Below I give the model leading to the most commonly occurring such distributions and consider where one can obtain numerical values of their probabilities and their uses in quality control.

The Continuous Uniform Distribution. If all values between a and b ($a <$ b) are possible, and the chances of the value being in a subinterval are proportional to its length, then the uniform distribution is appropriate. The probability function is flat over the interval (a,b), where $y = 1/(b - a)$. Thus the probability the value is in a subinterval of length c is $c/(b - a)$. Since the probabilities are so simple, no special tables are needed.

Model Leading to a Uniform Distribution. The model leading to a continuous uniform distribution over the range (a,b) is random selection of a value between a and b. For example, if a valve on a water line is spun at random between pressure 0 lb/in^2 (closed) and 100 lb/in^2 (fully open), then the resulting pressure will be a uniform random variable on (0, 100).

Uses of Random Numbers. Uniform random variables on the range (0, 1) are called *random numbers.* Such variables are often used to drive digital computer simulation models and are of great importance in simulation studies of quality systems. Full details on sources of high-quality random numbers may be found in Dudewicz and Ralley (1981). Special considerations for simulation uses on microcomputers are given in Dudewicz et al. (1985).

The Exponential Distribution. The exponential probability function is

$$y = \frac{1}{\mu} e^{-X/\mu}$$

Figure 44.7 shows the shape of an exponential distribution curve. Note that the normal and exponential distributions have distinctly different shapes. An examination of the tables of areas shows that 50 percent of a normally distributed population occurs above the mean value and 50 percent below it. In an exponential population, 36.8 percent are above the mean and 63.2 percent below the mean. (This refutes the intuitive idea that the average is always the 50 percent point.) The property of a higher percentage below the average sometimes helps to indicate applications of the exponential. For example, the exponential describes the loading pattern for some structural members because smaller loads are more numerous than larger loads. The exponential is also useful in describing the distribution of failure times of certain complex equipment.

Model Leading to an Exponential Distribution. It can be shown that the exponential distribution of failure times arises when failures occur "at random" (and are not due to wearout but to such items as random shocks). In fact, the exponential distribution is characterized as the only continuous distribution with the "lack of memory property" that the chances of the item living an additional t_0 time units depend only on the length t_0 and not on how long the item has already been in use [see Dudewicz (1976), pp. 88, 106, for details].

Predictions with Exponential Distributions. Predictions based on an exponentially distributed population require only an estimate of the population mean. For example, the time between successive failures of a complex piece of equipment is measured, and the resulting histogram is found to

TABLE 44.16 Blemishes per Item in Past Work

No. of blemishes	No. of occurrences
0	50
1	25
2	25

TABLE 44.17 Blemish Probabilities via Relative Frequency Estimation

No. of blemishes	Probability
0	.50
1	.25
2	.25

resemble the exponential probability curve. The results of a sample of measurements indicate that the average time between failures (commonly called *MTBF,* or *mean time between failures*) is 100 h. What is the probability that the time between two successive failures of this equipment will be at least 20 h?

The problem is one of finding the area under the curve beyond 20 h (Figure 44.10). Table C in Appendix II gives the area under the curve beyond any particular ratio X/μ. In this problem,

$$\frac{X}{\mu} = \frac{20}{100} = 0.20$$

and from Table C in Appendix II the area under the curve beyond 20 h is thus 0.8187. The probability that the time between two successive failures is greater than 20 h is 0.8187; i.e., there is about an 82 percent chance that the equipment will operate without failure continuously for 20 or more hours. Similar calculations would give a probability of 0.9048 for 10 or more hours. In Section 48, Reliability Concepts and Data Analysis, this probability is calculated for the specified mission time of a product, and the result is called *reliability.* These analyses also could be made using exponential probability paper.

Example: The Relationship Between Part and System Reliability: It is often assumed that the probability of survival P_s (the system reliability) is the product of the individual reliabilities of the n parts within the system: $P_s = P_1P_2\ldots P_n$. This is known as the *product rule.* The formula assumes (1) that the failures of any part will cause failure of the system and (2) that the reliabilities of the parts are independent of each other, i.e., that the reliability of one part is not dependent on the reliability of another part. [Evans (1966) gives a good discussion of this and other assumptions in reliability calculations.] A set of lights in series on a Christmas tree demonstrates the product rule. These assumptions are usually not 100 percent correct. However, the formula is a convenient approximation that should be refined as information becomes available on the interrelationships of parts and their relationship to the system. [The redundancy formula (see below under Example: Redundancy) is an example of this.] I will now illustrate the product rule.

Suppose that the following reliability requirements have been set on the subsystems of a communications system:

Subsystem	Reliability (for a 4-h period), %
Receiver	0.970
Control system	0.989
Power supply	0.995
Antenna	0.996

What is the expected reliability of the overall system if the preceding requirements are met?

$$P_s = (0.970)(0.989)(0.995)(0.996) = 0.951$$

The chance that the overall system will perform its function without failure for a 4-h period is 95 percent.

If it can be assumed that each part follows the exponential distribution, then

$$P_s = e^{-t_1\lambda_1}e^{-t_2\lambda_2}\cdots e^{-t_n\lambda_n}$$

Further, if t is the same for each part,

$$P_s = e^{-t\Sigma\lambda}$$

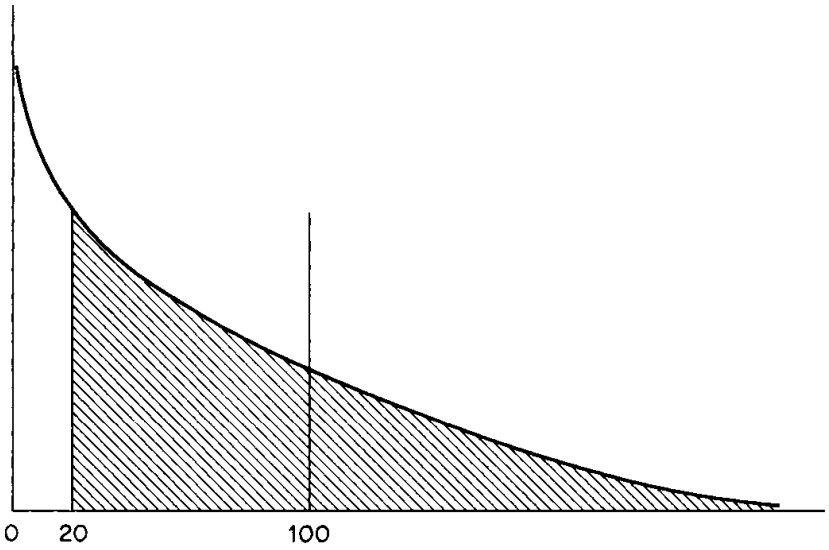


FIGURE 44.10 Distribution of time between failures.

Thus, when the failure rate is constant (and therefore the exponential distribution applies), a "reliability prediction" of a system can be made based on the addition of the part failure rates. This is illustrated in Section 19, under Designing for Reliability, Maintainability, and Availability.

Example: Redundancy: A number of system designs have been devised that attempt to increase system reliability by introducing redundancy. The simplest such system is the *parallel system,* which operates if at least one of its components operates. If each component has reliability $R(t_0)$, then the parallel system consisting of m components has reliability equal to [see Dudewicz (1976, p. 39)]:

$$1 - [1 - R(t_0)]^m$$

which is greater than $R(t_0)$.

Many other designs have been introduced, with a view both to reliability and to cost, such as the k-out-of-n systems. For details on these and other aspects of reliability, see Zacks (1983), Ireson (1982), and Lloyd and Lipow (1982).

The Weibull Distribution. The Weibull distribution is a family of distributions having the general density function

$$y = \alpha\beta(X - \gamma)^{\beta-1}\, e^{-\alpha(X-\gamma)^{\beta}}$$

where α = scale parameter, β = shape parameter, and γ = location parameter.

The curve of the function (Figure 44.7) varies greatly depending on the numerical values of the parameters. Most important is the shape parameter β, which reflects the pattern of the curve. Note that when β is 1.0, the Weibull function reduces to the exponential and that when β is about 3.5 (and $\alpha = 1$ and $\gamma = 0$), the Weibull closely approximates the normal distribution. In practice, β varies from about $^1/_3$ to 5. The scale parameter α is related to the peakedness of the curve; i.e., as α changes, the curve becomes flatter or more peaked.

The location parameter γ is the smallest possible value of X. This is often assumed to be zero, thereby simplifying the equation. It is often unnecessary to determine the values of these parameters because predictions are made directly from Weibull probability paper. King (1971, pp. 136–140) gives procedures for graphically finding α, β, and γ.

The Weibull covers many shapes of distributions. This makes it popular in practice because it reduces the problem of examining a set of data and deciding which of the common distributions (e.g., normal or exponential) fits best. In particular, both IFR (increasing failure rate) and DFR (decreasing failure rate) cases are included, respectively, with $\beta > 1$ and $\beta < 1$ [see Dudewicz (1976, pp. 88–89)].

Model Leading to a Weibull Distribution. It can be shown that a Weibull distribution arises if an exponential variable is raised to a power; i.e., if Y is exponential, then $Y^{1/\beta}$ has a Weibull distribution (Dudewicz 1976, p. 89).

Predictions with Weibull Distributions. An analytical approach for the Weibull distribution (even with tables) is cumbersome, and the predictions are usually made with Weibull probability paper. For example, five heat-treated shafts were stress tested until each of them failed. The fatigue life (in terms of number of cycles to failure) is

10,263
12,187
16,908
18,042
23,271

The problem is to predict the percentage failure of the population for various values of fatigue life. The solution is to plot the data on Weibull paper, observe if the points fall approximately in a straight line, and if so, read the probability predictions (percentage failure) from the graph.

Although Weibull plotting can follow the mean rank procedure of normal probability paper (see below under Predictions with Normal Distributions), much of the literature on Weibull applications uses "median ranks." Table D in Appendix II gives, for various sample sizes, the values of the median rank. (Note that the mean rank procedure does not require a table.) The median ranks necessary for this particular example are based on a sample size of five failures and are as shown in Table 44.18. (The mean rank estimates are shown for comparison.) The cycles to failure are now plotted on the Weibull graph paper against the corresponding values of the median rank (see Figure 44.11). These points fall approximately in a straight line [King (1971, pp. 126–128) describes how to modify a plot to help obtain a straight line], so it is assumed that the Weibull distribution applies. The vertical axis gives the cumulative percentage of failures in the population corresponding to the fatigue life shown on the horizontal axis. For example, about 50 percent of the population of shafts will fail in less than 17,000 cycles. About 90 percent of the population will fail in less than 24,000 cycles. By appropriate subtractions, predictions can be made of the percentage of failures between any two fatigue life limits.

It is tempting to extrapolate on probability paper, particularly to predict life. For example, suppose the minimum fatigue life were specified as 8000 cycles and the five measurements above were from tests conducted to evaluate the ability of the design to meet 8000 cycles. Since all five tests

TABLE 44.18 Table of Median and Mean Ranks

Failure number i	Median rank	Mean rank $= \frac{i}{5+1}$
1	0.1294	0.1667
2	0.3147	0.3333
3	0.5000	0.5000
4	0.6853	0.6667
5	0.8706	0.8333

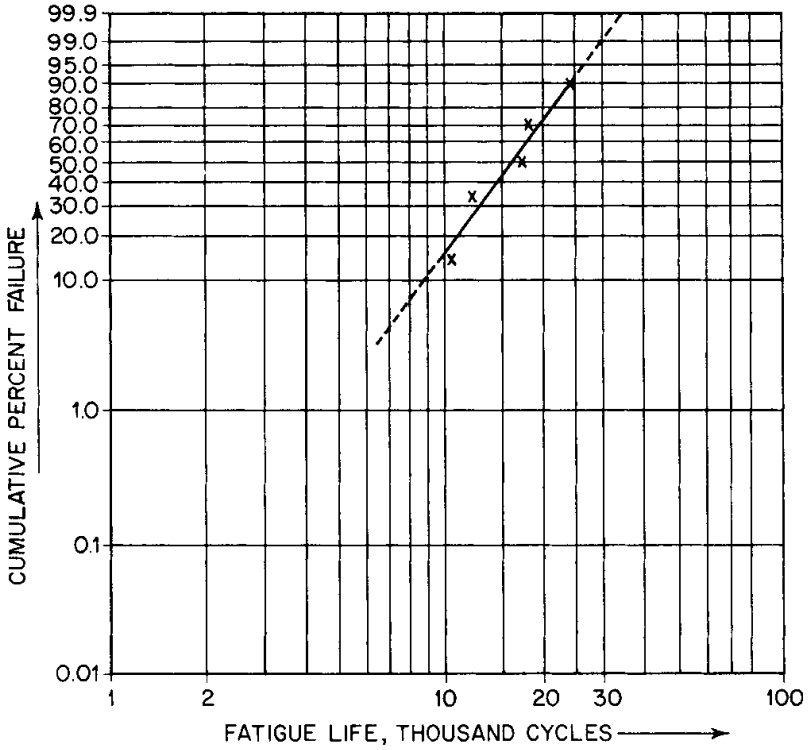


FIGURE 44.11 Distribution of fatigue life on Weibull probability paper.

exceeded 8000 cycles, the design seems adequate and therefore should be released for production. However, extrapolation on the Weibull paper predicts that about 8 percent of the *population* of shafts would fail in less than 8000 cycles. This suggests a review of the design before release to production. Thus the small *sample* (all *within* specifications) gave a deceiving result.

Extrapolation can go in the other direction. Note that a probability plot of life-test data does not require that all tests be completed before the plotting starts. As each unit fails, the failure time can be plotted against the median rank. If the early points appear to be following a straight line, then it is tempting to draw in the line *before* all tests are finished. The line can then be extrapolated beyond the actual test data, and life predictions can be made without accumulating a large amount of test time. The approach has been applied to predicting, *early in a warranty period,* the "vital few" components of a complex product that will be most troublesome. However, extrapolation has dangers. It requires the judicious melding of statistical theory and engineeirng experience and judgment.

Moult (1963) describes the use of a Weibull plot in comparing the suitability of two types of steel for use in bearings. The plot is shown in Figure 44.12. Nelson (1982) discusses Weibull paper. Probability graph paper is available for the normal, exponential, Weibull, and other probability distributions. (A source is TEAM, Technical and Engineering Aids for Management, Box 25, Tamworth, NH 03886.) Although the mathematical functions and tables provide the same information, the graph paper reveals *relationships* between probabilities and values of X that are not readily apparent from the calculations. For example, the reduction in percentage defective in a population as a function of wider and wider tolerance limits can be easily portrayed by the graph paper.

The Normal Distribution. Many engineering characteristics can be approximated by the normal distribution:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2}$$

where $e = 2.718$, $\pi = 3.141$, μ = population average, σ = population standard deviation.

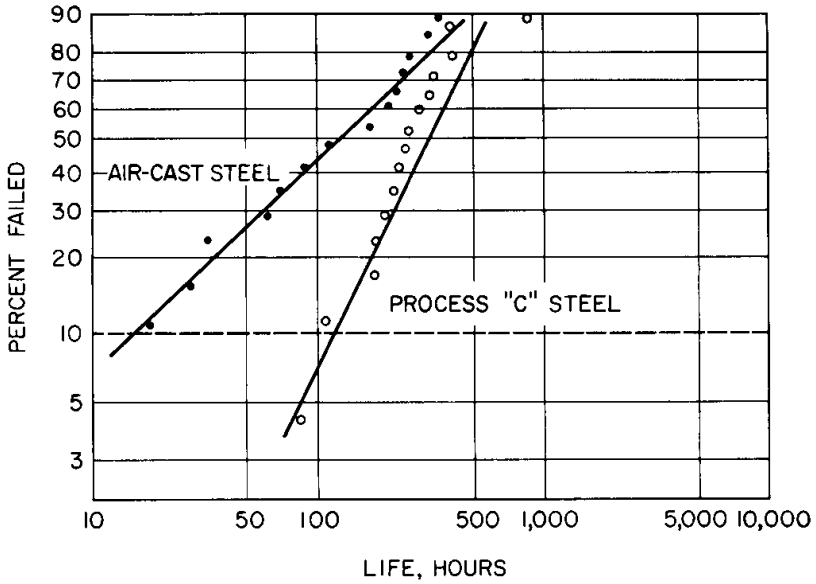


FIGURE 44.12 Composite fatigue endurance—process "C" vacuum degassing versus air cast AISI 8620.

Problems are solved with a table, but note that the distribution requires estimates of only the average μ and standard deviation σ of the population (unless otherwise indicated, Greek symbols will be used for population values and Roman symbols for sample values) in order to make predictions about the population. The curve for the normal probability distribution is related to a frequency distribution and its histogram. As the sample becomes larger and larger and the width of each cell becomes smaller and smaller, the historgram approaches a smooth curve. If the entire population (in practice, the population is usually considered infinite, e.g., the potential production from a process) were measured, and if it were normally distributed, the result would be as shown in Figure 44.7. Thus the shape of a histogram of sample data provides some indication of the probability distribution for the population. If the histogram resembles the "bell" shape shown in Figure 44.7, this is a basis for assuming that the population follows a normal probability distribution. (It is not necessary that the sample histogram be perfectly normal—the assumption of normality is applied only to the population, and small deviations from normality are expected in random samples.) Hahn (1971) gives a practical discussion of assuming normality. (The name *normal* distribution dates back to a time when all other distributions were erroneously thought to be abnormal. Today, some prefer the name *Gaussian* distribution.)

Model Leading to a Normal Distribution (Additive Errors, Central Limit Theorem). It can be shown that if a variable Y is the result of adding many other variables and those variables are not highly dependent on each other, then Y will have approximately a normal distribution. [This result is called the *central limit theorem*; see Dudewicz (1976, p. 149).] Statisticians usually recommend that 10 or more terms be added before this result is relied on to produce normality; however, a number of applied studies have shown good results with as few as three terms being added.

Predictions with Normal Distributions. Predictions require just two estimates and a table. The estimates are

$$\text{Estimate of } \mu = \overline{X} \qquad \text{and} \qquad \text{estimate of } \sigma = s$$

The calculations of the sample $\overline{X}$ and s are made by one of the methods previously discussed.

For example, from past experience, a manufacturer concludes that the burnout time of a particular light bulb it manufactures is normally distributed. A sample of 50 bulbs has been tested and the average life found to be 60 days, with a standard deviation of 20 days. How many bulbs in the entire population of light bulbs can be expected to be still working after 100 days of life?

The problem is to find the area under the curve beyond 100 days (see Figure 44.13). The area under a distribution curve between two stated limits represents the probability of occurrence. Therefore, the area beyond 100 days is the probability that a bulb will last more than 100 days. To find the area, calculate the difference between a particular value X and the average of the curve in units of standard deviation:

$$K = \frac{X - \mu}{\sigma}$$

In this problem, $K = (100 - 60) \div 20 = +2.0$. Table B in Appendix II shows, for $K = 2$, a probability of 0.9773. Applied to this problem, the probability that a bulb will last 100 days or less is 0.9773. The normal curve is symmetrical about the average, and the total area is 1.000. The probability of a bulb's lasting more than 100 days then is 1.0000 − 0.9773, or 0.0227, or 2.27 percent of the bulbs in the population will still be working after 100 days.

Similarly, if a characteristic is normally distributed, and if estimates of the average and standard deviation of the population are obtained, this method can estimate the total percentage of production that will fall within engineering specification limits.

Figure 44.14 shows representative areas under the normal distribution curve (these can be derived from Table B in Appendix II). Thus 68.26 percent of the *population* will fall between the average of the population plus or minus 1 standard deviation of the population, 95.46 percent of the population will fall between the average $\pm 2\sigma$, and finally, $\pm 3\sigma$ will include 99.73 percent of the population. The percentage of a *sample* within a set of limits can be quite different from the percentage within the same limits in the population. This important fact is crucial in testing hypotheses (covered later in this section).

Another way of making predictions based on a normal distribution employs probability paper. Probability paper is so constructed that data from a particular kind of distribution plots as a straight line; i.e., a sample of data from a normally distributed population plots approximately as a straight line on normal probability paper. (Small deviations from a straight line are expected because the data represent a *sample* of the population.) The following are the steps taken to plot a set of individual observations on probability paper:

1. Arrange the observations in ascending values. The smallest value is given a rank i of 1 and the largest value a rank of n.
2. For each value, calculate the cumulative frequency.

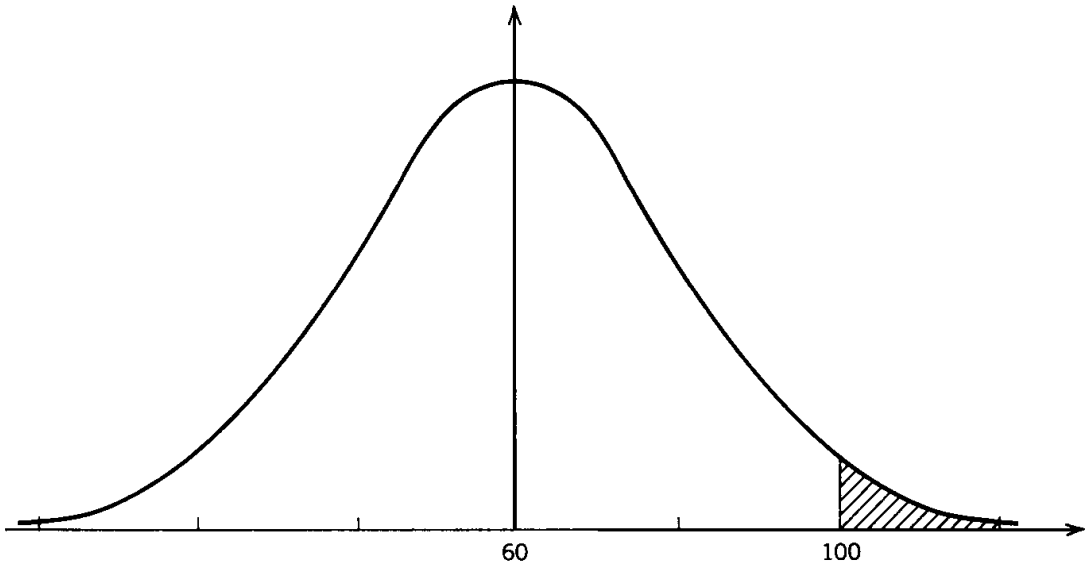


FIGURE 44.13 Distribution of light bulb life.

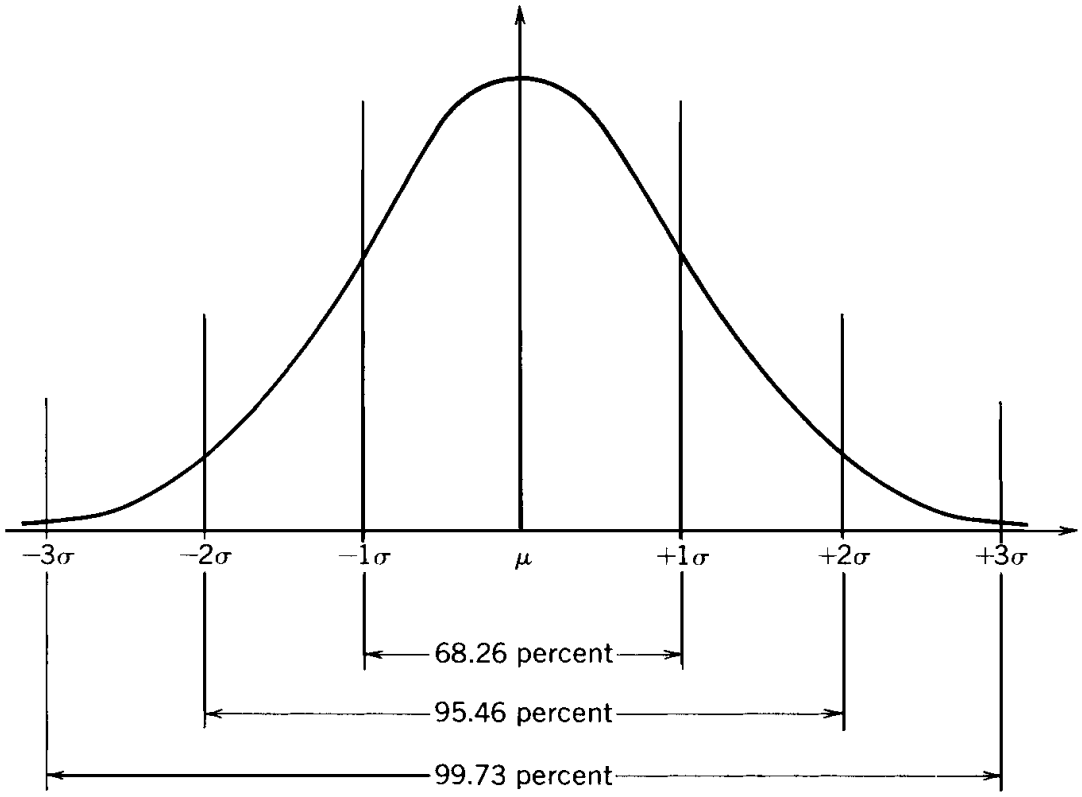


FIGURE 44.14 Areas of a normal curve.

3. For each value, calculate

$$\frac{\text{Cumulative frequency}}{n+1} \times 100$$

This provides the mean rank probability estimate, in percent, for plotting the data.

4. Plot the observed values against their mean rank probability estimate.

If the observations are in frequency distribution form, the procedure is the same, except that instead of using the observed values, the probability estimates are plotted against the cell boundaries. This is illustrated for the resistance data (see Table 44.19).

The plot is shown in Figure 44.15. Lower cell boundaries are plotted against the last column of Table 44.19 using the upper (Percent Over) scale. The line has been drawn in by eye, and the fit appears reasonable. This line represents an estimate of the population, and predictions like those obtained from the normal probability table can be read directly from the graph. For example, 5 percent of the population of coils will have resistance values greater than about 3.39. Also, 95 percent will have values greater than about 3.29. (Therefore, 95 − 5, or 90 percent, will have values between 3.29 and 3.39.)

Figure 44.16 shows a form that incorporates probability paper plotting with further analysis such as confidence limits and control limits. (This type of form was originally developed by E. F. Taylor.)

King (1971) gives a practical description of probability paper procedures for the normal and other important distributions. While fit is often evaluated by eye when we have clearly good fit (as in Figure 44.15), statistical tests are available and should be used in cases that are not so clearcut; see Iman (1982) for details and graphs on which this analysis can be performed.

The Lognormal Distribution. If $Y = e^Z$, where Z has a normal distribution, Y is said to have a *lognormal distribution* (since the logarithm of Y has a normal distribution).

TABLE 44.19 Resistance Data

Cell boundaries	Frequency	Cumulative frequency	$\frac{\text{Cumulative frequency}}{100+1}(100)$
3.415–3.445	1	1	0.99%
3.385–3.415	8	9	8.90
3.355–3.385	27	36	35.60
3.325–3.355	36	72	71.30
3.295–3.325	23	95	94.10
3.265–3.295	5	100	99.00
	100		

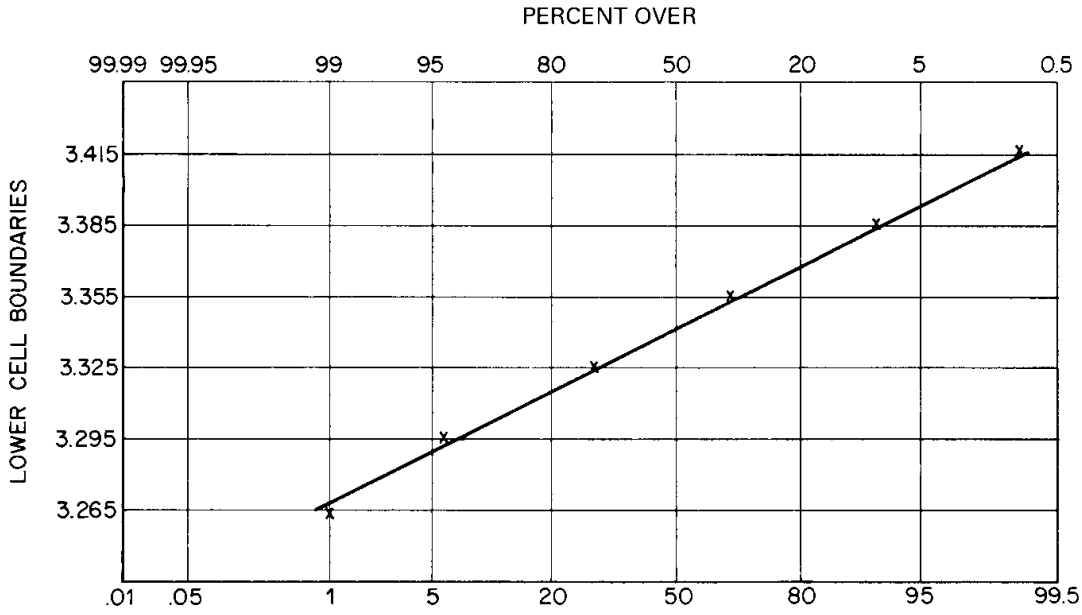


FIGURE 44.15 Cumulative probability plot of Table 44.19.

Model Leading to a Lognormal Distribution (Multiplicative or Percentage Effects). Lognormal variables arise when effects are percentages or multiplicative, which is common in biological and many other applications. For example, if a random percentage of a stock of items goes bad each time period, the percentage still good after a large number (10 or more) of time periods will be lognormally distributed. (This result follows from the central limit theorem already discussed when considering the normal distribution, since the logarithm of a product is the sum of the logarithms of the items in the product.)

Predictions with Lognormal Distributions. As with the normal distribution, predictions require two estimates. For details, see Cohen and Whitten (1981). Note that the lognormal distribution is positively skewed and is widely employed as a model for distribution of life spans, reaction times, income distributions, and other economic data.

Often the mean, standard deviation, and probabilities of the lognormal variable Y itself are of basic interest; thus, while one can easily estimate these for Z (the logarithm of $Y = e^Z$), this does not answer the real problem. For example, if Y is the lifetime of some system, one may want to estimate the average life of the system—not the mean of the logarithm of Y. This is why the special methods referred to above have been developed for this distribution.

Mixture Distributions. Y is said to have a *mixture distribution* if Y results from source i a percentage $100p_i$ of the time ($i = 1, 2,\ldots$).

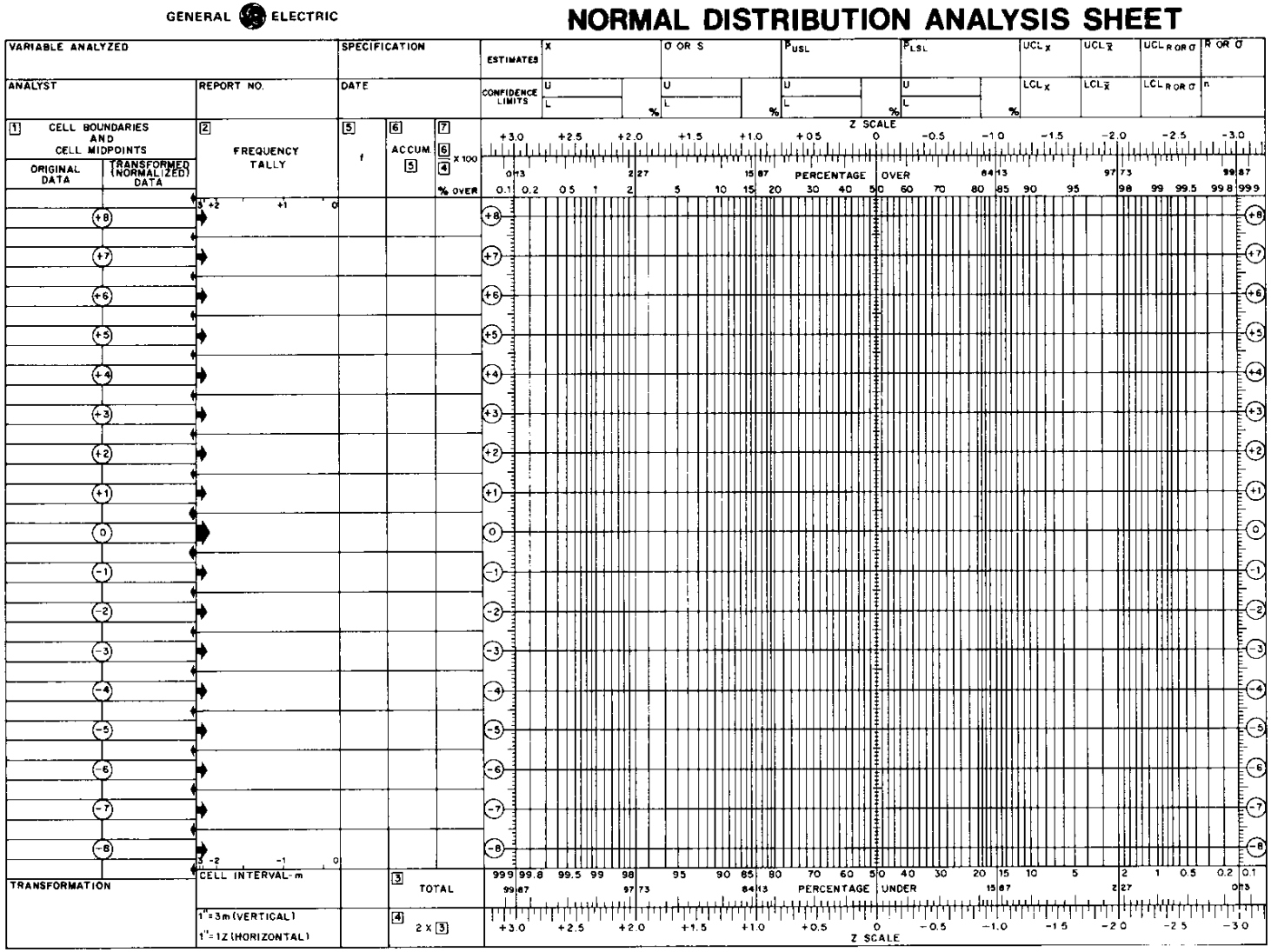


FIGURE 44.16 Normal distribution analysis sheet.

Model Leading to a Mixture Distribution. When output from several sources is mixed (e.g., output from several suppliers, several plants, several machines, several workers, and so on), the quality characteristics of the resulting mix have mixture distributions. If each of the components coming into the mix has *exactly* the same distribution, then the mix also will have that distribution. However, if the components coming into the mix have different distributions, then the mix will have a *mixture distribution.*

Fitting of Mixtures. As can be seen from Figure 44.7, the distributions we have considered so far are unimodal (have one peak). When in practice one sees two or more peaks in the histogram, one suspects that a mixture underlies the data. In some cases, this itself leads to a study of the items coming into the mix, often to find a problem in one of the streams of what should be homogeneous product.

In other cases, the streams coming into the mix cannot be separated (or it is not desired to separate them), but rather one wishes to fit the density of Y as $p_1 f(Z_1) + p_2 f(Z_2) + \ldots + p_k f(Z_k)$ for some k (2 or more). Here the p_i's add to 1 (100 percent of the mix), and often the Z_i's are known to be normal. To fit the distribution of the Y, one must then select $k - 1$ p's (since they add to 1, the last one is then determined), k means, and k variances. This process requires use of modern computer software such as LABONE. As an example, consider the data of Table 44.20. A histogram of these data shows that a mixture (of two terms, since there are two peaks) may be involved. The distributions seem to the eye to be normal. Using LABONE Expert Statistical Programs (ESP), we are able to easily fit a mixture of normal distributions.

The Multinormal Distribution. The continuous probability distributions discussed up to this point (the uniform, exponential, Weibull, normal, lognormal, and mixture) all relate to situations that are *univariate,* that is, where the outcome of interest has one component (such as lifetime). If there are additional components of interest (such as weight and height), then the outcome is *multivariate* (in this case of three, *trivariate*). The multinormal distribution is appropriate when each of the components has a normal distribution and is the most widely used continuous multivariate distribution.

Model Leading to a Multinormal Distribution. The same sort of additive process that leads to a univariate normal distribution leads to a multivariate normal distribution when more than one component is being measured.

Predictions with Multinormal Distributions. Predictions with multinormal distributions require computer packages in most cases. The details, with computer code and examples, are discussed by Siotani et al. (1985).

The Extended Generalized Lambda Distribution. A one-parameter lambda distribution was proposed in 1960 by J. Tukey, generalized in 1972 and 1974 by J. S. Ramberg and B. Schmeiser,

TABLE 44.20 Data for Mixture Analysis

3.37	3.34	3.48	3.32	3.33	3.38	3.34	3.31	3.43	3.34
3.29	3.46	3.30	3.31	3.43	3.34	3.34	3.46	3.39	3.34
3.45	3.36	3.30	3.42	3.33	3.35	3.45	3.34	3.32	3.48
3.32	3.37	3.44	3.38	3.36	3.47	3.36	3.31	3.43	3.30
3.35	3.43	3.38	3.37	3.54	3.32	3.36	3.42	3.29	3.35
3.48	3.39	3.34	3.42	3.30	3.39	3.46	3.40	3.32	3.43
3.29	3.41	3.37	3.36	3.41	3.47	3.36	3.37	3.43	3.36
3.31	3.43	3.35	3.34	3.45	3.34	3.31	3.46	3.37	3.35
3.50	3.35	3.37	3.42	3.35	3.36	3.48	3.35	3.31	3.44
3.35	3.36	3.49	3.31	3.31	3.40	3.35	3.33	3.45	3.31

and developed and tabled by Ramberg et al. (1979). While it fits a wide variety of curve shapes and can fit any mean and any variance, it cannot fit all combinations of skewness and kurtosis. Thus it is useful when it fits, but for some data sets it does not fit. Recently, Karian et al. (1996) gave an extension called the *extended generalized lambda distribution* (EGLD), which can fit any and all values of the mean, variance, skewness, and kurtosis; tables needed in applications were given in Dudewicz and Karian (1996). Thus now the benefits of this empirical family are available in all univariate data sets. These methods are extended to bivariate GLD cases in Karian and Dudewicz (1999).

Selecting a Continuous Distribution. Selection of which continuous distribution to use is usually made either by knowledge of the underlying situation or by fitting a model to the histogram (often via plots on probability papers of the most usual distributions). In either case, a test of the model selected is desirable to check its validity.

Selection from a Model of Reality. In many cases one will know (or assume) that the model that leads to one of the distributions we have discussed underlies the practical situation. For example, if the life distribution of the equipment under study has, in the past, always been adequately fitted by a Weibull model (though with parameters that change from application to application), one will usually start with a Weibull assumption.

Testing Distributional Assumptions (Probability Plotting, Tests for Specific Distributions). In practice, a distribution is assumed by evaluating a sample of data. Often it is sufficient to evaluate the shape of the histogram or the degree to which a plot on probability paper follows a straight line. These convenient methods do require judgment (e.g., how "straight" must the line be?) because the sample is never a perfect fit; quantitative tests for probability plots should be used (see Iman 1982). Be suspicious of the data if the fit is "perfect." "Goodness of fit" tests (see Tests of Hypotheses later in this section) evaluate any distribution assumption using quantitative criteria.

Fitting Empirical Probability Distributions. If there is little or no reason to suggest one of the specific models discussed, or if they are rejected (e.g., because of a poor probability paper fit), an alternative is to fit an empirical model. Such models can adapt to a wide range of distributional shapes, including many of those of the specific models discussed above. One of the most widely used empirical families has been the *generalized lambda distribution* (GLD) family. However, as discussed earlier, the new EGLD enhances its capabilities and should be used instead; often the two give the same result, but when the GLD has difficulty, the EGLD can still match the population's sample moments.

As an example, Table 44.21 presents data (from p. 219 of Hahn and Shapiro, 1967) on the coefficient of friction of a metal in 250 samples. Using procedures and tables given in Ramberg et al. (1979), a probability function can be developed.

TABLE 44.21 Coefficient-of-Friction Data

Range	Frequency
Less than .015	1
0.015–0.020	9
0.020–0.025	30
0.025–0.030	44
0.030–0.035	58
0.035–0.040	45
0.040–0.045	29
0.045–0.050	17
0.050–0.055	9
0.055–0.060	4
More than 0.060	4
Total	250

As a check on the goodness of the fit, it is recommended that the probability function always be plotted on the same graph as the histogram for a visual assessment of the fit (which can be supplemented by a chi square test if desired). This is done in Figure 44.17, and we see that the fit is excellent (a chi square test comes to the same conclusion).

Bootstrap Methods. Bootstrap methods suggest that one fit a model such as the EGLD to the data, and then, assuming that the fit passes testing, use that model in a "bootstrapping" analysis. (See The Generalized Bootstrap and Bootstrap Method later in this section.)

STATISTICAL ESTIMATION

In statistical estimation, we make inferences about parameters of a population from data on a sample. For example, if we have a random sample of 100 items from a large lot, information on the sample can be used to infer information about the proportion of defectives p in the lot. This inference takes the form of either a single number (a *point estimate*) or a pair of numbers (an *interval estimate*); there are several types of interval estimates, depending on our goals. If we found that 15 of the 100 items in the sample were defective, we would estimate p as being $15/100 = 0.15$ (point estimate); a typical interval would be to state that we are 95 percent confident that p is between 0.08 and 0.22 (confidence interval estimate). Thus a *confidence interval* sets limits on the unknown parameter, here the proportion p. Two other types of intervals often needed are *prediction intervals* and *tolerance intervals.* Let X denote the number of defectives in a future sample; a prediction interval sets limits on X, such as

$$P[L_1 \le X \le U_1] = 0.95$$

In this example, a *prediction interval* would state that X would be between 5 and 25; these are limits within which one could be 95 percent sure the number of defectives in a future sample of 100 items would lie. A *tolerance interval* sets limits (L, U) such that one can be 95 percent sure that at

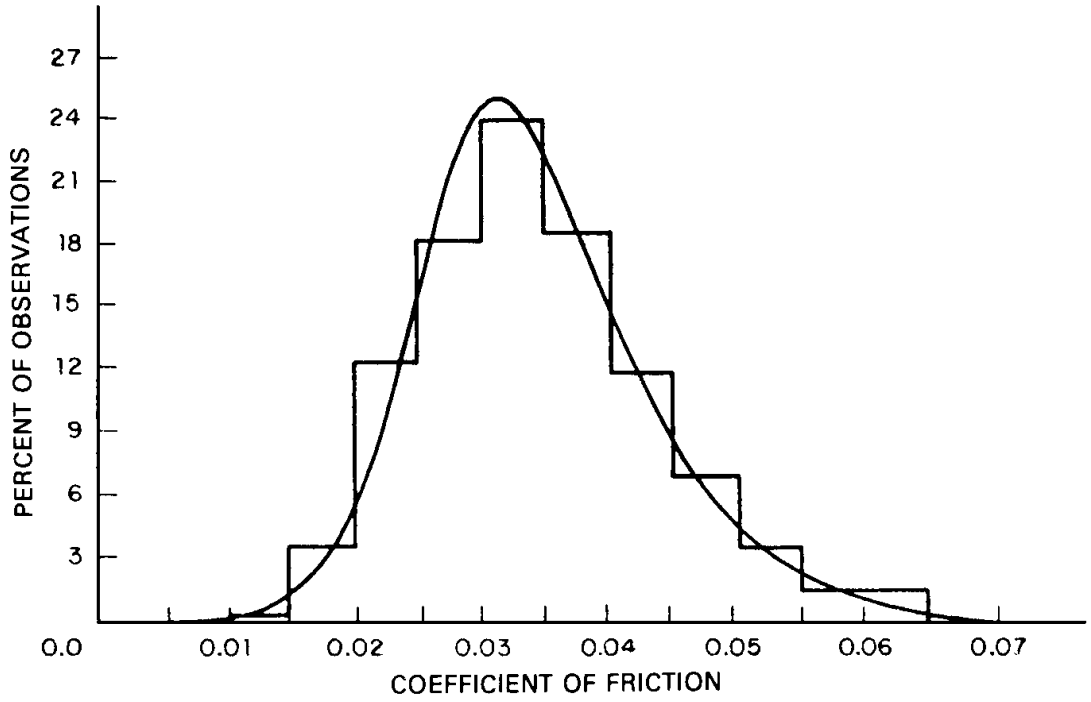


FIGURE 44.17 Coefficient of friction relative frequency histogram and the fitted distribution.

least 99 percent of the population will be included within these limits. In this example, the population is all lots of 100 items drawn from the same process; at least 99 percent of the lots will have a proportion of defectives between $L = 0.05$ and $U = 0.25$. These items are discussed in more detail below.

Point Estimates. Point estimates are customarily the points at which the interval estimates are centered. In many cases, it is preferable to give an interval estimate, since that estimate tells us how much uncertainty is associated with the estimate. For example, if we observe that 15 of 100 items chosen at random from a very large lot are defective, then we will estimate the proportion of defectives in the lot as 15/100 = 0.15. Similarly, if we observe three defectives in a sample of 20 items, we will estimate the proportion of defectives as 3/20 = 0.15. However, in the first case the interval estimate (at 95 percent confidence) will be that p is between 0.08 and 0.22, while in the latter case the interval estimate (at 95 percent confidence) will be that p is between 0.00 and 0.31. In either case, if forced to estimate proportion defective in the lot by a single number, that number would be 0.15; however, the uncertainty in that estimate is much greater with the smaller sample size (where as much as 31 percent of the population might be defective) than with the larger sample size (where as much as 22 percent of the population might be defective). The typical point estimates are covered below, as the centers of the respective intervals.

Confidence Interval Estimates. *Estimation* is the process of analyzing a sample result in order to predict the corresponding value of the population parameter. For example, a sample of 12 insulators has an average impact strength of 4.952 ft·lb (6.7149 N·m). If this is a representative sample from the process, what estimate can be made of the true average impact strength of the entire population of insulators?

1. The *point estimate* is a single value used to estimate the population parameter. For example, 4.952 ft·lb (6.7149 N·m) is the point estimate of the average strength of the population.
2. The *confidence interval* is a range of values that includes (with a preassigned probability called *confidence level*) the true value of a population parameter. *Confidence limits* are the upper and lower boundaries of the confidence interval. Confidence level is the proportion of times in the long run that an assertion about the value of a population parameter is correct.

Duncan (1974) provides a thorough discussion of confidence limits. The explanation here indicates the concept behind the calculations.

If the population mean is μ, the probability that the sample mean will be between

$$\mu \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

is equal to 0.95:

$$P\left(\mu - 1.96 \frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

This is algebraically equivalent to saying that the sample mean plus 1.96 standard deviations of means lies above μ and the sample mean minus 1.96 standard deviations of means lies below μ:

$$P\left(\mu \le \overline{X} + 1.96 \frac{\sigma}{\sqrt{n}} \text{ and } \overline{X} - 1.96 \frac{\sigma}{\sqrt{n}} \le \mu\right) = 0.95$$

or

$$P\left(\overline{X} - 1.96 \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Or the 95 percent confidence interval on μ is $\overline{X} \pm 1.96\ (\sigma/\sqrt{n})$. Before the sample is taken, this interval has a 0.95 probability of including the population value: 95 percent of the set of such intervals would include the population value. In practice, this is interpreted to mean that there is a 95 percent "confidence" that the limits based on one sample will include the true value.

For the sample of 12 insulators, suppose that $\sigma = 0.25$. Then, the 95 percent confidence limits are

$$\overline{X} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 4.952 \pm 1.96 \frac{(0.25)}{\sqrt{12}} = 4.811 \text{ and } 5.093$$

This is interpreted to mean that there is 95 percent confidence that μ is between 4.811 and 5.093. The 95 percent is the confidence level (confidence levels of 90, 95, or 99 percent are usually assumed in practice, and some statisticians call these the "holy" numbers), and 4.811 and 5.093 are the limits of the confidence interval. A confidence level is associated with an assertion based on actual measurements and measures the proportion of times that the assertion will be true in the long run. Confidence limits are limits that include the true value with a preassigned degree of confidence (the confidence level).

Table 44.22 summarizes confidence limit formulas for common parameters. The following examples illustrate some of these formulas.

Example: Sixty-one specimens of brass have a mean hardness of 54.62 and an estimated standard deviation of 5.34. Determine the 95 percent confidence limits on the mean.

Solution:

$$\text{Confidence limits} = \overline{X} \pm t \frac{s}{\sqrt{n}}$$

$$= 54.62 \pm 2.00 \frac{5.34}{\sqrt{61}}$$

$$= 53.25 \text{ and } 55.99$$

There is 95 percent confidence that the true mean hardness of the brass is between 53.25 and 55.99.

Example: A radar system has been operated for 1200 h, during which time eight failures occurred. What are the 90 percent confidence limits on the mean time between failures for the system?

Solution:

$$\text{Estimated } m = \frac{1200}{8} = 150 \text{ h}$$

$$\text{Upper confidence limit} = \frac{2(1200)}{7.962} = 301.4$$

$$\text{Lower confidence limit} = \frac{2(1200)}{26.296} = 91.3$$

There is 90 percent confidence that the true mean time between failures is between 91.3 and 301.4 h. [Epstein (1960) discusses several cases of making estimates from life test data.]

Confusion has arisen on the application of the term *confidence level* to a reliability index such as mean time between failures. Using a different example, suppose the numerical portion of a reliability requirement reads as follows: "The MTBF shall be at least 100 h at the 90 percent confidence level." This means that

1. The minimum MTBF must be 100 h.

TABLE 44.22 Summary of Confidence Limit Formulas and Graphs

Parameters	Formulas
1. Mean of a normal population (standard deviation known)	$\overline{X} \pm K_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ where $\overline{X}$ = sample average K = normal distribution coefficient (see Appendix II, Table B) σ = standard deviation of population n = sample size
2. Mean of a normal population (standard deviation unknown)	$\overline{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$ where t = Student's distribution coefficient (with n − 1 degrees of freedom) (see Appendix II, Table G) s = estimated σ
3. Standard deviation of a normal population *a.* Using sample standard deviation *b.* Using sample range	Upper confidence limit = $B_U s$ Lower confidence limit = $B_L s$ where B_U and B_L are numerical factors given in Natrella (1963, p. T-34), Dixon and Massey (1969, p. 140)
4. Population fraction defective based on attribute data (fraction defective in sample)	See Appendix II, Chart N
5. Population fraction defective based on variables data ($\overline{X}$ and s in sample)	See Kirkpatrick (1970)
6. Difference between the means of two normal populations (standard deviations σ_1 and σ_2 known)	$(\overline{X}_1 - \overline{X}_2) \pm K_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ where K = normal distribution coefficient (see Appendix II, Table B)
7. Difference between the means of two normal populations ($\sigma_1 = \sigma_2$ but unknown)	$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \times \sqrt{\frac{\sum (X - \overline{X}_1)^2 + \sum (X - \overline{X}_2)^2}{n_1 + n_2 - 2}}$ where t = Student's distribution coefficient (with degrees of freedom = $n_1 + n_2 - 2$)(see Appendix II, Table G)
8. Difference between the means of two normal populations (σ_1 and σ_2 both unknown)	$(\overline{X}_1 - \overline{X}_2) \pm t^*_{\alpha/2}\sqrt{\frac{\sum (X - \overline{X}_1)^2}{n_1(n_1 - 1)} + \frac{\sum (X - \overline{X}_2)^2}{n_2(n_2 - 1)}}$ where t^* = Student's distribution coefficient (with the smaller of $n_1 - 1$ and $n_2 - 1$ as degrees of freedom)

TABLE 44.22 Summary of Confidence Limit Formulas and Graphs (***Continued***)

Parameters	Formulas
9. Mean time between failures (based on an exponential population of time between failures)	Upper confidence limit = $2T/\chi^2_{\alpha/2}$ Lower confidence limit = $2T/\chi^2_{1-\alpha/2}$ where T = total test time on all units and DF = $2r$, where r is a preassigned number of failures and where χ^2 = chi square distribution coefficient (see Appendix II, Table L)
10. Reliability (based on a Weibull population)	See Thoman, Bain, and Antle (1970)
11. Availability (based on an exponential population of time between failures and log normal population of repair time)	See Gray and Lewis (1967)

2. Actual tests shall be conducted on the product to demonstrate with 90 percent confidence that the 100-h MTBF has been met.

3. The test data shall be analyzed by calculating the observed MTBF and the lower one-sided 90 percent confidence limit on MTBF.

4. The lower one-sided confidence limit must be ≥100 h.

The term *confidence level,* from a statistical viewpoint, has great implications on a test program. Note that the observed MTBF must be *greater than* 100 if the lower confidence limit is to be ≥ 100. Confidence level means that sufficient tests must be conducted to demonstrate, with statistical validity, that a requirement has been met. Confidence level does *not* refer to the qualitative opinion about meeting a requirement. Also, confidence level does *not* lower a requirement; i.e., a 100-h MTBF at a 90 percent confidence level does not mean that 100 h is desired but that 0.90 × 100, or 90 h, is acceptable. Such serious misunderstandings have occurred. When the term is used, a clear understanding should be verified and not assumed.

Determination of the Sample Size Required to Achieve a Specified Precision in an Estimate. Additional tests will increase the precision of the estimates obtained from a test program. The increase in precision usually does not vary linearly with the number of tests—doubling the number of tests usually does not double the precision (even approximately). Further, if the sample is selected randomly and if the sample size is less than 10 percent of the population size, then precision depends primarily on the absolute size of the sample rather than on the sample size expressed as a percentage of the population size. Thus a sample size that is 1 percent of a population of 100,000 may be more precise than a 10 percent sample from a population of 1000 (see Hahn 1972).

The cost of additional tests must be evaluated against the value of the additional precision. Confidence limits can help to determine the size of a test program required to estimate a product characteristic within a specified precision. Suppose it is desired to estimate the true mean life of a battery. The estimate must be within 2.0 h of the true mean if the estimate is to be of any value. The variability is known as $\sigma = 10.0$. A 95 percent confidence level is desired on the confidence statement. The 2.0 h is the desired confidence interval half-width, so

$$2.0 = \frac{(1.96)(10)}{\sqrt{n}} \qquad n = 96$$

A sample of 96 batteries will provide a mean that is within 2.0 h of the true mean (with 95 percent confidence). Notice the type of information required: (1) desired width of the confidence interval

(the precision desired in the estimate), (2) confidence level desired, and (3) variability of the characteristic under investigation. The number of tests required cannot be determined until the engineer furnishes these items of information.

Table 44.23 summarizes formulas and graphs useful in determining the sample size required to estimate a population parameter with a specified precision. The following examples illustrate some of the formulas.

Example: A sample must be selected to estimate the population mean length of a part. It appears reasonable to assume that length is normally distributed. An estimate of the standard deviation is not available, but process knowledge suggests that "almost all" production falls between 2.009 and 2.027 in. As a first approximation, the standard deviation is estimated as (2.027−2.009) divided by 6, or 0.003 in. It is desired that the estimate of μ be within 0.001 in of the true μ and that the estimation statement be made at the 95 percent confidence level. Referring to Appendix II, Chart S, $E/s = 0.001/0.003 = 0.33$, and the required sample size is about 37. It is instructive to calculate n for other values of E and s (see Table 44.24). Such a *sensitivity analysis* is helpful in evaluating the cost of extra tests against the value of extra precision.

Example: It is desired to estimate the standard deviation σ of a population within 20 percent of the true value at the 95 percent confidence level. Referring to Appendix II, Chart T, the required degrees of freedom is about 46 and, therefore, the sample size is 46 + 1, or 47.

TABLE 44.23 Summary of Sample Size Formulas and Graphs

Parameters	Formulas
1. Mean of a normal population (σ known)	$n = \dfrac{K^2_{\alpha/2}\sigma^2}{E^2}$ where K = normal distribution coefficient E = maximum allowable error in estimate (desired precision)
2. Mean of a normal population (σ estimated)	See Appendix II, Chart S
3. Standard deviation of a normal population	See Appendix II, Chart T
4. Fraction defective of a population	$n = p(1-p)\left(\dfrac{K_{\alpha/2}}{E}\right)^2$ where p = estimate of the population fraction defective. If no estimate of p is available, assume "worst case" of $p = 0.5$

TABLE 44.24 Effect of E and s on n

Maximum error E	Standard deviation s		
	0.002	0.003	0.004
0.0008	27	56	98
0.0010	18	37	64
0.0020	7	12	18

Prediction Intervals. A prediction interval is used when the desire is not an estimate of population characteristics directly but rather a prediction of what we will find when we take a future item from the population. For example, in the example of $n = 61$ specimens of brass with a mean hardness of 54.62 and an estimated standard deviation of 5.34, we previously found that 95 percent confidence limits on the mean were

$$\overline{X} + t\frac{s}{\sqrt{n}} = 53.25 \text{ and } 55.99$$

Now we ask: What limits can we be 95 percent sure the next item sampled will have its hardness within? The appropriate interval here is

$$\overline{X} + ts\sqrt{1 + \frac{1}{n}} = 54.62 \pm (2.00)(5.34)(1.0082)$$

$$= 43.85 \text{ and } 65.39$$

For further considerations and tables, see Hahn (1970*a,* 1970*b*).

Tolerance Intervals. *Statistical tolerance limits* are similar to *process capability*; i.e., they show the practical boundaries of process variability (see Section 22, under Operations) and therefore can be a valuable input in the determination of engineering tolerance limits (which specify the allowable limits for product acceptance). Methods for calculating statistical tolerance limits are of two types—those which assume a normal distribution and those which do not require any distributional assumption. Table 44.25 summarizes these methods.

Table 44.26 shows data I will use to illustrate these methods. Five samples of four each were taken and an outside dimension of a cathode pole recorded. A confidence level of 95 percent and a population percentage of 99 percent have been chosen.

Using method 1 and the standard deviation *s,* the statistical tolerance limits are

$$\overline{X} \pm ks = 1.00287 \pm 3.615(0.00034) = 1.00164 \text{ and } 1.00410$$

Using method 2 and the overall range R of the combined data, the limits are

$$\overline{X} \pm K_1R = 1.00287 \pm 1.005(0.00134) = 1.00152 \text{ and } 1.00422$$

Using method 3 and the average of the ranges R, the limits are

$$\overline{X} \pm K_2\overline{R} = 1.00287 \pm 1.783(0.00078) = 1.00148 \text{ and } 1.00426$$

These methods assume that the characteristic is normally distributed. Method 4 is "distribution-free" and assumes only that the distribution is continuous and the sample is a random one (these assumptions apply to all methods). The statistical tolerance limits by this method are simply the extreme observations in the combined sample, i.e., 1.00231 and 1.00365. Appendix II, Table W indicates that at least 78.4 percent of the population will be included within these limits. (Note that Appendix II, Table X provides the sample size required to include 99 percent of the population; i.e., a sample of 473 is needed to be 95 percent confident that the sample extremes would include 99 percent of the population.)

When it is feasible to assume a normal distribution, method 1 is preferred because it usually provides the narrowest set of limits while recognizing the variation in the sample. Methods 2 and 3 are good approximations. If normality cannot be assumed, then method 4 is appropriate but at the cost of a larger sample size. In practice, a partial sample can first be obtained to evaluate the assumption of normality. If normality can be assumed, the partial sample is then used to determine statistical tolerance limits. Otherwise, the full sample should be taken and the distribution-free approach (method 4) applied to determine the limits.

Tolerance intervals also have been developed for other cases, such as where the distribution is exponential. Ranganathan and Kale (1983) give such intervals that are resistant to the presence of an

TABLE 44.25 Methods for Calculating Statistical Tolerance Limits

Method	Distribution assumption	Formula for limits	Source of factor
1. Measure a sample of n items and calculate the average $\overline{X}$ and standard deviation s	Normal	Two-sided limits: $\overline{X} \pm ks$ One-sided limit: $\overline{X} + ks$ or $\overline{X} - ks$	Appendix II, Table V Appendix II, Table V
2. Measure a sample of n items and calculate the average $\overline{X}$ and range R	Normal	Two-sided limits: $\overline{X} \pm K_1 R$	Appendix II, Table U
3. Measure N samples of n items each and calculate the grand average $\overline{\overline{X}}$ and average range $\overline{R}$	Normal	Two-sided limits: $\overline{\overline{X}} \pm K_2 \overline{R}$	Bingham (1962, p. 37)
4. Define the population percentage P which must be included between the tolerance limits. Measure a sample of n and observe the largest and smallest values	None	Two-sided limits: The probability is γ that at least P % of the population will be between the sample extremes. One-sided limit: The probability is γ that at least P % of the population will be less than the largest value in the sample (or greater than the smallest value).	Appendix II, Table W Natrella (1963, p. T-76)

TABLE 44.26 Data on Cathode Pole Dimension

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	
1.00263	1.00306	1.00293	1.00291	1.00310	
1.00298	1.00328	1.00343	1.00247	1.00281	$\overline{X} = 1.00287$
1.00293	1.00274	1.00239	1.00268	1.00256	$s = 0.00034$ (standard deviation of the
1.00285	1.00303	1.00274	1.00365	1.00231	20 observations about $\overline{X}$)
$\overline{X}$ 1.00285	1.00303	1.00287	1.00293	1.00269	$\overline{R} = 0.00078$
R 0.00035	0.00054	0.00104	0.00118	0.00079	$R = 0.00134$

outlier in the observations. Their Section 5 includes an example with data on reliability of air conditioning on a Boeing 727 jet aircraft.

All the preceding methods involve two probabilities, i.e., a confidence level γ and the probability P of falling within limits. This is confusing, but these two probabilities are needed to obtain a mathematically correct statement concerning the limits. An approximation uses the sample average $\overline{X}$ and standard deviation s and regards these as highly reliable estimates of μ and σ. If normality is assumed, then 99 percent statistical tolerance limits are calculated as

$$\overline{X} \pm 2.58s$$

where the value of 2.58 is obtained from the normal distribution table (Appendix II, Table B):

$$\overline{X} \pm 2.58 = 1.00287 \pm 2.58\ (0.000274)$$

$$= 1.00216 \text{ and } 1.00358$$

These limits are then interpreted to mean that 99 percent of the population is within 1.00217 and 1.00357. Another approach sets the limits at simply $\overline{X} \pm 3s$. At best, these are only approximate because $\overline{X}$ and s are not exactly equal to μ and σ. Bingham (1962) discusses the $X \pm 3s$ approximation. (Confidence limit calculations could indicate the size of the possible error.)

Another approach to simplify the probability statement uses the Chebyshev inequality theorem, which holds for any continuous distribution. The theorem states that the probability of obtaining a value that deviates from μ by more than k standard deviations is less than $1/k^2$. For the limits to include 99 percent,

$$0.01 = \frac{1}{k^2}$$

or

$$k = 10$$

The 99 percent limits would be calculated at $\overline{X} \pm 10s$. These limits are distribution-free, and the prediction statement is simple; i.e., at least 99 percent of the population is within $\overline{X} \pm 10s$. However, the multiple of 10 is highly conservative and results in limits much wider than any of the other methods.

For the methods listed in Table 44.25, no provision is made for the division of the remaining $100(1 - P)$ percent between the upper and lower tails of the distribution. Owen and Frawley (1971) give a procedure and tables for setting limits that do provide for controlling the percentage outside each of the two limits.

Statistical tolerance limits are sometimes confused with other limits used in engineering and statistics. Table 44.27 summarizes the distinctions among five types of limits. Hahn (1970*a,* 1970*b*) gives an excellent discussion with examples and tables to illustrate the differences among several types of limits. Also see Harter (1983), under "Tolerance Limits."

Tolerance Limits for Interacting Dimensions. *Interacting dimensions* are those which mate or merge with other dimensions to create a final result. Setting tolerance limits on such dimensions is discussed in the following paragraphs. Setting tolerance limits on noninteracting dimensions makes use of the methods presented under Statistical Tolerance Limits in this section.

Conventional Method Relating Tolerances on Interacting Dimensions. Consider the simple mechanical assembly shown in Figure 44.18. The lengths of components *A, B,* and *C* are interacting dimensions because they determine the overall assembly length.

The conventional method of relating interacting dimensions is simple addition. For the example of Figure 44.18,

$$\text{Nominal value of the result} = \text{nominal value}_A + \text{nominal value}_B + \text{nominal value}_C$$

TABLE 44.27 Distinction among Limits

Name of limits	Meaning
Tolerance limits	Set by the engineering design function to define the minimum and maximum values allowable for the product to work properly.
Statistical tolerance limits	Calculated from process data to define the amount of variation that the process exhibits. These limits will contain a specified proportion of the total population.
Prediction limits	Calculated from process data to define the limits which will contain all of k future observations.
Confidence limits	Calculated from data to define an interval within which a population parameter lies.
Control limits	Calculated from process data to define the limits of chance (random) variation around some central value.

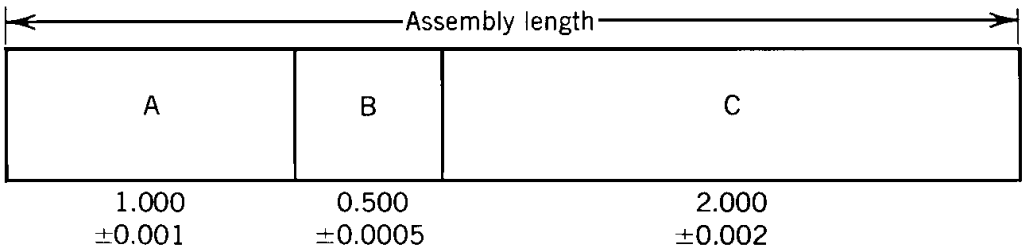


FIGURE 44.18 Mechanical assembly.

$$\text{Tolerance T of the result} = T_A + T_B + T_C$$

$$\text{Nominal value of assembly length} = 1.000 + 0.500 + 2.000 = 3.500$$

$$\text{Tolerance of assembly length} = 0.0010 + 0.0005 + 0.0020 = \pm 0.0035$$

This method assumes 100 percent interchangeability of components making up the assembly. If the component tolerances are met, then all assemblies will meet the assembly tolerance determined by the simple arithmetic addition.

The approach of adding component tolerances is mathematically correct but often too conservative. Suppose that about 1 percent of the pieces of component *A* are expected to be below the lower tolerance limit for component *A,* and suppose the same for components *B* and *C.* If a component *A* is selected at random, there is, on average, 1 chance in 100 that it will be on the low side, and similarly for components *B* and *C.* The key point is this: If assemblies are made at random, and if the components are manufactured independently, then the chance that an assembly will have all three components simultaneously below the lower tolerance limit is

$$\frac{1}{100} \times \frac{1}{100} \times \frac{1}{100} = \frac{1}{1{,}000{,}000}$$

There is only about one chance in a million that all three components will be too small, resulting in a small assembly. Thus, setting component and assembly tolerances based on the simple addition formula is conservative in that it fails to recognize the extremely low probability of an assembly containing all low (or all high) components.

Statistical Method of Relating Tolerances on Interacting Dimensions. This method states for the example Figure 44.18:

$$\text{Nominal value of the result} = \text{nominal value}_A + \text{nominal value}_B + \text{nominal value}_C$$

$$\text{Tolerance of the result} = \sqrt{T_A^2 + T_B^2 + T_C^2}$$

Then:

$$\text{Nominal value of the assembly} = 1.000 + 0.500 + 2.000 = 3.500$$

$$T \text{ of the assembly} = \sqrt{(0.001)^2 + (0.0005)^2 + (0.002)^2} = \pm 0.0023$$

Practically all (but not 100 percent) of the assemblies will fall within 3.500 ± 0.0023. This is narrower than 3.500 ± 0.0035 (the result by the arithmetic method).

In practice, the problem often is to start with a defined end result (e.g., assembly length specification) and set tolerances on the parts. Suppose the assembly tolerance was desired to be ±0.0035. Listed in Table 44.28 are two possible sets of component tolerances that when used with the quadratic formula will yield an assembly tolerance equal to ±0.0035. The tolerance set using the conventional formula is also shown.

The advantage of the statistical formula is larger component tolerances. With alternative 1, the tolerance for component *A* has been doubled, the tolerance for component *B* has been quadrupled, and the tolerance for component *C* has been kept the same as the original component based on the simple addition approach. If alternative 2 is chosen, similar significant increases in the component tolerances may be achieved. This formula, then, may result in a larger component tolerance with no change in the manufacturing processes and no change in the assembly tolerance. Note that the *largest single* tolerance has the greatest effect on the overall result.

The disadvantage of the quadratic formula is that it involves several assumptions that, even if met, will still result in a small percent (theoretically 0.27 percent) of results not conforming to the limits set by the formula. The assumptions are

1. The component dimensions are independent and the components are assembled randomly. This assumption is usually met in practice.
2. Each component dimension should be normally distributed. Some departure from this assumption is permissible.
3. The actual average for each component is equal to the nominal value stated in the specification. For the original assembly example, the actual averages for components *A, B,* and *C* must be 1.000, 0.500, and 2.000, respectively. Otherwise, the nominal value of 3.500 will not be achieved for the assembly, and tolerance limits set about 3.500 will not be realistic. Thus it is important to control the *average* value for interacting dimensions. This means that process control techniques are needed using variables measurement rather than go no-go measurement.

A summary of the two methods of tolerance is given in Table 44.29.

The statistical tolerancing formula applies both to assemblies made up of physically separate components and to a chain of several interacting dimensions within one physical item. Further, the result of the interacting dimensions can be an outside dimension (assembly length) or an internal result (clearance between a shaft and hole).

TABLE 44.28 Comparison of Statistical and Conventional Methods

	Statistical		
Component	Alternative 1	Alternative 2	Conventional
A	±0.002	±0.001	±0.0010
B	±0.002	±0.001	±0.0005
C	±0.002	±0.003	±0.0020

TABLE 44.29 Comparison of Conventional and Statistical Tolerancing

Factor	Conventional	Statistical
Risk of items not interacting properly	No risk; 100% interchangeability of items	Small percent of final results will fall outside limits (but these can sometimes be corrected with selective assembly)
Utilization of full tolerance range	Method is conservative; tolerances on interacting dimensions are smaller than necessary	Permits larger tolerances on interacting dimensions
Special process control techniques	None	Average of each interacting dimension must be controlled using variables measurement
Statistical assumptions	None	Interacting dimensions must be independent and each must be normally distributed
Lot size for components	Any size	Lot size should be moderately large (to assure balancing effect on extreme interacting dimensions)

Further Applications of Statistical Tolerancing. It is easy to be deceived into concluding that the statistical method of tolerancing is merely a change from an expression of tolerances in the form of limits on each component to a form of

1. Upper and lower limits on the average $\overline{X}$ of the mass of components
2. An upper limit to the scatter σ of the components

The change is much more profound than mere form of the specification. It affects the entire cycle of manufacturing planning, production, inspection, quality control, service, etc. It is, in effect, *a new philosophy of manufacture.*

The first published example of a *large-scale application* of statistical tolerancing appears to be that of the L-3 coaxial system (a broad-band transmission system for multiple telephone or television channels). Dodge et al. (1953) discuss the application.

The general plan was

1. Discovery of the key quality characteristics of each component element of the system.
2. Determination of the precision of measurement to separate measurement variability from process and product variability.
3. Collection of data on process capability for the key qualities, to aid in establishing realistic tolerances.

The foregoing were preliminary to

4. Establishment of tolerances for the key quality characteristics in the dual form of a maximum on the standard deviation σ and limits on the average $\overline{X}$. The limits on X were established as $\pm(1/3)\sigma$ around the nominal.
5. Establishment of control procedures.

It was recognized that the limits on σ and $\overline{X}$ required further interpretation if the intent of the designers was to be carried out by the manufacturers. To this end, three forms of product acceptance were established:

1. *Control charts.* Shewhart control charts for $\bar{X}$ and σ could be used for product acceptance, provided "eligibility" was established (seven consecutive subgroups, of five pieces each, all met the control limits for $\bar{X}$ and σ) and provided subsequent statistical control was maintained (based on chart results plus absence of major changes in process).
2. *Batch control.* This was based on examination of a sample of (normally) 50 pieces by variables measurements, with limits on $\bar{X}$ and σ appropriate to the sample size of 50. Each batch stood or fell on its own measurements.
3. *Detailed classification.* Product that did not qualify under 1 or 2 was measured in detail. The resulting conforming units were classified into one of three variable classes. The packaging was then done by selecting classified units in such a way that each package contained an assortment of product which conformed to the intent of the design as to $\bar{X}$ and σ.

Grant and Leavenworth (1980) discuss statistical tolerancing, including an application to shafts and holes. The Western Electric Company, Inc. (1982), in its *Statistical Quality Control Handbook* (pp. 122–127), presents examples and discusses the assumptions. Peters (1970) discusses statistical tolerancing, including a method for recognizing cost differences among components. Choksi (1971) discusses the use of computer simulation to determine optimum tolerances.

The concept may be applied to several interacting variables in an engineering relationship. The nature of the relationship need *not* be additive (assembly example) or subtractive (shaft and hole example). The formula can be adapted to predict the variation of results that are the product and/or the division of several variables. Mouradian (1966) discusses these applications.

Bayesian Estimates. Bayesian estimation can be used when the parameter to be estimated can be considered to be a random variable for which we know the distribution. For example, in sampling inspection, the proportion of defectives p in a lot may be a random variable about which we can fit a distribution by our sampling inspection over time. If so, then in the future that information can be used to provide quality assurance with less sampling (see Lenz and Rendtel 1984).

When it is not possible to cumulate information about a stable process, some have proposed that we use our "feelings" about the parameter to choose a statistical distribution for it and then proceed as if the parameter were a random variable with that distribution. This is called the *personal probability approach,* and those who use it are called *Bayesians.* Some of the proponents of this approach say that it is the only method that any sensible person should use, and this has been cause for bitter debates and ill-will. In my view, while a person in a management position might find this a reasonable way to express his or her insights quantitatively, in most cases this will be an unscientific way of simply incorporating prejudices into the decision process, resulting in costly errors. This approach is of some use in general statistical decision theory (see Chapter 12 of Dudewicz 1976), but there it is used to generate a set of decision rules that contains all good rules, not just one rule based on one's "feelings." The Bayesian approach should be considered whenever information can be gathered over time on a stable process.

Intervals with Fixed Width and Precision. The intervals considered up to now typically either had a random width (e.g., parts 7 and 8 of Table 44.22) or required that one know such parameters as variances (e.g., part 6 of Table 44.22). If one can take observations in two stages, then one can control both the width and the confidence. Let me illustrate for two normal means when we do not know the variances (and do not know that they are equal). Here we can proceed in two stages as follows:

Sample n_0 observations from each of $k = 2$ populations (n_0 at least 10 is desirable).

Determine the total sample size for population i as

$$n_i = max\,[n_0 + 1, (ws_i)^2] \qquad \text{with } w = h_{n_0}(2, (1 + P^*)/2)/d$$

where h is from Table 44.30, d is the desired half-width, $k = 2$ populations, and P^* is the confidence desired.

TABLE 44.30 Multipliers $h=h_{n_0}(k, P^*)$ Needed for Solving Two-Stage Testing, Confidence Interval, and Selection of the Best Problems*

k	P^*	n_0												
		2	3	4	5	6	7	8	9	10	15	20	25	30
2	.75	2.00	1.37	1.21	1.14	1.10	1.07	1.05	1.04	1.03	1.00	0.99	0.98	0.98
	.80	2.75	1.76	1.54	1.44	1.38	1.35	1.32	1.30	1.29	1.25	1.24	1.23	1.22
	.85	3.93	2.27	1.94	1.80	1.72	1.68	1.64	1.62	1.60	1.55	1.53	1.51	1.51
	.90	6.16	3.04	2.50	2.29	2.18	2.11	2.06	2.02	2.00	1.93	1.90	1.88	1.87
	.95	12.63	4.57	3.50	3.11	2.91	2.79	2.71	2.66	2.61	2.50	2.45	2.42	2.41
	.975	25.42	6.54	4.59	3.94	3.63	3.45	3.33	3.24	3.18	3.02	2.95	2.91	2.88
	.99	63.7	10.28	6.31	5.14	4.60	4.30	4.11	3.98	3.89	3.64	3.54	3.48	3.45
3	.75	3.52	2.15	1.86	1.74	1.67	1.63	1.60	1.57	1.56	1.51	1.49	1.48	1.47
	.80	4.59	2.59	2.20	2.04	1.95	1.89	1.85	1.83	1.80	1.75	1.72	1.71	1.70
	.85	6.31	3.17	2.62	2.40	2.28	2.21	2.16	2.13	2.10	2.03	1.99	1.98	1.96
	.90	9.64	4.05	3.22	2.90	2.73	2.63	2.57	2.52	2.48	2.39	2.34	2.32	2.30
	.95	19.40	5.86	4.29	3.75	3.48	3.32	3.21	3.14	3.08	2.94	2.87	2.84	2.81
	.975	38.7	8.25	5.50	4.63	4.22	3.98	3.82	3.72	3.64	3.43	3.35	3.30	3.27
	.99	96.2	12.83	7.44	5.91	5.23	4.86	4.62	4.46	4.34	4.04	3.92	3.85	3.81
4	.75	4.77	2.66	2.25	2.08	1.99	1.93	1.89	1.86	1.84	1.78	1.75	1.74	1.73
	.80	6.16	3.13	2.59	2.38	2.26	2.19	2.14	2.11	2.08	2.01	1.98	1.96	1.95
	.85	8.41	3.77	3.03	2.75	2.60	2.51	2.45	2.41	2.37	2.28	2.24	2.22	2.21
	.90	12.80	4.75	3.66	3.26	3.06	2.93	2.85	2.80	2.75	2.63	2.58	2.55	2.54
	.95	25.76	6.80	4.80	4.14	3.81	3.62	3.50	3.41	3.34	3.17	3.10	3.06	3.03
	.975	51.4	9.53	6.10	5.05	4.57	4.29	4.11	3.99	3.90	3.67	3.57	3.51	3.48
	.99	128	14.79	8.21	6.40	5.62	5.19	4.92	4.74	4.60	4.27	4.13	4.05	4.01
5	.75	5.95	3.05	2.53	2.32	2.21	2.14	2.09	2.06	2.03	1.96	1.93	1.91	1.90
	.80	7.65	3.56	2.89	2.63	2.49	2.40	2.35	2.30	2.27	2.19	2.15	2.13	2.12
	.85	10.43	4.25	3.34	3.00	2.83	2.72	2.65	2.60	2.56	2.46	2.41	2.39	2.37
	.90	15.90	5.32	4.00	3.53	3.29	3.15	3.05	2.99	2.94	2.81	2.75	2.72	2.69
	.95	32.04	7.58	5.20	4.42	4.05	3.84	3.70	3.60	3.53	3.34	3.26	3.21	3.18
	.975	64.1	10.61	6.58	5.37	4.83	4.52	4.32	4.18	4.08	3.83	3.72	3.66	3.62
	.99	160	16.47	8.82	6.78	5.90	5.43	5.13	4.93	4.79	4.43	4.28	4.20	4.14
6	.75	7.10	3.39	2.76	2.52	2.38	2.30	2.25	2.21	2.18	2.10	2.06	2.04	2.03
	.80	9.12	3.93	3.13	2.82	2.66	2.57	2.50	2.45	2.42	2.32	2.28	2.26	2.24
	.85	12.43	4.66	3.60	3.21	3.01	2.89	2.81	2.75	2.70	2.59	2.54	2.51	2.49
	.90	18.96	5.82	4.28	3.74	3.47	3.31	3.21	3.13	3.08	2.93	2.87	2.84	2.81
	.95	38.29	8.26	5.53	4.66	4.25	4.01	3.86	3.75	3.67	3.46	3.38	3.33	3.30
	.975	76.7	11.56	6.97	5.64	5.03	4.69	4.48	4.33	4.22	3.95	3.83	3.77	3.73
	.99	192	17.97	9.34	7.09	6.13	5.62	5.30	5.09	4.93	4.55	4.39	4.30	4.25
7	.75	8.23	3.68	2.96	2.67	2.53	2.44	2.37	2.33	2.30	2.21	2.17	2.14	2.13
	.80	10.58	4.25	3.33	2.99	2.81	2.70	2.63	2.58	2.54	2.43	2.38	2.36	2.34
	.85	14.42	5.03	3.81	3.37	3.15	3.02	2.93	2.87	2.82	2.70	2.64	2.61	2.59
	.90	22.01	6.27	4.51	3.92	3.62	3.45	3.33	3.25	3.19	3.04	2.97	2.93	2.91
	.95	44.53	8.88	5.81	4.85	4.41	4.15	3.98	3.87	3.79	3.57	3.47	3.42	3.39
	.975	89.3	12.43	7.32	5.86	5.20	4.84	4.61	4.45	4.34	4.05	3.93	3.86	3.82
	.99	223	19.34	9.79	7.35	6.32	5.78	5.44	5.21	5.05	4.64	4.48	4.39	4.33
8	.75	9.36	3.95	3.13	2.81	2.65	2.55	2.48	2.43	2.40	2.30	2.25	2.23	2.21
	.80	12.02	4.55	3.51	3.13	2.93	2.81	2.73	2.68	2.63	2.52	2.47	2.44	2.42
	.85	16.40	5.37	4.01	3.52	3.28	3.13	3.04	2.97	2.92	2.78	2.72	2.69	2.67
	.90	25.05	6.68	4.72	4.07	3.75	3.56	3.44	3.35	3.29	3.12	3.05	3.01	2.98
	.95	50.76	9.45	6.06	5.03	4.54	4.27	4.09	3.97	3.88	3.65	3.55	3.50	3.46
	.975	101.9	13.24	7.63	6.05	5.35	4.97	4.72	4.56	4.44	4.13	4.00	3.93	3.89
	.99	256	20.62	10.20	7.58	6.49	5.91	5.56	5.32	5.15	4.73	4.55	4.46	4.40
9	.75	10.49	4.20	3.28	2.93	2.75	2.64	2.57	2.52	2.48	2.37	2.33	2.30	2.28
	.80	13.47	4.82	3.67	3.25	3.04	2.91	2.82	2.76	2.72	2.60	2.54	2.51	2.49
	.85	18.37	5.68	4.18	3.65	3.39	3.23	3.13	3.06	3.00	2.86	2.80	2.76	2.74
	.90	28.08	7.06	4.91	4.21	3.86	3.66	3.53	3.44	3.37	3.20	3.12	3.08	3.05
	.95	57.0	9.99	6.29	5.18	4.66	4.37	4.19	4.06	3.96	3.72	3.62	3.56	3.53
	.975	114.5	13.99	7.91	6.22	5.48	5.08	4.82	4.65	4.52	4.20	4.07	4.00	3.95
	.99	287	21.8	10.58	7.79	6.64	6.03	5.66	5.41	5.23	4.80	4.62	4.53	4.46

TABLE 44.30 Multipliers $h=h_{n_0}(k, P^*)$ Needed for Solving Two-Stage Testing, Confidence Interval, and Selection of the Best Problems* *(Continued)*

k	P^*	n_0												
		2	3	4	5	6	7	8	9	10	15	20	25	30
10	.75	11.60	4.43	3.42	3.04	2.85	2.73	2.65	2.60	2.55	2.44	2.39	2.36	2.35
	.80	14.90	5.08	3.82	3.36	3.13	3.00	2.90	2.84	2.79	2.66	2.60	2.57	2.55
	.85	20.34	5.98	4.33	3.77	3.48	3.32	3.21	3.13	3.08	2.92	2.86	2.82	2.80
	.90	31.12	7.41	5.09	4.33	3.96	3.75	3.62	3.52	3.45	3.26	3.18	3.14	3.11
	.95	63.2	10.49	6.50	5.32	4.77	4.47	4.27	4.14	4.04	3.79	3.68	3.62	3.58
	.975	127.1	14.70	8.17	6.38	5.60	5.17	4.91	4.73	4.60	4.26	4.13	4.05	4.01
	.99	318	22.9	10.92	7.98	6.77	6.14	5.75	5.49	5.31	4.86	4.68	4.58	4.51
11	.75	12.72	4.64	3.54	3.14	2.93	2.81	2.72	2.66	2.62	2.50	2.45	2.42	2.40
	.80	16.34	5.32	3.95	3.46	3.22	3.07	2.98	2.91	2.86	2.72	2.66	2.63	2.61
	.85	22.31	6.25	4.48	3.87	3.57	3.40	3.28	3.20	3.14	2.98	2.91	2.87	2.85
	.90	34.15	7.75	5.25	4.44	4.06	3.83	3.69	3.59	3.51	3.32	3.24	3.19	3.16
	.95	69.4	10.97	6.70	5.44	4.87	4.55	4.35	4.21	4.10	3.84	3.73	3.67	3.63
	.975	139.7	15.38	8.41	6.52	5.71	5.26	4.99	4.80	4.66	4.32	4.18	4.10	4.05
	.99	350	24.0	11.25	8.16	6.90	6.24	5.84	5.57	5.38	4.91	4.73	4.62	4.56
12	.75	13.84	4.85	3.66	3.23	3.01	2.88	2.79	2.73	2.68	2.55	2.50	2.47	2.45
	.80	17.77	5.54	4.07	3.56	3.30	3.14	3.04	2.97	2.92	2.77	2.71	2.68	2.66
	.85	24.27	6.52	4.61	3.97	3.65	3.47	3.35	3.26	3.20	3.03	2.96	2.92	2.90
	.90	37.17	8.07	5.40	4.54	4.14	3.91	3.76	3.65	3.57	3.37	3.28	3.24	3.21
	.95	75.6	11.42	6.88	5.56	4.96	4.63	4.42	4.27	4.16	3.89	3.78	3.72	3.68
	.975	152.3	16.02	8.63	6.66	5.08	5.34	5.06	4.86	4.72	4.37	4.23	4.15	4.10
	.99	382	25.0	11.55	8.32	7.01	6.32	5.91	5.64	5.44	4.96	4.77	4.67	4.60
13	.75	14.95	5.04	3.77	3.31	3.08	2.94	2.85	2.78	2.73	2.60	2.54	2.51	2.49
	.80	19.21	5.76	4.19	3.64	3.37	3.21	3.10	3.03	2.97	2.82	2.76	2.72	2.70
	.85	26.24	6.77	4.74	4.06	4.73	3.53	3.41	3.32	3.25	3.08	3.01	2.97	2.94
	.90	40.20	8.38	5.54	4.64	4.22	3.97	3.82	3.71	3.63	3.42	3.33	3.28	3.25
	.95	81.8	11.86	7.05	5.67	5.05	4.70	4.48	4.33	4.22	3.94	3.82	3.76	3.72
	.975	164.8	16.64	8.85	6.78	5.90	5.42	5.12	4.92	4.78	4.42	4.27	4.19	4.14
	.99	413	26.0	11.84	8.47	7.11	6.41	5.98	5.70	5.50	5.01	4.81	4.71	4.64
14	.75	16.06	5.23	3.87	3.39	3.15	3.00	2.90	2.83	2.78	2.65	2.59	2.55	2.53
	.80	20.64	5.97	4.30	3.72	3.44	3.27	3.16	3.08	3.02	2.86	2.80	2.76	2.74
	.85	28.20	7.01	4.86	4.14	3.80	3.60	3.46	3.37	3.30	3.12	3.05	3.01	2.98
	.90	43.23	8.67	5.67	4.73	4.29	4.04	3.87	3.76	3.68	3.46	3.37	3.32	3.29
	.95	88.1	12.27	7.22	5.77	5.12	4.76	4.54	4.38	4.27	3.98	3.86	3.80	3.75
	.975	177.4	17.24	9.05	6.90	5.98	5.49	5.18	4.98	4.83	4.46	4.31	4.22	4.17
	.99	445	27.0	12.11	8.61	7.21	6.48	6.05	5.76	5.55	5.05	4.85	4.74	4.68
15	.75	17.17	5.41	3.97	3.46	3.21	3.05	2.95	2.88	2.83	2.69	2.63	2.59	2.57
	.80	22.07	6.17	4.40	3.80	3.50	3.32	3.21	3.13	3.07	2.91	2.84	2.80	2.77
	.85	30.17	7.24	4.97	4.22	3.86	3.65	3.52	3.42	3.35	3.16	3.09	3.01	3.01
	.90	46.25	8.96	5.80	4.82	4.36	4.09	3.93	3.81	3.72	3.50	3.41	3.35	3.32
	.95	94.3	12.68	7.37	5.87	5.20	4.83	4.59	4.43	4.32	4.02	3.90	3.83	3.79
	.975	190.0	17.81	9.24	7.01	6.06	5.55	5.24	5.03	4.88	4.50	4.34	4.26	4.20
	.99	476	27.9	12.37	8.75	7.30	6.55	6.11	5.81	5.60	5.09	4.88	4.78	4.71
16	.75	18.28	5.58	4.06	3.53	3.26	3.11	3.00	2.93	2.87	2.72	2.66	2.63	2.60
	.80	23.50	6.36	4.50	3.87	3.56	3.37	3.26	3.17	3.11	2.94	2.87	2.83	2.81
	.85	32.13	7.46	5.07	4.29	3.92	3.70	3.56	3.47	3.39	3.20	3.12	3.07	3.05
	.90	49.28	9.23	5.92	4.90	4.42	4.15	3.97	3.85	3.77	3.54	3.44	3.39	3.35
	.95	100.5	13.07	7.52	5.96	5.27	4.88	4.64	4.48	4.36	4.06	3.93	3.86	3.82
	.975	202.6	18.37	9.42	7.11	6.14	5.62	5.29	5.08	4.92	4.53	4.38	4.29	4.24
	.99	508	28.8	12.62	8.88	7.39	6.62	6.16	5.86	5.65	5.12	4.92	4.81	4.74
17	.75	19.40	5.74	4.15	3.59	3.32	3.15	3.05	2.97	2.91	2.76	2.69	2.66	2.63
	.80	24.93	6.55	4.59	3.94	3.61	3.42	3.30	3.21	3.15	2.98	2.90	2.86	2.84
	.85	34.09	7.68	5.17	4.37	3.98	3.75	3.61	3.51	3.43	3.24	3.15	3.11	3.08
	.90	52.30	9.50	6.03	4.97	4.48	4.20	4.02	3.90	3.81	3.57	3.47	3.42	3.38
	.95	106.7	13.45	7.66	6.04	5.33	4.94	4.69	4.52	4.40	4.09	3.96	3.89	3.85
	.975	215.3	18.90	9.60	7.21	6.21	5.67	5.34	5.12	4.96	4.57	4.41	4.32	4.27
	.99	542	29.6	12.86	9.00	7.47	6.68	6.22	5.91	5.69	5.16	4.95	4.84	4.76

TABLE 44.30 Multipliers $h=h_{n_0}(k, P^*)$ Needed for Solving Two-Stage Testing, Confidence Interval, and Selection of the Best Problems* *(Continued)*

k	P^*	n_0												
		2	3	4	5	6	7	8	9	10	15	20	25	30
18	.75	20.51	5.90	4.23	3.66	3.37	3.20	3.09	3.01	2.95	2.79	2.72	2.69	2.66
	.80	26.36	6.73	4.68	4.00	3.67	3.47	3.34	3.25	3.19	3.01	2.93	2.89	2.87
	.85	36.05	7.89	5.27	4.43	4.03	3.80	3.65	3.55	3.47	3.27	3.18	3.14	3.10
	.90	55.3	9.76	6.14	5.04	4.54	4.25	4.06	3.94	3.84	3.60	3.50	3.45	3.41
	.95	112.9	13.81	7.80	6.13	5.39	4.99	4.73	4.56	4.44	4.12	3.99	3.92	3.87
	.975	227.7	19.43	9.77	7.31	6.28	5.73	5.39	5.16	5.00	4.60	4.43	4.35	4.29
	.99	571	30.5	13.09	9.12	7.55	6.74	6.27	5.95	5.73	5.19	4.98	4.86	4.79
19	.75	21.61	6.06	4.32	3.71	3.42	3.24	3.13	3.04	2.98	2.82	2.75	2.71	2.69
	.80	27.79	6.91	4.77	4.06	3.72	3.51	3.38	3.29	3.22	3.04	2.96	2.92	2.89
	.85	38.01	8.09	5.37	4.50	4.08	3.85	3.69	3.59	3.51	3.30	3.21	3.16	3.13
	.90	58.4	10.01	6.25	5.11	4.59	4.29	4.10	3.97	3.88	3.63	3.53	3.47	3.44
	.95	119.1	14.17	7.93	6.20	5.45	5.04	4.78	4.60	4.47	4.15	4.02	3.95	3.90
	.975	240.2	19.94	9.93	7.40	6.34	5.78	5.43	5.20	5.04	4.63	4.46	4.37	4.32
	.99	604	31.3	13.31	9.23	7.62	6.80	6.31	5.99	5.77	5.22	5.00	4.89	4.81
20	.75	22.72	6.21	4.39	3.77	3.46	3.28	3.16	3.08	3.02	2.85	2.78	2.74	2.72
	.80	29.22	7.08	4.85	4.12	3.76	3.56	3.42	3.32	3.26	3.07	2.99	2.95	2.92
	.85	39.98	8.29	5.46	4.56	4.13	3.89	3.73	3.62	3.54	3.33	3.24	3.19	3.16
	.90	61.4	10.25	6.35	5.18	4.64	4.34	4.14	4.01	3.91	3.66	3.56	3.50	3.46
	.95	125.3	14.52	8.05	6.28	5.51	5.08	4.82	4.64	4.51	4.18	4.04	3.97	3.92
	.975	252.9	20.43	10.09	7.49	6.40	5.83	5.48	5.24	5.07	4.66	4.49	4.40	4.34
	.99	635	32.1	13.52	9.34	7.69	6.86	6.36	6.03	5.80	5.25	5.03	4.91	4.84
21	.75	23.83	6.36	4.47	3.82	3.51	3.32	3.20	3.11	3.05	2.88	2.81	2.77	2.74
	.80	30.65	7.24	4.93	4.17	3.81	3.59	3.45	3.36	3.29	3.10	3.02	2.97	2.94
	.85	41.94	8.48	5.54	4.62	4.18	3.93	3.77	3.65	3.57	3.35	3.26	3.21	3.18
	.90	64.4	10.49	6.44	5.24	4.69	4.38	4.18	4.04	3.94	3.69	3.58	3.52	3.48
	.95	131.5	14.86	8.17	6.35	5.56	5.13	4.86	4.67	4.54	4.21	4.07	3.99	3.95
	.975	265.7	20.92	10.24	7.57	6.46	5.88	5.52	5.28	5.11	4.68	4.51	4.42	4.36
	.99	667	32.8	13.72	9.44	7.76	6.91	6.40	6.07	5.84	5.27	5.05	4.93	4.86
22	.75	24.94	6.50	4.54	3.88	3.55	3.36	3.23	3.15	3.08	2.91	2.83	2.79	2.76
	.80	32.08	7.40	5.01	4.23	3.85	3.63	3.49	3.39	3.32	3.12	3.04	2.99	2.96
	.85	43.90	8.67	5.63	4.67	4.22	3.97	3.80	3.69	3.60	3.38	3.29	3.24	3.20
	.90	67.4	10.72	6.54	5.30	4.74	4.42	4.22	4.08	3.98	3.71	3.60	3.54	3.51
	.95	137.7	15.19	8.29	6.42	5.61	5.17	4.89	4.71	4.57	4.23	4.09	4.02	3.97
	.975	278	21.39	10.39	7.65	6.52	5.92	5.56	5.31	5.14	4.71	4.53	4.44	4.38
	.99	700	33.6	13.93	9.54	7.83	6.96	6.44	6.11	5.87	5.30	5.07	4.95	4.88
23	.75	26.05	6.64	4.61	3.92	3.59	3.39	3.27	3.18	3.11	2.93	2.85	2.81	2.78
	.80	33.51	7.56	5.08	4.28	3.89	3.67	3.52	3.42	3.34	3.15	3.06	3.02	2.99
	.85	45.86	8.85	5.71	4.73	4.27	4.00	3.83	3.72	3.63	3.41	3.31	3.26	3.22
	.90	70.4	10.95	6.63	5.36	4.78	4.46	4.25	4.11	4.00	3.74	3.63	3.57	3.53
	.95	143.9	15.51	8.40	6.49	5.66	5.21	4.93	4.74	4.60	4.26	4.11	4.04	3.99
	.975	291	21.85	10.53	7.73	6.57	5.96	5.59	5.35	5.17	4.73	4.56	4.46	4.40
	.99	730	34.3	14.12	9.63	7.89	7.01	6.48	6.14	5.90	5.32	5.10	4.97	4.90
24	.75	27.16	6.78	4.68	3.97	3.63	3.43	3.30	3.20	3.14	2.95	2.88	2.83	2.81
	.80	34.93	7.72	5.15	4.33	3.93	3.70	3.55	3.45	3.37	3.17	3.08	3.04	3.01
	.85	47.82	9.03	5.78	4.78	4.31	4.04	3.87	3.75	3.66	3.43	3.33	3.28	3.24
	.90	73.5	11.17	6.72	5.42	4.82	4.49	4.28	4.14	4.03	3.76	3.65	3.59	3.55
	.95	150.1	15.83	8.51	6.55	5.71	5.25	4.96	4.77	4.63	4.28	4.14	4.06	4.01
	.975	303	22.30	10.67	7.80	6.63	6.01	5.63	5.38	5.20	4.76	4.58	4.48	4.42
	.99	762	35.1	14.31	9.73	7.95	7.05	6.52	6.18	5.93	5.37	5.12	4.99	4.92
25	.75	28.27	6.91	4.74	4.02	3.67	3.46	3.33	3.23	3.16	2.98	2.90	2.85	2.83
	.80	36.36	7.87	5.22	4.38	3.97	3.74	3.58	3.48	3.40	3.19	3.11	3.06	3.03
	.85	49.78	9.21	5.86	4.83	4.35	4.07	3.90	3.77	3.68	3.45	3.35	3.30	3.26
	.90	76.5	11.38	6.80	5.47	4.87	4.53	4.31	4.17	4.06	3.78	3.67	3.61	3.57
	.95	156.3	16.14	8.62	6.62	5.75	5.28	4.99	3.80	4.66	4.30	4.16	4.08	4.03
	.975	316	22.74	10.80	7.88	6.68	6.05	5.66	5.41	5.23	4.78	4.60	4.50	4.44
	.99	794	35.8	14.49	9.82	8.01	7.10	6.56	6.21	5.96	5.37	5.14	5.01	4.94

*The table entries are from Table 4 on pp. 17–23 of "New Tables for Multiple Comparisons with a Control (Unknown Variances)," by E. J. Dudewicz, J. S. Ramberg, and H. J. Chen, *Biometrische Zeitschrift,* vol. 17 (1975), pp. 13–26. Reprinted with the permission of Akademie-Verlag, Berlin.

Take $n_1 - n_0$ more observations on population 1 and $n_2 - n_0$ more on population 2.

Compute the sample means $\overline{X}_1$ of n_1 observations, and $\overline{X}_2$ of n_2 observations and the interval is $\overline{X}_1 - \overline{X}_2 \pm d$ (i.e., we are $100P^*$ percent sure that the difference of the population means is within d units of the difference of the sample means). For an example of the calculations involved, see Example of Selection of the Best below.

STATISTICAL TESTS OF HYPOTHESES

A *statistical hypothesis* is an assertion about a population, often about some parameter of a population. *Tests of hypotheses* (also called *tests of significance*) were designed so that experimenters would not ascribe causes to variations in data that were in fact due simply to random variation (and thus did not need a cause to explain them). Thus statistical hypothesis testing is a modern-day version of the medieval Occam's razor principle that "one should not multiply causes without reason." (William of Occam was an English Franciscan philosopher who died about 1349.) For example, if a process has a mean weight of 14.90 lb per item produced, a change is made with a view to increasing the weight per item produced, and a sample of 10 items (taken after the process change) has a mean weight of 15.10 lb, this *does not* necessarily mean that the process mean has been shifted up: It could be that it has remained the same (or has even decreased) and that we are simply seeing the results of the randomness of the process. Making correct inferences in the face of such possibilities is the gist of the area of hypothesis testing.

Basic Concepts, Types of Errors. *Hypothesis* as used here is an assertion made about a population. Usually the assertion concerns the numerical value of some parameter of the population. For example, a hypothesis might state that the mean life of a population of batteries equals 30.0, written as $H{:}\mu_0 = 30.0$. This assertion may or may not be correct. A *hypothesis test* is a test of the validity of the assertion and is carried out by analysis of a sample of data.

There are two reasons why sample results must be evaluated carefully. First, there are many other samples that, by chance alone, could be drawn from the population. Second, the numerical results in the sample actually selected can easily be compatible with several different hypotheses. These points are handled by recognizing two types of error which can be made in evaluating a hypothesis:

1. *Reject* the hypothesis when it is *true.* This is called the *type 1 error*; its probability is called the *level of significance* and is denoted by α.
2. *Accept* the hypothesis when it is *false.* This is called the *type II error*; its probability is usually denoted by β (though some authors call it $1-\beta$).

These error probabilities can be controlled to desired values.

The type I error is shown graphically in Figure 44.19 for the hypothesis $H{:}\mu_0 = 30.0$. The area between the vertical lines represents the *acceptance region* for the hypothesis test: If the sample result falls within the acceptance region, the hypothesis is accepted. Otherwise, it is rejected. Notice that there is a small portion of the curve that falls outside the acceptance region. This portion (α) represents the probability of obtaining a sample result outside the acceptance region, even though the hypothesis is correct.

Suppose it has been decided that the type I error must not exceed 5 percent. This is the probability of rejecting the hypothesis when, in truth, the true average life is 30.0. The acceptance region can be obtained by locating values of average life that have only a 5 percent chance of being exceeded when the true average life is 30.0. Further, suppose a sample n of four measurements is taken and $\sigma = 10.0$.

Remember that the curve represents a population of sample averages because the decision will be made on the basis of a sample average. Sample averages vary less than individual measurements according to the relationship $\sigma_{\overline{X}} = \sigma/\sqrt{n}$.

Further, the distribution of sample averages is approximately normal even if the distribution of the individual measurements (going into the averages) is not normal [see Grant and Leavenworth

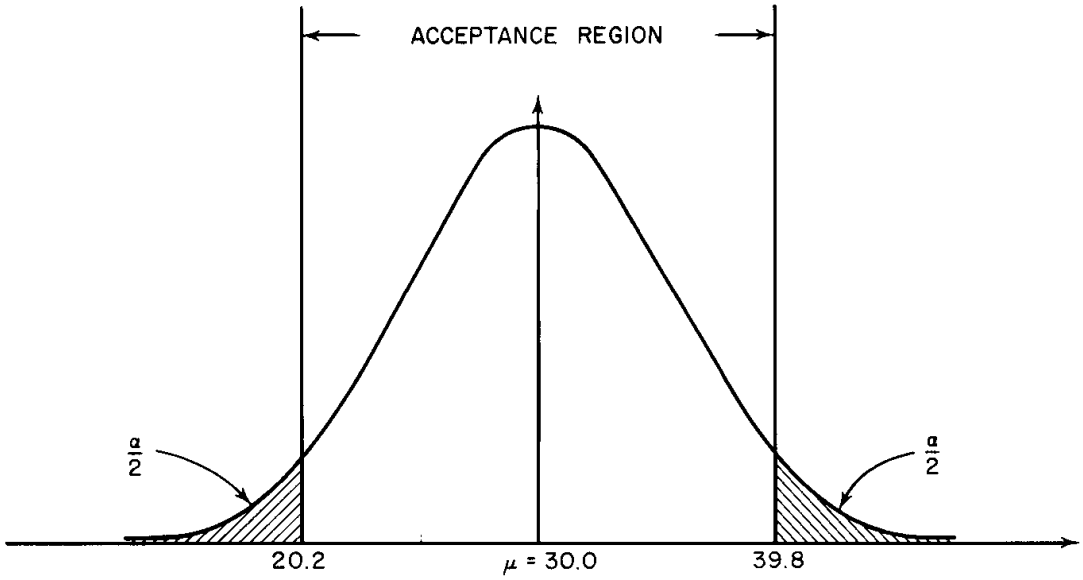


FIGURE 44.19 Acceptance region for H:$\mu_0 = 30.0$.

(1972, pp. 69–71)]. The approximation holds best for large values of n but is adequate for n as low as 4.

Table B in Appendix II shows that a 2.5 percent area in each tail is at a limit that is 1.96 standard deviations from 30.0. Then under the hypothesis that $\mu_0 = 30.0$, 95 percent of sample averages will fall within $\pm 1.96\sigma_{\bar{x}}$ of 30.0, or

$$\text{Upper limit} = 30.0 + 1.96\frac{10}{\sqrt{4}} = 39.8$$

$$\text{Lower limit} = 30.0 - 1.96\frac{10}{\sqrt{4}} = 20.2$$

The acceptance region is thereby defined as 20.2 to 39.8. If the average of a random sample of four batteries is within this acceptance region, the hypothesis is accepted. If the average falls outside the acceptance region, the hypothesis is rejected. This decision rule provides a type I error of 0.05.

The type II or β error, the probability of accepting a hypothesis when it is false, is shown in Figure 44.20 as the shaded area. Notice that it is possible to obtain a sample result within the acceptance region, even though the population has a true average that is not equal to the average stated in the hypothesis. The numerical value of β depends on the true value of the population average (and also on n, σ, and α). This is depicted by an *operating characteristic (OC) curve.*

The problem now is to construct an operating characteristic curve to assess the magnitude of the type II (β) error. Since β is the probability of *accepting* the original hypothesis ($\mu_0 = 30.0$) when it is *false,* the probability that a sample average will fall between 20.2 and 39.8 must be found when the true average of the population is something other than 30.0. This has been done for many values of the true average, and the result is shown in Figure 44.21. [This curve should not be confused with that of a normal distribution of measurements. In some cases the shape is similar, but the meanings of an OC curve and a distribution curve are entirely different; Juran and Gryna (1980, pp. 410–412) give the detailed calculations; also see Dudewicz (1976, pp. 272–275).] Thus the OC curve is a plot of the probability of accepting the original hypothesis as a function of the true value of the population parameter (and the given values of n, σ, and α).

Use of the Operating Characteristic Curve in Selecting an Acceptance Region. The acceptance region was determined by dividing the 5 percent allowable α error into equal parts

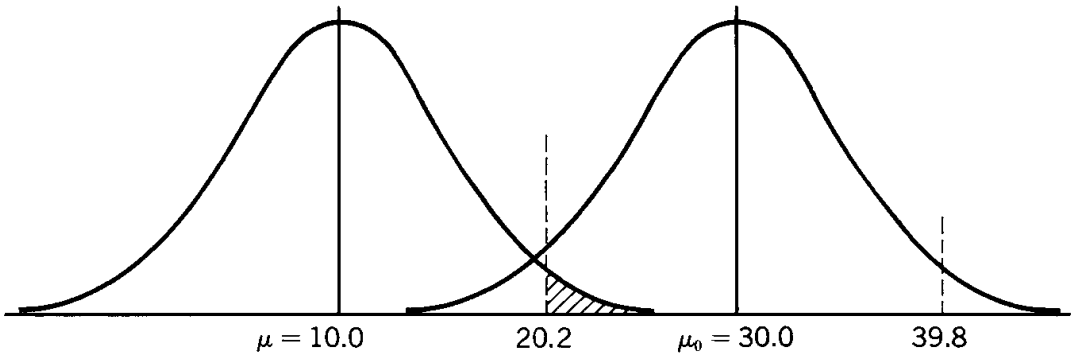


FIGURE 44.20 Type II or β error.

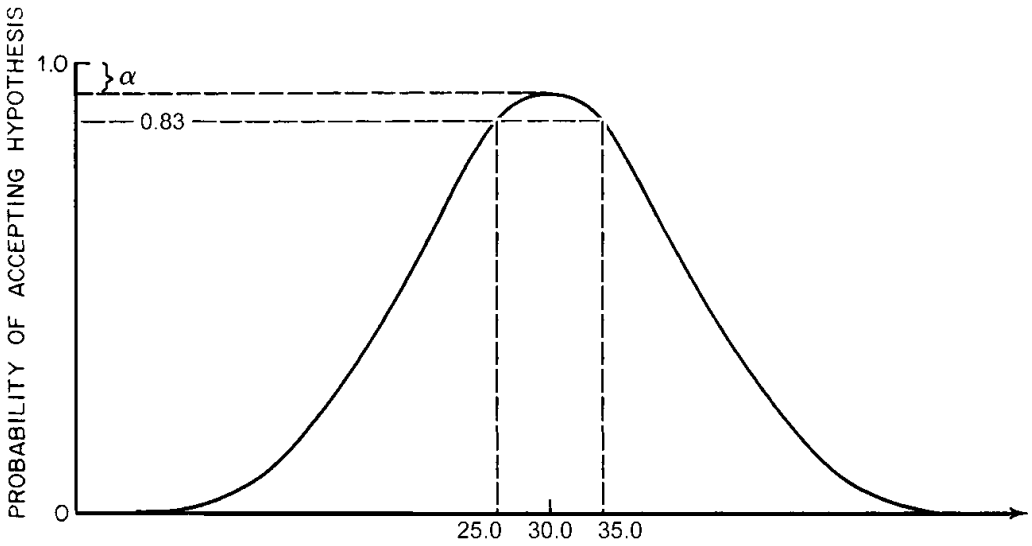


FIGURE 44.21 Operating characteristic curve.

(see Figure 44.19). This is called a *two-tailed test.* The entire 5 percent also could be placed at either the left or the right tail of the distribution curve (Figure 44.22). These are *one-tailed tests.*

Operating characteristic curves for tests having these one-tailed acceptance regions can be developed following the approach used for the two-tailed region. Although the α error is the same, the β error varies for the three tests. (See Figure 9.2-2 on p. 275 of Dudewicz 1976.)

In some problems, knowledge is available to indicate that if the true average of the population is *not* equal to the hypothesis value, then it is on one side of the hypothesis value. For example, a new material of supposedly higher average strength will have an average equal to or *greater than* that of the present material. Such information will help select a one-tailed or two-tailed test to make the β error as small as possible. The following guidelines are based on the analysis of OC curves:

Use a one-tailed test with the entire α risk in the right tail if (1) it is known that (if μ_0 is not true) the true mean is $>\mu_0$ or (2) values of the population mean $<\mu_0$ are acceptable and we are interested only in detecting a population mean $>\mu_0$. [Use a one-tailed test with the entire α risk in the left tail if (1) it is known that (if μ_0 is not true) the true mean is $<\mu_0$ or (2) values of the population mean $>\mu_0$ are acceptable and we are interested only in detecting a population mean $<\mu_0$.]

Use a two-tailed test if (1) there is no prior knowledge on the location of the true population mean or (2) we are interested in detecting a true population mean $<$ or $>$ the μ_0 stated in the original hypothesis. (With a two-tailed test, the hypothesis is sometimes stated as the original hypothesis $H_0:\mu_0 = 30.0$ against the alternative hypothesis $H_1:\mu_0 \neq 30.0$. With a one-tailed test, $H_0:\mu_0 = 30.0$ against the alternative $H_1:\mu_1 < 30.0$ if α is placed in the left tail or $H_1:\mu_1 > 30.0$ if α is placed in the right tail.)

Every test of hypothesis has an OC curve. Duncan (1974) and Natrella (1963) are good sources of OC curves. [Some references present "power" curves, but power is simply 1 − (the probability of acceptance) = 1 − β.]

With this background, our discussion now proceeds to the steps for testing a hypothesis.

Testing a Hypothesis When the Sample Size Is Fixed in Advance. Ideally, desired values for the type I and type II errors are defined in advance and the required sample size determined (see later discussion on Determining the Sample Size Required for Testing a Hypothesis). If the sample size is fixed in advance because of cost or time limitations, then usually the desired type I error is defined and the following procedure is followed:

1. State the hypothesis.
2. Choose the type I error. Common values are 0.01, 0.05, or 0.10.
3. Choose the test statistic for testing the hypothesis.
4. Determine the acceptance region for the test, i.e., the range of values of the test statistic that result in a decision to accept the hypothesis.
5. Obtain the sample of observations, compute the test statistic, and compare the value to the acceptance region to make a decision to accept or reject the hypothesis.
6. Draw an engineering conclusion.

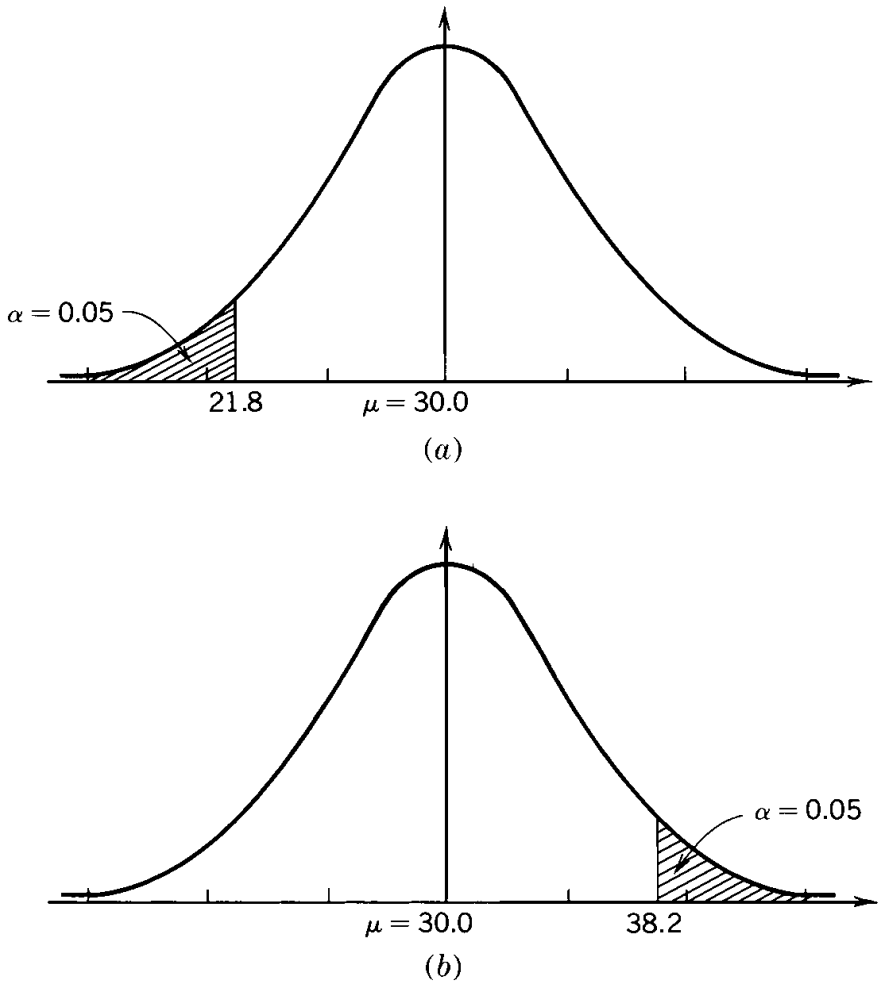


FIGURE 44.22 (*a*) Entire 5 percent error on left tail. (*b*) Entire 5 percent on right tail.

In the case of fixed sample size, the hypothesis is said to be "accepted" in the sense that there is insufficient evidence for the hypothesis to be rejected. However, this does not mean that the hypothesis is true. (See the introductory paragraph of Statistical Tests of Hypotheses above and Drawing Conclusions from Tests of Hypotheses below.) For this reason, one often uses the terminology "fail to reject" rather than "accept."

Table 44.31 summarizes some common tests of hypotheses. [These are tests commonly mentioned in the literature and used in practice. A number of them assume a normal distribution, Harter and Dubey (1967) define tests for the mean and variance but assuming a Weibull distribution—which really covers a family of distributions.] The procedure is illustrated through the following examples. Further examples and elaboration of the procedure are provided in Dixon and Massey (1969), Johnson and Leone (1964), Duncan (1974), Dudewicz (1976), and Natrella (1963). Table 44.31 lists a few unique or additional references for certain tests. For nonparametric tests, see especially Gibbons (1997).

Example: Tests on eight units of an experimental engine showed that they operated, respectively, for 28, 26, 31, 29, 25, 27, 28, and 27 min with 1 liter of a certain kind of fuel. A proposed specification states that the engine must operate for an average of at least 30 min. Does the engine meet the requirement? Assume a 5 percent significance level.

Solution: Using Test 1*b* of Table 44.31,

$$H_0: \quad \mu = 30.0$$

$$H_1: \quad \mu < 30.0$$

Test statistic:

$$t_1 = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

Acceptance region:

$$\text{Degrees of freedom DF} = 8 - 1 = 7$$

$$t \geq -1.895$$

A mathematical derivation of *degrees of freedom* is beyond the scope of this handbook, but the underlying concept can be stated. *Degrees of freedom* is a measure of the assurance involved when a sample standard deviation is used to estimate the true standard deviation of a universe. When the true standard deviation is known, DF = ∞. More generally, DF equals the number of measurements used to determine the sample standard deviation minus the number of constants estimated from the data in order to compute the standard deviation. In this example, it was necessary to estimate only one constant (the sample average) in order to compute the standard deviation, therefore DF = 8 − 1.

Analysis:

$$\overline{X} = 27.6 \qquad s = 1.86$$

$$t = \frac{27.6 - 30.0}{\dfrac{1.86}{\sqrt{8}}} = -3.65$$

Conclusion: Reject the hypothesis. There is sufficient evidence to conclude that the engine does not meet the requirement.

TABLE 44.31 Summary Table of Tests of Hypotheses

Test statistic and its distribution	Assumptions	Remarks
Test 1. The mean of a population is equal to μ_0 (H:$\mu = \mu_0$).		
(*a*) $U = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$ Normal distribution (Appendix II, Table B)	X is distribution-free but should be continuous and have only one mode.	Standard deviation of population is known.
(*b*) $t = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$ t distribution with DF = $n - 1$ (Appendix II, Table G)	X is normally distributed.	Standard deviation of population is estimated by sample s.
(*c*) $\tau_1 = \dfrac{\overline{X} - \mu_0}{R}$ τ_1 distribution (Appendix II, Table H)	X is normally distributed.	Dispersion of the population is estimated from sample range.
(*d*) r = number of occurrences of less frequent sign for a two-tail test (or number of positive or negative signs for a one-tail test) Distribution of r (Appendix II, Table I)	Distribution-free but population should be continuous and symmetrical.	Data are analyzed by evaluating signs of ($X_i - \mu_0$). (This is a "sign" test; see Gibbons, 1997, p. 112.)

TABLE 44.31 Summary Table of Tests of Hypotheses *(Continued)*

Test statistic and its distribution	Assumptions	Remarks
Test 2. The means of two populations are equal (H:$\mu_1 = \mu_2$).		
(*a*) $U = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$ Normal distribution (Appendix II, Table B)	X_1 and X_2 are distribution-free but should be continuous and have only one mode. If the populations are not normally distributed, sample sizes n_1 and n_2 should be large so that sampling distribution of U is approximately normal.	Standard deviations of populations are known.
(*b*) $t = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}\ \sqrt{\dfrac{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{n_1 + n_2 - 2}}}$ t distribution with DF $= n_1 + n_2 - 2$ (Appendix II, Table G)	$\sigma_1 = \sigma_2$. X_1 and X_2 are normally distributed.	Standard deviations of populations are estimated by sample s_1 and s_2.
(*c*) $\tau_d = \dfrac{\overline{X}_1 - \overline{X}_2}{0.5(R_1 + R_2)}$ Distribution of τ_d (Appendix II, Table J)	X_1 and X_2 are normally distributed.	Dispersions of populations are estimated by sample ranges.

TABLE 44.31 Summary Table of Tests of Hypotheses *(Continued)*

Test statistic and its distribution	Assumptions	Remarks
Test 2. The means of two populations are equal ($H:\mu_1 = \mu_2$).		
(*d*) $t' = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$ t distribution with DF = $\min(n_1 - 1, n_2 - 1)$ (Appendix II, Table G)	X_1 and X_2 are normally distributed.	Standard deviations of populations are estimated by sample s_1 and s_2 (no assumption that $\sigma_1 = \sigma_2$).
(*e*) $t = \dfrac{\overline{d}}{s_d/\sqrt{n}}$ t distribution with DF = number of pairs − 1 (Appendix II, Table G)	Populations are normally distributed.	Data are taken in n pairs, and difference d within each pair is calculated.
(*f*) r = number of occurrences of less frequent sign for a two-tail test (or number of positive or negative signs for a one-tail test) Distribution of r (Appendix II, Table I)	n_1 and n_2 each ≥ 6. Populations must be continuous and symmetrical; each of the two observations in a pair has been obtained under similar conditions.	Data are analyzed by evaluating signs of $(X_1 - X_2)$. (This is a "sign" test.) Distribution of r (Appendix II, Table I)
Test 3. Two characteristics on one product have specified population means ($H:\mu_X = \mu_{X0}$ and $\mu_Y = \mu_{Y0}$) (Johnson and Leone, 1964).		
$T^2 = \dfrac{s_Y^2}{s_X^2 s_Y^2 - s_{XY}}(X - \mu_{X0})^2 - \dfrac{2s_{XY}}{s_X^2 s_Y^2 - s_{XY}} \cdot (X - \mu_{X0})(Y - \mu_{Y0}) + \dfrac{s_X^2}{s_X^2 s_Y^2 - s_{XY}}(Y - \mu_{Y0})^2$ where $s_{XY} = \dfrac{n\Sigma(XY) - \Sigma X \Sigma Y}{n(n-1)}$ where critical value for $T^2 = \dfrac{2(n-1)F}{n-2}$ and where the F distribution has $DF_1 = 2$; $DF_2 = n - 2$ (Appendix II, Table K)	Bivariate normal distribution.	Population standard deviations are estimated by s_X and s_Y (T^2 is called Hotelling's T^2).

TABLE 44.31 Summary Table of Tests of Hypotheses *(Continued)*

Test statistic and its distribution	Assumptions	Remarks
Test 4. The standard deviation of a population is equal to σ_0 (H:$\sigma = \sigma_0$).		
$\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$ Chi-square distribution with DF = $n - 1$ (Appendix II, Table L)	Population is normally distributed.	Standard deviation of population is estimated by sample s.
Test 5. The standard deviations of two populations are equal (H:$\sigma_1^2 = \sigma_2^2$).		
(*a*) $F = \dfrac{s_1^2}{s_2^2}$ F distribution with $DF_1 = n_1 - 1$ and $DF_2 = n_2 - 1$ (Appendix II, Table K)	Populations are normally distributed.	Standard deviations of populations are estimated by sample s_1 and s_2.
(*b*) $F' = \dfrac{R_1}{R_2}$ Distribution of F' (Appendix II, Table M)	Distribution-free but X_1 and X_2 should be continuous.	Dispersions of populations are estimated by sample ranges.
Test 6. The proportion of a population exhibiting a certain characteristic is p_0 (H:$p = p_0$).		
(*a*) $U = \dfrac{X - np_0}{\sqrt{np_0(1-p_0)}}$ Normal distribution (Appendix II, Table B)	$n \geq 100$. Only for large sample sizes.	Proportion of population is estimated by sample proportion.
(*b*) Determine confidence limits (Appendix II, Table N) and observe if p_0 falls within the limits		Proportion of population is estimated by sample proportion (useful when $n < 100$).

TABLE 44.31 Summary Table of Tests of Hypotheses *(Continued)*

Test statistic and its distribution	Assumptions	Remarks
Test 7. The proportions in two populations are equal (H:$p_1 = p_2$).		
$U = \dfrac{X_1/n_1 - X_2/n_2}{\sqrt{\hat{p}\,(1-\hat{p})(1/n_1 + 1/n_2)}}$ where $\hat{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$ Normal distribution (Appendix II, Table B)	$np > 5$ for each population. Sample sizes n_1 and n_2 must be large so that sampling distribution of U is approximately normal.	Proportions in populations are estimated by sample proportions.
Test 8. Proportion of correct decisions on a sensory evaluation is p_0 (H:$p = p_0$).		
(*a*) $U = \dfrac{X/n - 0.50}{\sqrt{0.25/n}}$ Normal distribution (Appendix II, Table B)	$p = 0.5$, n should be $>$ 30.	Judge is asked to identify which of two specimens is the same as a control specimen originally given to him or her. (This is a "duo-trio" test.)
(*b*) $U = \dfrac{X/n - 0.33}{\sqrt{0.22/n}}$ Normal distribution (Appendix II, Table B)	$p = 0.33$, n should be $>$ 30.	Judge is asked to identify which of three specimens is different from the other two. (This is a "triangle" test.)

TABLE 44.31 Summary Table of Tests of Hypotheses *(Continued)*

Test statistic and its distribution	Assumptions	Remarks
Test 9. Samples are from identically distributed populations [H:$F(X_1) = F(X_2)$ where $F(X)$ = distribution function] (Johnson and Leone, 1964).		
(*a*) For evaluation of means T' = sum of ranks in smaller sample Distribution of T' (Appendix II, Table O)	Distribution-free; if $n_1 > 8$ and $n_2 > 8$, the distribution of statistic T' can be closely approximated by normal.	Data are evaluated by ranking the combined observations from the two samples (1 for the smallest, 2 for the next smallest, etc.). Then calculate the sum of the ranks in the smaller sample. A rejected hypothesis leads to the conclusion that the means are different. (This is a "rank sum" test.)
(*b*) For evaluation of standard deviations T' = sum of ranks in smaller sample Distribution of T' (Appendix II, Table O)	Distribution-free; if $n_1 > 8$, and $n_2 > 8$, the distribution of statistic T' can be closely approximated by normal.	Data are evaluated by ranking the combined observations but ranking assigns 1 to smallest observation, 2 to largest observation, 3 to next smallest, etc. The sum of the ranks for the smaller sample is calculated. A rejected hypothesis means that the standard deviations are different. (This is a "rank sum" test.)

TABLE 44.31 Summary Table of Tests of Hypotheses *(Continued)*

Test statistic and its distribution	Assumptions	Remarks
Test 10. The observations in a sample have been randomly drawn from a single population (Bennett and Franklin, 1954, and Dixon and Massey, 1969).		
(a) u = number of runs Distribution of runs (Appendix II, Table P)	$n_1 \geq 10$, $n_2 \geq 10$.	Data are evaluated in terms of number of sequences or "runs" above and below the median. (This is a "runs" test.)
(b) $M = \dfrac{\sum_{i=1}^{n-1} (X_{i+1} - \overline{X}_i)^2}{\Sigma(X_i - \overline{X})^2}$ $U = \dfrac{1 - M/2}{\sqrt{(n-2)/[(n-1)(n+1)]}}$ Normal distribution (Appendix II, Table B)	1. $n \geq 4$. 2. Normal population.	Data are evaluated in terms of differences between successive observations in a sequence $(X_{i+1} - X_i)$. (This is the "mean square successive difference, M, test.")
Test 11. An observation does belong to the same population as the other observations in a sample (Dixon and Massey, 1969; Grubbs, 1969).		
Both X_1 and X_2 are to be evaluated if extreme observations in either direction are undesirable. $3 \leq n \leq 7$: $r_{10} = \dfrac{X_2 - X_1}{X_n - X_1}$ $8 \leq n \leq 10$: $r_{11} = \dfrac{X_2 - X_1}{X_{n-1} - X_1}$ $11 \leq n \leq 13$: $r_{21} = \dfrac{X_3 - X_1}{X_{n-1} - X_1}$ $14 \leq n \leq 25$: $r_{22} = \dfrac{X_3 - X_1}{X_{n-2} - X_1}$ Distribution of r (Appendix II, Table Q)	X is normally distributed. Population mean and standard deviation are unknown.	Data are evaluated by arranging data in order of magnitude and comparing the distance of one extreme observation from other observations with a measure of variability. (This is an outlier test.)

TABLE 44.31 Summary Table of Tests of Hypotheses (*Continued*)

Test statistic and its distribution	Assumptions	Remarks
Test 12. A sample of data comes from a population with the specified probability function.		
(*a*) D = largest deviation of actual % cumulative frequency from theoretical % cumulative frequency Distribution of D (Gibbons, 1997, Table C)	Distribution should be continuous.	Data are evaluated by first plotting on probability paper. The largest deviation of a plotted point from a straight line is then evaluated. This test applies particularly where $n < 30$. (This is the "Kolmogorov-Smirnov goodness-of-fit test.")
(*b*) $\chi^2 = \Sigma \frac{(f_a - f_e)^2}{f_e}$ Distribution — DF Normal — Number of cells minus 3 Exponential — Number of cells minus 2 Weibull — Number of cells minus 4 Poisson — Number of cells minus 2 Binomial — Number of cells minus 2 Test statistics for all distributions follow the chi square distribution (Appendix II, Table L)	$n > 30$ and preferably > 100.	Data are evaluated by first constructing a frequency distribution. Theoretical frequencies (based on the distribution assumption) are calculated for each cell. The actual (f_a) and theoretical (f_e) frequencies are then compared. This test can be used for continuous and discrete distributions. If any theoretical frequency is less than 5, the cell involved should be combined with one or more adjacent cells. (This is the x^2 goodness-of-fit test.)

Example: *Solve the previous example using the range instead of the standard deviation.*

Solution: *Using Test 1c of Table 44.31,*

$$H_0: \quad \mu = 30.0$$

$$H_1: \quad \mu < 30.0$$

Test statistic:

$$\tau_1 = \frac{\overline{X} - \mu_0}{R}$$

Acceptance region:

$$\tau_1 \geq -0.230$$

Analysis:

$$\tau_1 = \frac{27.6 - 30.0}{6} = -0.40$$

Conclusion: Reject the hypothesis. There is sufficient evidence to conclude that the engine does not meet the requirement.

Example: Solve the previous example using the sign test.

Solution: Using Test 1d of Table 44.31,

$$H_0: \quad \mu = 30$$

$$H_1: \quad \mu < 30$$

Test statistic: Number of positive signs r.

Acceptance region:

$$r > 1 \text{ (one-tailed test)}$$

Analysis:

X	$X - \mu_0$	
28	−	
26	−	
31	+	
29	−	$r = 1$
25	−	
27	−	
28	−	
27	−	

Conclusion: Reject the hypothesis. There is sufficient evidence to conclude that the engine does not meet the requirement.

Example: Five batches of rubber were made by each of two recipes and tested for tensile strength with the following results:

Recipe 1	*Recipe 2*
3067	3200
2730	2777
2840	2623
2913	3044
2789	2834

Test the hypothesis that average strength is the same for the two recipes. Assume a 5 percent significance level.

Solution: First, Test 5a of Table 44.31 tests the assumption of equal variances. The outcome of this is used to decide whether to use Test 2b or 2d to evaluate the question about average strength.

$$H_0: \quad \sigma_1^2 = \sigma_2^2$$

$$H_1: \quad \sigma_1^2 \neq \sigma_2^2$$

Test statistic:

$$DF_1 = 5 - 1 = 4 \qquad DF_2 = 5 - 1 = 4$$

$$F = \frac{(s_1)^2}{(s_2)^2}$$

Acceptance region:

$$\frac{1}{9.60} \leq F \leq 9.60$$

Analysis:

$$s_1^2 = 16{,}923.7$$

$$s_2^2 = 51{,}713.3$$

$$F = \frac{16{,}923.7}{51{,}713.3} = 0.33$$

Conclusion: Accept the hypothesis. This is used to satisfy the assumption of equal variances in the following test of hypothesis. Now, using Test 2b,

$$H_0: \quad \mu_1 = \mu_2$$

$$H_1: \quad \mu_1 \neq \mu_2$$

Test statistic:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{n_1 + n_2 - 2}}}$$

Acceptance region:

$$\text{DF} = 5 + 5 - 2 = 8$$

$$-2.306 \le t \le +2.306$$

Analysis:

$$t = \frac{2867.8 - 2895.6}{\sqrt{\frac{1}{5} + \frac{1}{5}}\sqrt{\frac{(5-1)16{,}923.7 + (5-1)51{,}713.3}{5+5-2}}} = -0.2373$$

Conclusion: Accept the hypothesis. There is insufficient evidence to conclude that the recipes differ in average strength.

Example: Solve the previous example using ranges instead of standard deviations.

Solution: First, Test 5b of Table 44.31 tests the assumption of equal variances.

$$H_0: \quad \sigma_1^2 = \sigma_2^2$$

$$H_1: \quad \sigma_1^2 \neq \sigma_2^2$$

Test statistic:

$$F' = \frac{R_1}{R_2}$$

Acceptance region:

$$0.32 < F' < 3.2$$

Analysis:

$$R_1 = 3067 - 2730 = 337$$

$$R_2 = 3200 = 2623 = 577$$

$$F' = \frac{337}{577} = 0.58$$

Conclusion: Accept the hypothesis. This is used to satisfy the assumption of equal variances in the following test of hypothesis. Now, using Test 2c,

$$H_0: \quad \mu_1 = \mu_2 \quad H_1: \quad \mu_1 \neq \mu_2$$

Test statistic:

$$\tau_d = \frac{\bar{X}_1 - \bar{X}_2}{0.5(R_1 + R_2)}$$

Acceptance region:

$$-0.493 \le \tau_d \le +0.493$$

$$R_1 = 3{,}067 - 2{,}730 = 337$$

$$R_2 = 3{,}200 - 2{,}623 = 577$$

Analysis:

$$\tau_d = \frac{2867.8 - 2895.6}{0.5(337 + 577)} = -0.061$$

Conclusion: Accept the hypothesis. There is insufficient evidence to conclude that the recipes differ in average strength.

Testing a Hypothesis When the Sample Size Is Not Fixed in Advance. As noted earlier under Testing a Hypothesis When the Sample Size Is Fixed in Advance, ideally the desired values of type I and type II errors are defined in advance and the required sample size determined. However, if the sample size is fixed in advance, this cannot be done in most cases, and one can control only the type I error (not the type II error). One can control both when the sample size is not fixed in advance, as illustrated now for the case of testing if the means of two populations are equal ($H : \mu_1 = \mu_2$).

For type I error equal to α and type II error equal to β when the means differ by δ^* (a positive number) in absolute value, one proceeds as follows:

Sample n_0 observations from each of the ($k = 2$) populations (it is desirable that n_0 be at least 10 if possible).

Determine the total sample size for population i as

$$n_i = max[n_0 + 1, (ws_i)^2]$$

where w solves the equation (solve by trial and error)

$$P_{n_0}(-h - \delta^* w) + P_{n_0}(-h + \delta^* w) = \beta$$

where $h = h_{n_0}(2,1 - \alpha/2)$ is found from Table 44.30, and $P_{n_0}(t)$ is tabled in Table 44.32.

Take $n_1 - n_0$ more observations from population 1 and $n_2 - n_0$ more observations on population 2.

Reject the hypothesis that the means are equal if the sample means (based on all the data) differ by more than h/w.

As an example, suppose we have an initial sample size of $n_0 = 15$ and desire type I error $\alpha = .05$ and type II error $\beta = .10$ when the means differ by $\delta^* = 4.0$. Then from Table 44.30, $h = h_{n_0}(2,1 - \alpha/2) = h_{15}(2, .975) = 3.02$. We find the w that solves

$$P_{15}(-3.02 - 4w) + P_{15}(-3.02 + 4w) = .10$$

in two steps. First, we find w approximately by solving $P_{15}(-3.02 + 4w) = .10$ using Table 44.32. Since Table 44.32 has entries of .4999 and larger, we first convert the equation using the fact that $P_{n_0}(-v) = 1 - P_{n_0}(v)$ for all v, obtaining

$$1 - P_{15}(3.02 - 4w) = .10 \qquad \text{or} \qquad P_{15}(3.02 - 4w) = .90$$

Since $P_{15}(1.9) = .8968$ and $P_{15}(2.0) = .9079$, w will (to a first approximation) be in the range of the solutions of

$$3.02 - 4w = 1.9 \qquad \text{and} \qquad 3.02 - 4w = 2.0$$

namely, $w = (3.02 - 1.9)/4 = 0.28$ and $w = (3.02 - 2.0)/4 = 0.255$. Thus let us try a value of $w = 0.255$. For this value,

$$P_{15}[-3.02 - (4)(0.255)] + P_{15}[-3.02 + (4)(0.255)] = P_{15}(-4.04) + P_{15}(-2.0) = 1 - P_{15}(4.04) + 1 - P_{15}(2.0) = 1 - .9946 + 1 - .9079 = .0054 + .0921 = .0975 = .10$$

TABLE 44.32 Probabilities $P_{n_0}(t)$ Needed for Testing if Two Means Are Equal in Two Stages*

	n_0							
t	2	3	4	5	6	7	8	9
0.0	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
0.1	.5158	.5208	.5229	.5241	.5248	.5254	.5257	.5260
0.2	.5317	.5415	.5457	.5481	.5496	.5506	.5514	.5520
0.3	.5473	.5620	.5684	.5719	.5742	.5757	.5768	.5777
0.4	.5628	.5822	.5907	.5954	.5984	.6004	.6019	.6030
0.5	.5779	.6021	.6126	.6184	.6221	.6246	.6265	.6279
0.6	.5927	.6215	.6340	.6410	.6453	.6483	.6505	.6521
0.7	.6071	.6404	.6549	.6629	.6679	.6713	.6738	.6757
0.8	.6211	.6588	.6751	.6841	.6897	.6936	.6964	.6986
0.9	.6345	.6765	.6946	.7046	.7108	.7151	.7182	.7206
1.0	.6475	.6935	.7134	.7242	.7311	.7357	.7391	.7417
1.1	.6600	.7099	.7313	.7431	.7504	.7554	.7591	.7618
1.2	.6720	.7255	.7485	.7610	.7688	.7742	.7780	.7810
1.3	.6834	.7405	.7648	.7781	.7863	.7919	.7960	.7991
1.4	.6943	.7547	.7803	.7942	.8029	.8087	.8130	.8162
1.5	.7048	.7681	.7950	.8094	.8184	.8245	.8289	.8322
1.6	.7147	.7809	.8088	.8238	.8331	.8393	.8439	.8473
1.7	.7242	.7930	.8218	.8372	.8467	.8532	.8578	.8613
1.8	.7332	.8044	.8340	.8498	.8595	.8660	.8707	.8742
1.9	.7418	.8151	.8454	.8615	.8714	.8780	.8827	.8863
2.0	.7499	.8252	.8561	.8725	.8824	.8891	.8938	.8974
2.1	.7577	.8347	.8661	.8826	.8926	.8993	.9040	.9075
2.2	.7651	.8436	.8754	.8920	.9020	.9087	.9134	.9169
2.3	.7721	.8520	.8841	.9007	.9107	.9173	.9220	.9254
2.4	.7788	.8599	.8922	.9087	.9187	.9252	.9298	.9332
2.5	.7852	.8673	.8996	.9162	.9260	.9324	.9369	.9402
2.6	.7912	.8742	.9066	.9230	.9327	.9390	.9434	.9466
2.7	.7970	.8807	.9130	.9293	.9388	.9450	.9493	.9524
2.8	.8025	.8868	.9190	.9350	.9444	.9504	.9546	.9576
2.9	.8078	.8925	.9245	.9403	.9495	.9553	.9594	.9623
3.0	.8128	.8979	.9297	.9452	.9541	.9598	.9637	.9665
3.1	.8176	.9029	.9344	.9497	.9583	.9638	.9676	.9703
3.2	.8221	.9076	.9388	.9537	.9622	.9675	.9711	.9736
3.3	.8265	.9121	.9429	.9575	.9657	.9708	.9742	.9766
3.4	.8307	.9162	.9466	.9609	.9688	.9737	.9770	.9793
3.5	.8347	.9201	.9501	.9640	.9717	.9764	.9795	.9817
3.6	.8385	.9238	.9533	.9669	.9743	.9788	.9817	.9838
3.7	.8422	.9273	.9563	.9695	.9766	.9809	.9837	.9857
3.8	.8457	.9305	.9591	.9719	.9788	.9829	.9855	.9874
3.9	.8491	.9336	.9617	.9741	.9807	.9846	.9871	.9889
4.0	.8524	.9365	.9640	.9761	.9824	.9861	.9885	.9902
4.1	.8555	.9392	.9662	.9779	.9840	.9875	.9898	.9913
4.2	.8585	.9417	.9683	.9796	.9854	.9888	.9909	.9923
4.3	.8614	.9442	.9702	.9812	.9867	.9899	.9919	.9932
4.4	.8641	.9464	.9720	.9826	.9879	.9909	.9928	.9940
4.5	.8668	.9486	.9736	.9839	.9890	.9918	.9936	.9947
4.6	.8694	.9506	.9751	.9851	.9899	.9926	.9943	.9953
4.7	.8719	.9526	.9765	.9862	.9908	.9933	.9949	.9959
4.8	.8743	.9544	.9779	.9872	.9916	.9940	.9954	.9964
4.9	.8766	.9561	.9791	.9881	.9923	.9946	.9959	.9968
5.0	.8788	.9578	.9803	.9889	.9930	.9951	.9964	.9972
5.1	.8810	.9593	.9813	.9897	.9935	.9956	.9967	.9975

TABLE 44.32 Probabilities $P_{n_0}(t)$ Needed for Testing if Two Means Are Equal in Two Stages* *(Continued)*

	n_0								
t	10	11	12	13	14	15	20	25	30
0.0	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
0.1	.5262	.5264	.5266	.5267	.5268	.5269	.5272	.5274	.5275
0.2	.5524	.5528	.5531	.5533	.5535	.5537	.5544	.5547	.5555
0.3	.5783	.5789	.5793	.5797	.5800	.5803	.5812	.5818	.5822
0.4	.6039	.6046	.6052	.6057	.6061	.6065	.6077	.6085	.6089
0.5	.6290	.6299	.6306	.6312	.6317	.6322	.6337	.6346	.6352
0.6	.6534	.6545	.6554	.6561	.6567	.6572	.6591	.6601	.6608
0.7	.6772	.6785	.6794	.6803	.6810	.6816	.6837	.6849	.6857
0.8	.7003	.7016	.7027	.7037	.7045	.7051	.7075	.7089	.7098
0.9	.7224	.7239	.7252	.7262	.7271	.7278	.7304	.7319	.7329
1.0	.7437	.7453	.7466	.7478	.7487	.7495	.7523	.7539	.7550
1.1	.7640	.7657	.7671	.7683	.7693	.7702	.7732	.7749	.7761
1.2	.7833	.7851	.7866	.7879	.7889	.7899	.7930	.7948	.7961
1.3	.8015	.8034	.8050	.8063	.8075	.8084	.8117	.8136	.8149
1.4	.8187	.8207	.8223	.8237	.8249	.8259	.8293	.8313	.8326
1.5	.8348	.8369	.8386	.8400	.8412	.8422	.8457	.8478	.8491
1.6	.8499	.8520	.8538	.8552	.8564	.8575	.8611	.8631	.8645
1.7	.8639	.8661	.8679	.8693	.8706	.8716	.8753	.8774	.8788
1.8	.8770	.8792	.8809	.8824	.8837	.8847	.8884	.8905	.8919
1.9	.8890	.8912	.8930	.8945	.8957	.8968	.9004	.9025	.9039
2.0	.9001	.9023	.9041	.9056	.9068	.9079	.9115	.9136	.9149
2.1	.9103	.9124	.9142	.9157	.9169	.9180	.9215	.9236	.9249
2.2	.9196	.9217	.9235	.9249	.9261	.9272	.9307	.9327	.9340
2.3	.9281	.9302	.9319	.9333	.9345	.9355	.9389	.9409	.9421
2.4	.9358	.9379	.9395	.9409	.9421	.9430	.9464	.9482	.9495
2.5	.9428	.9448	.9464	.9478	.9489	.9498	.9530	.9548	.9560
2.6	.9491	.9511	.9526	.9539	.9550	.9559	.9590	.9607	.9618
2.7	.9548	.9567	.9582	.9594	.9605	.9614	.9643	.9659	.9670
2.8	.9599	.9617	.9632	.9644	.9654	.9662	.9690	.9706	.9716
2.9	.9645	.9663	.9676	.9688	.9697	.9705	.9732	.9746	.9756
3.0	.9686	.9703	.9716	.9727	.9736	.9743	.9768	.9782	.9791
3.1	.9723	.9739	.9751	.9761	.9770	.9777	.9800	.9813	.9821
3.2	.9756	.9771	.9782	.9792	.9800	.9807	.9828	.9841	.9848
3.3	.9785	.9799	.9810	.9819	.9826	.9833	.9853	.9864	.9871
3.4	.9811	.9824	.9834	.9843	.9850	.9855	.9874	.9885	.9891
3.5	.9833	.9846	.9856	.9864	.9870	.9875	.9893	.9902	.9908
3.6	.9854	.9865	.9874	.9882	.9888	.9893	.9909	.9917	.9923
3.7	.9871	.9882	.9891	.9898	.9903	.9908	.9923	.9930	.9935
3.8	.9887	.9897	.9905	.9912	.9917	.9921	.9934	.9941	.9946
3.9	.9901	.9911	.9918	.9924	.9928	.9932	.9944	.9951	.9955
4.0	.9913	.9922	.9929	.9934	.9938	.9942	.9953	.9959	.9962
4.1	.9924	.9932	.9938	.9943	.9947	.9950	.9960	.9966	.9969
4.2	.9933	.9941	.9947	.9951	.9955	.9958	.9967	.9971	.9974
4.3	.9942	.9949	.9954	.9958	.9961	.9964	.9972	.9976	.9979
4.4	.9949	.9955	.9960	.9964	.9967	.9969	.9976	.9980	.9982
4.5	.9955	.9961	.9966	.9969	.9972	.9974	.9980	.9983	.9985
4.6	.9961	.9966	.9970	.9973	.9976	.9978	.9983	.9986	.9988
4.7	.9966	.9971	.9974	.9977	.9979	.9981	.9986	.9989	.9990
4.8	.9970	.9974	.9978	.9980	.9982	.9984	.9988	.9991	.9992
4.9	.9974	.9978	.9981	.9983	.9985	.9986	.9990	.9992	.9993
5.0	.9977	.9981	.9983	.9985	.9987	.9988	.9992	.9993	.9994
5.1	.9980	.9983	.9986	.9987	.9989	.9990	.9993	.9995	.9995

*The table entries are from p.52 of E.J. Dudewicz and S.R. Dalal (1975), "Allocation of Observations in Ranking and Selection With Unequal Variances." *Sankhyā*, vol. 73B, pp.28-78. Acknowledgment is made to the Indian Statistical Institute for permission to reproduce these tables.

(The approximation process could be carried further, but for most practical purposes this w will suffice.) Then, if $s_1 = 17.3$, we will need a total sample of size

$$n_1 = \max\{15 + 1, [(0.255)(17.3)]^2\} = \max(16, 19.46) = 20$$

(since sample sizes must be integers, we round up). Since we already have 15 observations, $20 - 15 = 5$ more will be required from population 1. Similarly, if $s_2 = 10.4$, we will need a total sample of size

$$n_2 = \max\{15 + 1, [(0.255)(10.4)]^2\} = \max(16, 7.03) = 16$$

Since we already have 15 observations from population 2, $16 - 15 = 1$ more will be required. If the sample means based on all the data are $\overline{X}_1$ and $\overline{X}_2$, we reject the hypothesis that $\mu_1 = \mu_2$ if the sample means differ by more than

$$\frac{h}{w} = \frac{3.02}{0.255} = 11.84$$

For example, if $\overline{X}_1 = 38.3$ and $\overline{X}_2 = 50.2$, the sample means differ by 11.9, and we reject the null hypothesis that the means are equal.

Drawing Conclusions from Tests of Hypotheses. The payoff for these tests of hypotheses comes from reaching useful conclusions. The meaning of "Reject the hypothesis" or "Accept the hypothesis" is shown in Table 44.33 along with some analogies to explain subtleties of the meanings.

When a hypothesis is rejected, the practical conclusion is that "the parameter value specified in the hypothesis is wrong." This conclusion is made with strong conviction—roughly speaking at a confidence level of 100 $(1-\alpha)$ percent. The key question then is: Just what is a good estimate of the value of the parameter for the population? Help can be provided on this question by calculating the "confidence limits" for the parameter discussed under Statistical Estimation: Confidence Interval Estimates.

TABLE 44.33 The Meaning of a Conclusion from Tests of Hypotheses

	If hypothesis is rejected	If hypothesis is accepted
Adequacy of evidence in the sample of observations	Sufficient to conclude that hypothesis is false	Not sufficient to conclude that hypothesis is false; hypothesis is a reasonable one but has *not* been proved to be true
Difference between sample result (e.g., $\overline{X}$) and hypothesis value (e.g., μ_0)	Unlikely that difference was due to chance (sampling) variation	Difference could easily have been due to chance (sampling) variation
Analogy of guilt or innocence in a court of law	Guilt has been established beyond a reasonable doubt	Have not established guilt beyond a reasonable doubt
Analogy of a batting average in baseball	If player got 300 base hits out of 1000 times at bat, this is sufficient to conclude that the overall batting average is about 0.300	If player got 3 hits in 10 times, this is not sufficient to conclude that the overall average is about 0.300

When a hypothesis is accepted, the numerical value of the parameter stated in the hypothesis has not been proved, but it has not been disproved. It is *not* correct to say that the hypothesis has been proved as correct at the 100 $(1 - \alpha)$ percent confidence level. Many other hypotheses could be accepted for the given sample of observations and yet only one hypothesis can be true. Therefore, an acceptance does *not* mean a high probability of proof that a specific hypothesis is correct. (All other factors being equal, the smaller the sample size, the more likely it is that the hypothesis will be accepted. Less evidence certainly does not imply proof.) For this reason, often today the wording used is "the hypothesis was not rejected at level of significance α" rather than "the hypothesis was accepted at level α." Only when the sample size is not fixed in advance do we have an indication that the true value does not differ from the hypothesized value by more than δ^* with risk β.

With an acceptance of a hypothesis, a key question then is: What conclusion, if any, can be drawn about the parameter value in the hypothesis? Two approaches are suggested:

1. Calculate confidence limits on the sample result (see the previous topic of Statistical Estimation). These confidence limits define an interval within which the true population parameter lies. If this interval is small, then an acceptance decision on the test of hypothesis means that the true population value is either equal to or close to the value stated in the hypothesis. Then it is reasonable to act as if the parameter value specified in the hypothesis is in fact correct. If the confidence interval is relatively wide, then this is a stern warning that the value stated in the hypothesis has not been proved and that the true value of the population might be far different from that specified in the hypothesis.

2. Construct and review the operating characteristic curve for the test of hypothesis. This defines the probability that other possible values of the population parameter could have been accepted by the test. Knowing these probabilities for values relatively close to the original hypothesis can help draw further conclusions about the acceptance of the original hypothesis. For example, Figure 44.21 shows the OC curve for a hypothesis that specified that the population mean is 30.0. Note that the probability of accepting the hypothesis when the population mean μ is 30.0 is 0.95 (or $1 - \alpha$). Also note that if μ really is 35.0, then the probability of accepting $\mu = 30.0$ is still high (about 0.83). If μ really is 42.0, the probability of accepting $\mu = 30.0$ is only about 0.33.

Care must always be taken in drawing engineering conclusions from the statistical conclusions, particularly when a hypothesis is accepted. [Rutherford (1971) discusses a procedure for drawing conclusions which requires that a choice be made between two policies for drawing conclusions, i.e., conservative and liberal.]

Determining the Sample Size Required for Testing a Hypothesis. The previous subsections assumed that the sample size was fixed by nonstatistical reasons and that the type I error only was predefined for the test. The ideal procedure is to predefine the desired type I and type II errors and calculate the sample size required to cover both types of errors.

The sample size required will depend on (1) the sampling risks desired (α and β), (2) the size of the smallest true difference that is to be detected, and (3) the variation in the characteristic being measured. The sample size can be determined by using the "operating characteristic" curve for the test. Table 44.34 summarizes methods useful in determining the sample size required for two-sided tests of certain hypotheses. [Further sources of OC curves are Duncan (1974) and Natrella (1963).]

Suppose it were important to detect the fact that the average life of the batteries cited previously was 35.0. Specifically, be 80 percent sure of detecting this change ($\beta = 0.2$). Further, if the true average was 30.0 (as stated in the hypothesis), there should be only a 5 percent risk of rejecting the hypothesis ($\alpha = 0.05$). In using Appendix II, Chart R, d is defined as

$$d = \frac{\mu - \mu_0}{\sigma} = \frac{35.0 - 30.0}{10} = 0.5$$

Entering with $d = 0.5$ and $P_a = 0.2$ (the β risk), the curves indicate that a sample size of about 30 is required.

TABLE 44.34 Summary of Sample Size Graphs and Tables

Hypothesis	Graph or table
1. Mean of a population = μ_0 (σ known)	Appendix II, Chart R
2. Mean of a population = μ_0 (σ estimated by s)	Duncan (1974, p. 539)
3. Means of two populations are equal (σ_1 and σ_2 known)	Natrella (1963, pp. T-16, T-17)
4. Means of two populations are equal ($\sigma_1 = \sigma_2$ but estimated by s_1 and s_2)	Natrella (1963, pp. T-16, T-17)
5. Standard deviation of a population = σ_0	Duncan (1974, p. 324)
6. Standard deviations of two populations are equal	Duncan (1974, p. 572)

Duncan (1974) discusses the calculation of the sample size required to meet the type I and II errors. In practice, however, one is often not sure of desired values of these errors. Reviewing the operating characteristic curves for various sample sizes can help to arrive at a decision on the sample size required to reflect the relative importance of both risks. It is especially important to consider β as well as α, lest meaningless results be obtained. (Note that randomizing so as to reject H_0 100α percent of the time yields a test with level of significance α. That in itself, without consideration of β, is trivial.)

Relation to Confidence Intervals. Confidence limits provide a set of limits within which a population parameter lies (with specified probability). Tests of hypotheses evaluate a specific statement about a population parameter. These procedures are related, and most hypothesis tests also can be made using confidence limit calculations.

Example: A sample of 12 insulators has an average strength of 4.95 ft·lb (6.7149 N·m). The standard deviation of the population is known to be 0.25 ft·lb (0.34 N·m). It is desired to test the hypothesis that the population mean is 5.15 ft·lb (6.9834N·m).

Solution using tests of hypotheses: Table 44.31 defines the test statistic 1*a* and $U = (\overline{X} - \mu_0)/(\sigma/\sqrt{n})$, and U is normally distributed. If $\alpha = 0.05$, the acceptance region is a U between ± 1.96. Then

$$H_0: \quad \mu = \mu_0 = 5.15$$

$$H_1: \quad \mu \neq \mu_0$$

$$U = \frac{4.95 - 5.15}{0.25/\sqrt{12}} = -2.75$$

Since the sample index is outside the acceptance region, the hypothesis is rejected. The procedure using confidence limits is:

1. State the hypothesis concerning the value of a population parameter.
2. Obtain a sample of data and calculate confidence limits for the population parameter.
3. If the hypothesis value falls within the confidence limits, accept the hypothesis. If the hypothesis value falls outside the confidence limits, reject the hypothesis.

Solution using confidence limits: From Table 44.22, parameter 1, the confidence limits are

$$\overline{X} \pm K_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

The 95 percent confidence limits are $4.95 \pm 1.96\ (0.25/\sqrt{12}) = 4.81$ and 5.09. As the hypothesis value falls outside the confidence limits, the hypothesis is rejected. This is the same conclusion reached by using the hypothesis testing procedure.

Confidence limit concepts and tests of hypotheses are therefore alternative approaches to evaluating a hypothesis. [For certain hypotheses, these two approaches will result in slightly different type I errors (see Barr 1969).] As discussed under Drawing Conclusions from Tests of Hypotheses, confidence limits are a valuable supplement to the test of hypothesis procedure. For example, in the preceding example, not only do confidence limits tell us that μ is not 5.15, they also tell us that μ is between 4.81 and 5.09.

Standard Cases. Some of the most important practical cases have been covered in Table 44.31, namely:

Binomial proportion	Tests 6, 8
Two binomial proportions	Test 7
Normal mean	Test 1
Two means	Test 2
Bivariate normal mean	Test 3
One normal standard deviation	Test 4
Two standard deviations	Test 5
Two distributions are equal	Test 9
Random order of observations	Test 10
Test for outliers	Test 11

Many of these problems also can be solved using sequential tests. There, the sample size is not set in advance, but based on the data we decide how many observations are needed. For example, if one decided to inspect a fixed number of items such as 100 items and reject the lot if 15 or more defectives were found, clearly one could stop sampling as soon as the fifteenth defective were found. Similar ideas allow savings in numbers of observations in most of the standard cases listed above and are especially important when sampling is costly or time-consuming. For details, see Govindarajulu (1981).

Paired versus Unpaired Data. In Test 2*e* of Table 44.31 a test is given that is appropriate when data are taken in pairs and the difference within each pair is used as the basic data. This procedure is often used in order to "wash out" the effects of variables that are believed to have effects but whose effects we do not wish to study.

For example, suppose that there is an effect of the operator of the machine, and we wish to compare two types of operation on that machine—but do not wish to evaluate the size of the operator effect. Then by letting each operator perform both operations and taking the difference, we wash out the effect of the operator. (Whereas if one operator performed all of one procedure and another operator performed all of the other procedure, differences observed might be due not to the procedures but to the operators.)

As another example, in testing of mailing lists to evaluate competing advertising copy, often an "A/B split" is used. That is, one type of copy goes to names 1, 3, 5, 7,…on the list, while the other type goes to names 2, 4, 6, …. Since (on ZIP-code-ordered lists) adjacent listings may be expected to be more similar than entries far apart, this is an appropriate pairing.

This technique of pairing is *not* used when there is no advance pairing of the data. For example, it is an error to pair items by their sequence in a data listing (where often they may be sorted by some other characteristic).

Statistical Significance versus Practical Significance. Suppose we are using Test 1*a* of Table 44.31 to test the hypothesis that the mean is 30, and wish a two-tailed test with level of sig-

nificance 0.05. Then we will reject H if U is outside the interval $(-1.96, 1.96)$. If we find $U = 3.15$, one way of reporting the result of the test is to state that H was rejected at level 0.05.

Another way of reporting is to find for which level of significance the acceptance interval would be $(-3.15, 3.15)$. From Table B in Appendix II we see that level is (2) (0.00082) = 0.00164. This is called the *significance probability* of the test we have conducted. It is the smallest level of significance at which we would reject H for the data we observed.

One advantage of the significance probability is that if we use it, then we can report the significance probability as 0.00164 without choosing a level of significance. Anyone reading our report can use the level of significance he or she believes is appropriate: If theirs is less than 0.00164, they do not reject, while if theirs is equal to or greater than 0.00164, they do reject the hypothesis.

One disadvantage of the significance probability is that it can be very small (indicating, one would think, a "very significant" result) even when the true mean is close to 30. For example, $U = 3.15$ when $\overline{X} = 33.15$ and $\sigma/\sqrt{n} = 1.00$; $U = 3.15$ also when $\overline{X} = 30.0315$ and $\sigma/\sqrt{n} = 0.01$. In each case the significance probability is 0.00164 (i.e., 0.164 percent). In the first case, the confidence interval on the mean at 95 percent confidence runs over (31.19, 35.11), while in the second, the interval runs over (30.0119, 30.0511). In terms of *practical significance,* the latter is much more likely to be a trivial difference to the practitioner than is the former. However, there is no way to tell these two situations apart by using the significance probability. For this reason, it is recommended that instead confidence intervals be computed and presented.

ADDITIONAL STATISTICAL TOOLS

Today, a large number of statistical tools are used in quality control. The statistical tool kit (see Table 44.1) stresses the statistical base of collection, analysis, and interpretation of data. *Transformations,* discussed below, are a method often used to ensure that data will meet the assumptions of statistical procedures, while *Monte Carlo sampling methods* and *clustering and discrimination procedures* are powerful methods whose use in quality control is now growing. They allow analysis with minimal assumptions and analysis of multivariate characteristics, respectively. *Bootstrap methods* are a relatively recent attempt to simplify modeling and analysis; in their *generalized bootstrap* form they achieve this with minimal drawbacks. *Selection of the best* is an alternative goal (versus hypothesis testing or confidence intervals) that should be used when one's experiment has a goal of selection.

Transformations of Data. Most of the statistical methodology presented in this section assumes that the quality characteristic follows a known probability distribution. The analysis and conclusions that result are, of course, strictly valid only to the extent that the distribution assumption is correct. Under Tests of Hypotheses, a "goodness-of-fit" test was presented for quantitatively evaluating a set of data to judge the validity of a distributional assumption. Moderate deviations of a sample of observations from a theoretical population assumption are to be expected because of sampling variation. The goodness-of-fit test determines whether the deviation of the sample from a theoretical assumption is likely to have been due to sampling variation. If it turns out as unlikely, then it is concluded that the assumption is wrong.

Sometimes a set of data does not fit one of the standard distributions such as the normal distribution. One approach uses "distribution-free" statistical methods for further analysis. Some of these were listed under Tests of Hypotheses, and Natrella (1963) and Gibbons (1997) present further material. However, these methods often require larger sample sizes than conventional methods for equivalent statistical risks. Some other approaches to analysis are

1. Examine the data to see if there is a nonstatistical explanation for the unusual distributional pattern. For example, the output of each of several supposedly identical machines may be normally distributed. If the machines have different means or standard deviations, then the combined output probably has an unusual distribution pattern such as the mixture distribution already discussed in this section. In this case, separate analyses could be made for each machine.

2. Analyze the data in terms of averages instead of individual values. As stated under Basic Concepts, Types of Error, sample averages closely follow a normal probability distribution even if the population of individual values from which the sample averages came is not normally distributed. If it is sufficient to draw a final conclusion on a characteristic in terms of the average value, the normal distribution assumption can be applied. However, the conclusions apply only to the average value and not to the individual values in the population. (Predicting the percentage of a population falling outside engineering limits illustrates the situation where analysis in terms of the average would not be sufficient because engineering limits refer to individual values rather than averages.)

3. Use the Weibull probability distribution. The Weibull distribution is really a group of many continuous distributions with each distribution uniquely defined by numerical values of the parameters of the Weibull probability function (e.g., a beta value of 1.0 indicates an exponential distribution). If a set of data yields an approximate straight-line plot on Weibull paper, the straight line then directly provides estimates of the probabilities for the population. Whether the exact form of the probability distribution is normal, or exponential, or another distribution becomes somewhat secondary because the straight-line plot provides the needed probability estimates.

4. Make a transformation of the original characteristic to a new characteristic that is normally distributed. Figure 44.23 summarizes several of these mathematical transformations. These transformations are useful for (*a*) achieving normality of measured results, (*b*) satisfying the assumption of equal population variances required in certain tests, and (*c*) satisfying the assumption of additivity of effects required in certain tests. Natrella (1963) discusses transformations for all these uses. Romeu and Ozturk (1996) provide graphic tests of normality (even in multivariate cases).

The most common transformations for achieving normality are

$$\xi_1(X_1) = \sqrt{X_1 - a}$$

$$\xi_2(X_1) = X_1^{1/3}$$

$$\xi_3(X_1) = \log_{10}(X_1)$$

$$\xi_4(X_1) = \arcsin \sqrt{X_1}$$

$$\xi_5(X_1) = \sinh^{-1} \sqrt{X_1}$$

If one of these, say, $\xi(X_1)$, is normally distributed, the mean and variance of $Y_i = \xi(X_i)$ may be estimated by

$$\overline{Y} = \sum_{j=1}^{n} \frac{Y_j}{n} \qquad S_Y^2 = \sum_{j=1}^{n} \frac{(Y_j - \overline{Y})^2}{(n-1)}$$

However, interest in many cases is not in the expected value $E\xi(X_1)$ and the variance Var $\xi(X_1)$ but in the original problem units EX_1 and Var(X_1). Simply using the inverse transformation—for example, to estimate EX_1 by $\overline{Y}^2 + a$ in the case of ξ_1—results in a biased estimate. Good estimators for the mean of the X's are given in Table 44.35. Good estimators of the variance of the X's allow us to find approximate 95 percent confidence intervals for the mean of the X's; such estimates are given in Table 44.36. For example, when using $\sqrt{X_1 - a}$, a 95 percent confidence interval for the mean of the X's is

$$\overline{Y}^2 + a + \left(1 - \frac{1}{n}\right) s_Y^2 \pm 2\sqrt{\lambda}$$

where

$$\lambda = \frac{4}{n} s_Y^2 \overline{Y}^2 + s_Y^4 \left\{ \left(1 - \frac{1}{n}\right)^2 - \frac{n-1}{n+1} \left[1 - 2\left(1 - \frac{1}{n}\right)^2 + 3\left(1 - \frac{1}{n}\right)^4 \right] \right\}$$

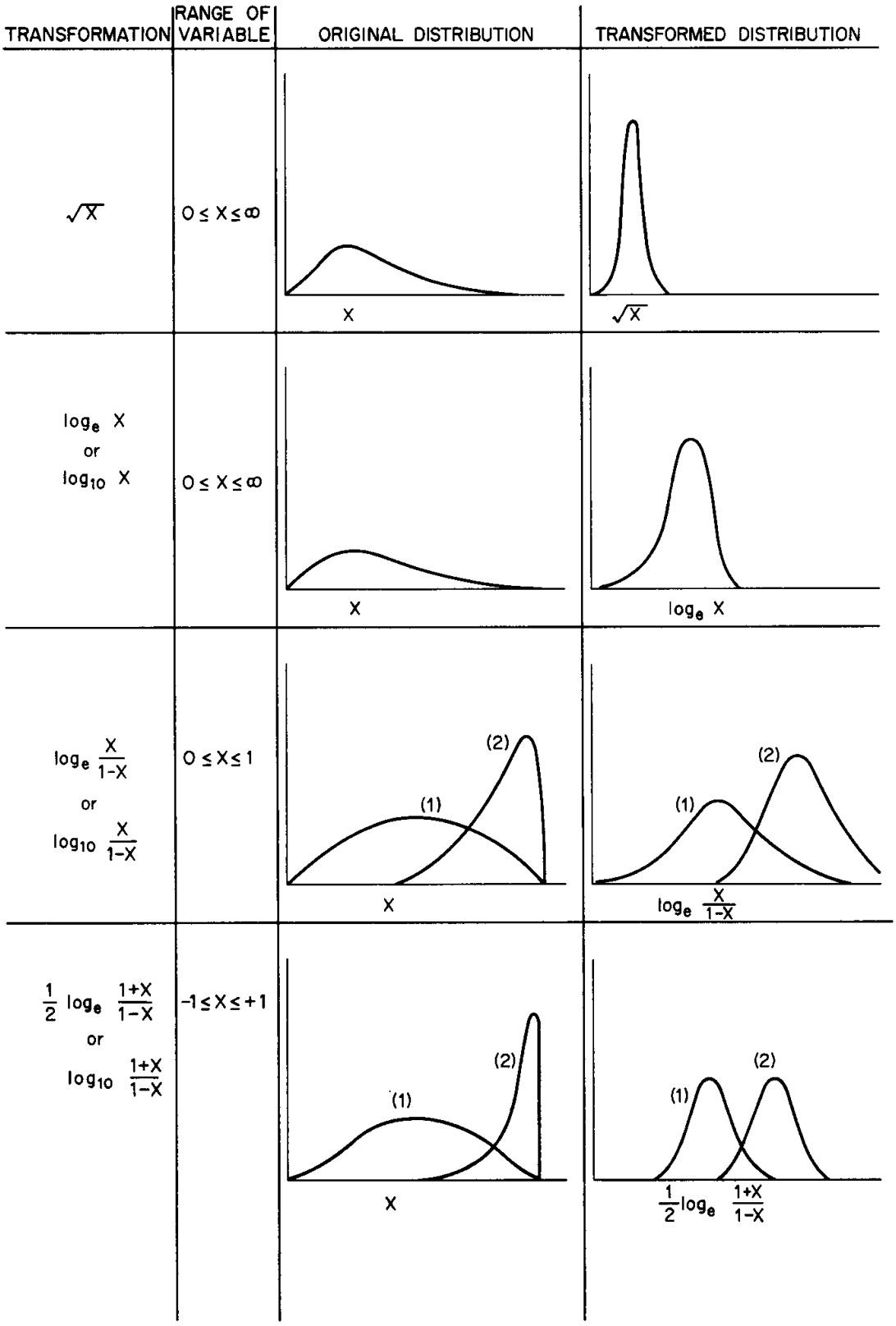


FIGURE 44.23 Summary of some transformations.

TABLE 44.35 Transformations and Estimators of $E(X_1)$†

$\xi(X_1)$	Estimator of $E(X_1)$
$\sqrt{X_1 - a}$	$\overline{Y}^2 + a + \left(1 - \frac{1}{n}\right) s_Y^2$

†For estimators in the cases $\log_{10}(X_1)$, arcsin $\sqrt{X_1}$, and $\sinh^{-1}\sqrt{X_1}$, see Dudewicz (1983).

TABLE 44.36 Estimators of Variances of Estimators of $E(X_1)$†

$\xi(X_1)$	Estimator of variance of estimator of $E(X_1)$
$\sqrt{X_1 - a}$	$\frac{4}{n} s_Y^2 \overline{Y}^2 + s_Y^4 \left\{ \left(1 - \frac{1}{n}\right)^2 - \frac{n-1}{n+1} \left[1 - 2\left(1 - \frac{1}{n}\right)^2 + 3\left(1 - \frac{1}{n}\right)^4 \right] \right\}$

†For estimators in the cases $\log_{10}(X_1)$, arcsin $\sqrt{X_1}$, and $\sinh^{-1}\sqrt{X_1}$, see Dudewicz (1983).

Full references are given by Dudewicz (1983), who also covers procedures for dealing with cases where variances are unequal and data are normal. It is standard to recommend that equality of variability should be investigated, even when data are normal. Procedures for dealing with variance inequality when it is found had not been available until recent years. Recently, Dudewicz and Dalal (1983) showed how to compare several new processes with a standard process in this setting; their paper also includes consideration of nonnormality and a data set (their Section 5) with numerical details and normal probability plots for an example arising with solvents.

Monte Carlo Sampling Methods. Monte Carlo sampling methods are finding increasing and important uses in quality control. For example, Gutt and Gruska (1977) use them to predict quality problems that may result from variation in manufacturing and assembly operations. This method is based on fitting distributions to data (such as the GLD distribution discussed above under Continuous Probability Distributions) and sampling from them using random numbers (also discussed previously), employing the resulting data to assess the performance of the simulated system (which allows optimization of the system before it is built or modified). Other uses occur in optimization (Golden et al. 1984), location modeling (Golden and Eiselt 1992), vehicle routing (Golden 1993), and inventory management (Dudewicz 1997). For example, a recent optimization method called ARSTI that uses random intervals (Edissonov 1994) compares very well with previous methods. Since one can choose to apply more than one method and then take the better result, one can only gain by incorporating such new methods into one's work [which is facilitated by the fact that a FORTRAN computer program for ARSTI is given by Edissonov (1994)].

Bootstrap Methods. In many quality control problems, full solutions have been developed under the assumption that one knows the underlying probability distribution. Since often one does not have this knowledge in practice, one needs to estimate the probability distribution (see Selecting a Discrete Distribution and Selecting a Continuous Distribution earlier in this section). In the *bootstrap method,* one does not estimate the probability distribution from a set of data like the tensile strengths of five batches of rubber made with recipe 1 used in the illustration of testing average strength: 3067, 2730, 2840, 2913, and 2789. That is, one does not test (for example) normality of the data and (if the test fails to reject) use procedures that assume a normal probability distribution. Instead, the bootstrap method takes samples at random with replacement from the data we have and uses them to try to answer the question of interest. If one is interested, for example, in a 90 percent

confidence interval for the mean tensile strength with recipe 1, then one proceeds as follows: Draw samples of size $n = 5$ repeatedly from the basic data points until one has N such samples (with $N = 500$ being a popular choice—see note 1 below); for each sample, calculate the sample mean, thus obtaining $N = 500$ sample means, say, $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_{500}$; delete the smallest 25 (5 percent) and largest 25 (5 percent) of these 500 sample means [since (.10)(500)/2 = 25]; and state that the mean tensile strength is between the smallest and largest of the remaining 450 sample means.

The bootstrap method is simple and attractive; it seems to yield (without any complicated statistics or mathematics) solutions to difficult problems, i.e., to give us "something for nothing." Since there is no free lunch, we should be suspicious. In fact, the method can behave badly if the sample is not large (and most of the theory is developed as the sample size becomes infinite). For example, suppose that one desires to study the maximum rainfall over 100 years and has data on 5 years. In resampling those 5 years, one will never observe a higher rainfall than the largest of the 5 measurements with the bootstrap method. Clearly, the study will be greatly misled by this (e.g., we may recommend an inadequately sized dam or levee). Thus while the bootstrap method can be useful in some settings, it is fraught with danger. A more robust version (which yields about the same results when the sample sizes are large but is not so fragile when the sample sizes are small) is the generalized bootstrap method discussed below.

The bootstrap type of method was used as early as 1967 but did not gain wide acceptance until it was given the name *bootstrap method* by B. Efron in 1979, after which it experienced an explosion of interest. It in some ways generalizes the method given by Quenouille in 1949, which gained wide acceptance when christened the *jackknife method* by J. Tukey in 1958. History, references, and examples of its flaws (and how to remedy them) are given by Dudewicz (1992) [also see Section 15.6 of Dudewicz and Mishra (1988)].

The Generalized Bootstrap. The generalized bootstrap was introduced by Dudewicz (1992) as a generalization of the bootstrap method with superior properties in the small-sample (few observations) case. [For a textbook discussion, see Section 6.6 of Karian and Dudewicz (1991).] Basically, with the generalized bootstrap, one takes the observations and fits an appropriate probability distribution from a broad class such as the extended generalized lambda distribution discussed earlier in this section. One's random samples are then taken from the fitted distribution. Thus, in the context of the five rubber batches discussed under Bootstrap Methods earlier, one fits an EGLD to the $n = 5$ data points. Then $N = 500$ random samples of size 5 are drawn from the fitted EGLD (not from the basic 5 data points), and analysis proceeds as in the bootstrap method. This method has been shown to do better when one has few data points (but to do as well when one has many) in recent studies [e.g., Sun and Dietland-Müller (1996) have an excellent exposition with real-data examples].

Bootstrap methods, especially in the form of the generalized bootstrap, would suggest that one fit a model to the data (e.g., a Poisson or other model) and then (assuming the fit passes testing; see Test of Model Validity above) use that model in one's analysis, which can proceed by bootstrapping. (The generalized bootstrap method allows for there being a possibility one could observe three or more blemishes sometime in the future, while the bootstrap method—which is not recommended—always assumes the probability of three or more is zero just because we did not observe three or more in any of the set of data we have.)

Note 1: The statistical literature recommends the use of at least $N = 200$ replications. Recent work by Sun and Müller-Schwarze (1996, pp. 482–483) suggests that considerable gains in accuracy can be had by requiring at least $N = 500$ replications. The number of replications made with one's data set should not be confused with one's actual data set's size (see Note 2).

Note 2: Bootstrap methods are widely used for sample sizes that are quite small [such as 9 in Sun and Müller-Schwarze (1996)] due to the need to draw reliable conclusions from small data sets (often gathered over a considerable period at considerable effort, such as 10 years for the Sun and Müller-Schwarze data). I would strongly recommend the generalized bootstrap for cases with fewer than 100 data points.

Clustering and Discrimination. *Clustering and discrimination* methods are a part of the area of statistics called *multivariate analysis* (Siotani et al. 1985). A typical type of problem where these methods are used in quality control is when several different kinds of malfunctions within a production facility cause product to fall outside engineering limits. It is often difficult to determine the causes of the malfunction in any one case. Then clustering a number of cases may reveal causal links via common factors over the clusters. (That is, this method allows one to ask, "What do the cases with malfunctions of each type have in common?")

As an example of the power of the "discrimination" method, Fisher (1936) gave the data excerpted in Table 44.37. [A convenient source of the full data set is Dixon (1983, p. 520).]This consists of two length and two width measurements on each of three distinct varieties that might be found in the same location. We wish to know: How well can the varieties (which can be classified by a more involved analysis without error) be classified by just use of the two length and two width measurements? After these data are entered into a computer, program 7M of the BMDP set of programs (Dixon 1983) may be used to answer this question. The program code is given in Figure 44.24. From the resulting output, of key interest are the so-called *canonical variables,* which are the linear combinations of L_1, L_2, W_1, and W_2 that best discriminate among the three groups. In this example, these turn out to be

$$V_1 = 2.10510 + 0.82938L_1 + 1.53447W_1 - 2.20121L_2 - 2.81046W_2$$

$$V_2 = -6.66147 + 0.02410L_1 + 2.16452W_1 - 0.93192L_2 + 2.83919W_2$$

A plot of the (V_1, V_2) values for the 150 data points is given in Figure 44.25, with $T = 1, 2, 3$ cases labeled *A, B, C,* respectively, and shows the excellent results obtained. These results are deemed excellent because they allow us to classify a future observation into the correct group with high probability of being correct. For example, if we find $V_1 = 7.20$ and $V_2 = 1.00$, we are virtually certain that group *A* is involved. (This plot is produced by program 7M.)

Heteroscedastic Discrimination. Traditional discrimination methods have assumed that one knows that the variances are equal and have not been able to specify that the misclassification proba-

TABLE 44.37 Two Length (L_1, L_2) and Two Width (W_1, W_2) Measurements on 150 Individuals (50 from Each of Types $T = 1, 2, 3$), in 0.01-cm Units*

L_1	W_1	L_2	W_2	T
50	33	14	02	1
64	28	56	22	3
65	28	46	15	2
67	31	56	24	3
63	28	51	15	3
46	34	14	03	1
69	31	51	23	3
62	22	45	15	2
59	32	48	18	2
46	36	10	02	1
61	30	46	14	2
60	27	51	16	2
65	30	52	20	3
56	25	39	11	2
65	30	55	18	3
⋮	⋮	⋮	⋮	⋮

*To save space, only the first 15 individuals' measurements are shown.

```
//   EXEC    BIMED, PROG =BMDP7M
/problem title is 'fisher data'.
/input variables are 5. format is '(4f3.1, f3.0)'.
/variable names are L1, w1, L2, w2, t.
grouping is t.
/group codes(5) are 1 to 3.
names(5) are set, ver, vir.
/end
```

FIGURE 44.24 BMDP 7M program for dataset analysis.

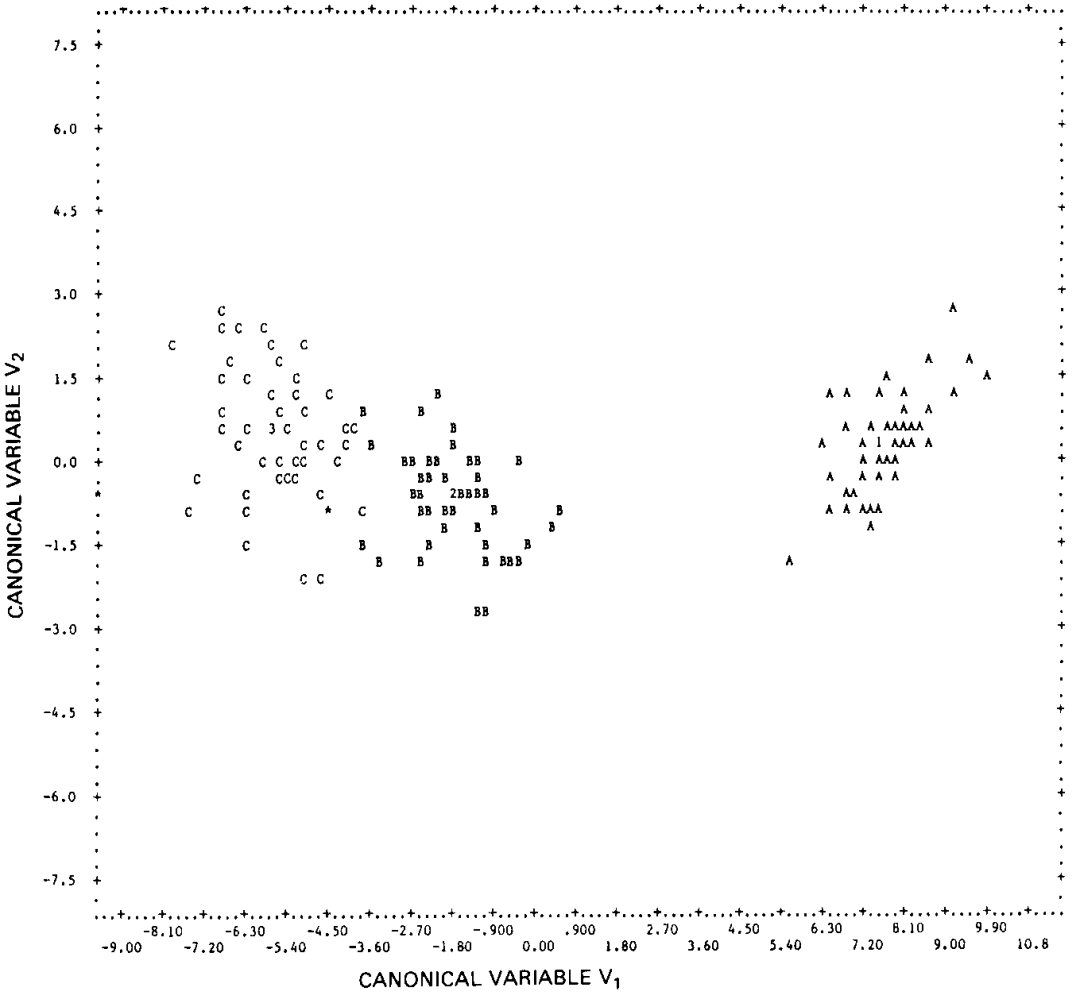


FIGURE 44.25 Plot showing groups (*A, B, C*) discriminated by "canonical variables" V_1 and V_2. Overlap of different groups is indicated by *.

bility will be estimated within .01 (say). However, owing to development of the heteroscedastic method [see Dudewicz (1995) for a review], this is now possible in some cases and expected to become more available in the next few years. In other cases, the goal may be to estimate the overlap (rather than directly discriminate the populations); for some methods here, see Mulekar and Mishra (1997).

Selection of the Best versus Testing Hypotheses. In hypothesis testing, one may (e.g., in the setting of Test 2 in Table 44.31) assess whether one can reject the assertion that the means of two populations are equal. While this is an appropriate question in some settings, there are

other settings where one knows a priori that one must select one or the other of the two populations and wants to select the one with the larger mean (called the "best" population). Procedures for this setting were first published in 1954 by R. E. Bechhofer, and since that time, over 500 publications and several books have been devoted to the problem. [For a categorized list and reviews, see Dudewicz and Koo (1982).]

Example of Selection of the Best: Suppose we are considering two recipes for rubber and want to select the one with the larger tensile strength. We make $n_0 = 5$ batches with each recipe and test for tensile strength with the following results:

Recipe 1	*Recipe 2*
3067	3200
2730	2777
2840	2623
2913	3044
2789	2834

We desire to be at least 95 percent sure that the one we select has a true mean tensile strength no further than $\delta^* = 120$ units from the best one. We then calculate the $s_1^2 = 16{,}923.7$ and $s_2^2 = 51{,}713.3$ in the first samples and take more observations (since we do not know the variances in advance of the first samples, we are not in a position to know how many total samples we will need). The total sample sizes are to be

$$n_1 = \max[n_0 + 1,(s_1 h/\delta^*)^2] = \max\{5 + 1,[(130.09)(3.11)/120]^2\}$$
$$= \max(6,11.4) = 12$$
$$n_2 = \max[n_0 + 1,(s_2 h/\delta^*)^2] = max\{5 + 1,[(227.41)(3.11)/120]^2\}$$
$$= \max(6,34.7) = 35$$

where $h = h_{n_0}(k, P^*) = h_5(2,0.95)$ comes from Table 44.30 (with $k = 2$ because we are seeking the best of two populations, $P^* = 0.95$ because we desire 95 percent certainty, and $n_0 = 5$ because the first samples were of five observations). We then take $n_1 - 5 = 12 - 5 = 7$ more observations from recipe 1 and $n_2 - 5 = 35 - 5 = 30$ more observations from recipe 2. We then compute the sample mean $\overline{X}_1$ of all 12 observations from recipe 1 and the sample mean $\overline{X}_2$ of all 35 observations from recipe 2, selecting the recipe that produces the larger of $\overline{X}_1$ and $\overline{X}_2$, asserting that that recipe has the larger mean (or a mean no further than 120 units of tensile strength from the best recipe). For example, if we find $\overline{X}_1 = 3067$ and $\overline{X}_2 = 2895$, we will select recipe 1. We will state that we are at least 95 percent sure that recipe 1 is either the best recipe or (in any case) has a mean tensile strength no further than 120 strength units from the best (if it is not the best).

Procedures are also available for selection of the best of more than two populations [see Section 6.3 of Karian and Dudewicz (1991)]; one can essentially use the same procedure with the appropriate k being used when h is looked up in Table 44.30.

REGRESSION AND CORRELATION ANALYSIS

Many quality control problems require estimation of the relationship between two or more variables. Often interest centers on finding an equation relating one particular variable to a set of one or more variables. For example, how does the life of a tool vary with cutting speed? Or how does the octane number of a gasoline vary with its percentage purity?

Regression analysis is a statistical technique for estimating the parameters of an equation relating a particular variable to a set of variables. (Some authors refer to this as *least squares* or *curve fitting.*) The resulting equation is called a *regression equation.*

Some experimental data for the tool life example are given in Table 44.38 [from Johnson and Leone (1964, p. 380)] and plotted in Figure 44.26. Tool life is the *response variable* (also called the *dependent variable* or the *predictand*), and cutting speed is the *independent variable* (also called the *predictor variable*). In this case, the independent variable is controllable; i.e., it is fixed by the experimenter or the operator of the machine. In the second example, both the octane number and the percentage purity are random. The data for this example (from Volk 1956) are given in Table 44.39 and plotted in Figure 44.27. Since the goal is to predict the octane number, it is regarded as the dependent variable, and the percentage purity is considered as the independent variable. (In many problems there are a number of independent variables, and in some cases this set of independent variables includes both random and controllable variables.)

The computations for two-variable regression problems can be done quite easily on a calculator, but when there are many variables, the number of computations becomes overwhelming. (Even in the two-variable case, many of the computer programs available fail to provide numerically accurate calculations. It is strongly recommended that one *not* write one's own regression program and that only major tested software packages such as SAS, BMDP, and the like be utilized.) With modern digital computer multiple regression programs, the number of variables is not a restriction. To under-

TABLE 44.38 Tool Life (Y in Minutes) versus Cutting Speed (X in Feet per Minute)

Y	X
41	90
43	90
35	90
32	90
22	100
35	100
29	100
18	100
21	105
13	105
18	105
20	105
15	110
11	110
6	110
10	110

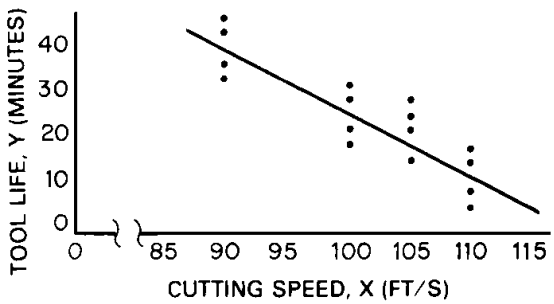


FIGURE 44.26 Tool life Y versus cutting speed X.

TABLE 44.39 Octane Number Y versus Percentage Purity X

Y	X
88.6	99.8
86.4	99.7
87.2	99.6
88.4	99.5
87.2	99.4
86.8	99.3
86.1	99.2
87.3	99.1
86.4	99.0
86.6	98.9
87.1	98.8

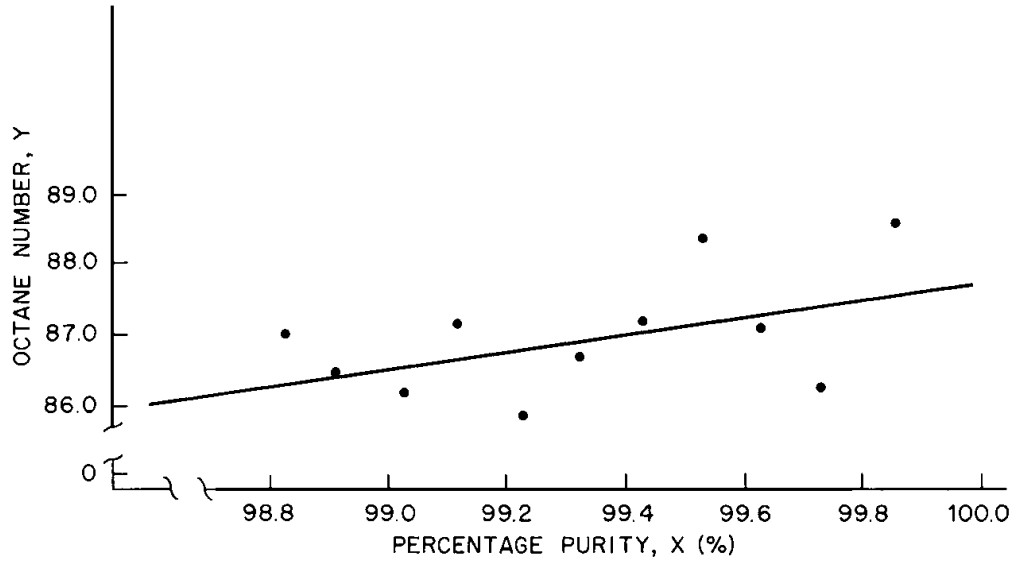


FIGURE 44.27 Octane number versus percentage purity.

stand and interpret the results of multidimensional problems, a thorough knowledge of the two- and three-variable cases is necessary.

There are many reasons for constructing regression equations. Although the motives do not affect the calculations, they do affect the interpretation of the results. In some cases, regression analysis is used to describe the nature of a relationship in a quantitative manner. Often the *goals* are more specific. In the first example, where the cutting speed is controllable, the objective might be to find the particular value of cutting speed which minimizes tool wear or some cost function based on tool wear. Least squares regression also can be used to determine the important independent variables in a process, e.g., whether process variables such as moisture, pressure, or temperature affect a quality characteristic of the product such as strength.

In other problems, where the independent variable is not controllable, the goal may be to predict the value of the dependent or response variable. This might be done because the independent variable is easier to measure than the dependent variable. Or the independent variable may be available before the dependent variable, and hence it would be desirable to forecast the value of the dependent variable before it occurs. In still other cases it might allow a destructive test to be replaced by a nondestructive test.

The following list includes a number of *uses of regression equations:*

1. Forecasting and prediction
2. Quantitatively describing the relationship between a particular variable and another set of variables
3. Interpolating between values of a function
4. Determining the important independent variables
5. Locating the optimum operating conditions
6. Discriminating between alternative models
7. Estimating particular regression coefficients

For any of these goals stated, the *basic steps in a regression study* are those of the checklist for planned experimentation given at the beginning of this section. A summary, specifically relating the steps to regression, is

1. Obtain a clear statement of the objectives of the study. Determine which variable is to be the response variable and which variables can be included as independent variables. In addition, obtain some measure of the precision of the results required—not necessarily in statistical terminology. (It is important to have a thorough understanding of what use will be made of the regression equation, since this may preclude the use of certain variables in the equation and will also help to give an understanding as to how much effort and money should be devoted to the project.)
2, 3. Specify collection procedures for the data. Collect the data. (The end results can only be as good as the data on which they are based. Careful planning at this stage is of considerable importance and can also simplify the analysis of the data.)
4. Prepare crossplots (plots of one variable versus another) of the data to obtain information about the relationships between the variables; screen the data; calculate the regression equation; and evaluate how well it fits the data (including looking at transformations of variables for a better fit, or the removal of variables from an equation if they do not improve the prediction). Give measures of the precision of the equation and any procedure for using the equation. Also specify procedures for updating the equation and checks to determine whether it is still applicable, including control charts for the residuals (observed value-predicted value). (Section 24, Statistical Process Control, discusses control charts.)
5, 6, 7. As in the checklist for planned experimentation.

A number of texts have been written on regression, including Daniel and Wood (1971) and Draper and Smith (1981). These include computer programs and output. Dudewicz and Karian (1985) also discuss design questions in detail. In addition to regression, other techniques have been devised for the analysis of multivariate data; see Kramer and Jensen (1969, 1970) and Siotani et al. (1985) for details. (One of these techniques is discriminant and cluster analysis, already discussed in this section.)

Our discussion of regression begins with a single predictor problem and then proceeds to problems with more than one predictor variable and a discussion of computer programs and outputs, with their interpretation. While many texts emphasize advanced mathematical aspects of regression, this is not needed for a practical understanding now that high-quality software is available; hence we find no need for such mathematics. This makes this important subject accessible to most quality practitioners.

Simple Linear Regression. Many problems involve only a single predictor variable X. (The dependent variable Y is often related to other predictor variables, which have either been held constant during the experiment or their effects judged to be much smaller than that of X.) These problems are often referred to as ones of *simple linear regression.*

Graphing the Data. A first step in any study of relationships between variables is to plot a graph of the data (often called a *scatter diagram*). The convention is to plot the response variable on the

vertical axis and the independent variable on the horizontal axis. A graph can provide a great deal of information concerning the relationship between variables and often suggests possible models for the data. The data plotted in Figure 44.26 suggest that Y is linearly related to X over the range of this experiment. (If this were not the case, various transformations of the data as well as curvilinear relationships could also be considered. Often the relationship can be "linearized" by taking the logarithm of one or both of the variables.)

A graph also can indicate whether any of the observations are outliers, i.e., observations that deviate substantially from the rest of the data. (Outliers may be due to measurement errors or recording errors, in which case they should be corrected or deleted. They may be due to process changes or other causes, and the investigation of these changes or causes may provide more information than the analysis of the rest of the data.) No outliers are apparent in Figure 44.26.

A closer inspection of the graph can give an indication of the variability of Y for fixed X. In addition, it may show that this variability remains constant over all X or that it changes with X. In the latter case the method of weighted least squares (see Draper and Smith 1981) may be preferred to the standard least squares technique discussed here.

The Model. After graphing the data, we want to obtain an equation relating Y to X. To do this, a model for the data must be postulated. (I emphasize that the proposed model may be modified during the course of the study and is just a starting point.)

A possible model for the data in Figure 44.26 is

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where β_0 and β_1 are the unknown intercept and slope, respectively, of the regression line. The model assumes that Y is a linear function of X plus a random error term, denoted by ϵ. This random error may be due to errors in the measurement of Y and/or to the effects of variables not included in the model, which is called *equation error.* The X's are assumed to be measured with negligible error. For the data in Figure 44.26, the X's are fixed; however, the same model can be used when the X's are random as in Figure 44.27.

Estimating the Prediction Equation. The objective is to find estimates (b_0, b_1) of the unknown parameters (β_0,β_1) and thus obtain a prediction equation

$$\hat{Y} = b_0 + b_1 X$$

where $\hat{Y}$ is the predicted value of Y for a given value of X.

Least squares provides a method for finding estimates of these parameters from a set of N observations $(Y_1,X_1),\ldots,(Y_N,X_N)$. The estimates are called *least squares estimates* because they minimize the sum of the squared deviations between the observed and predicted values of the response variable $\sum(Y_m - \hat{Y}_m)^2 = \sum(Y_m - b_0 - b_1X_m)^2$. These ideas are illustrated in Figure 44.28. (For a mathematical derivation of the estimates, see any of the texts mentioned in the introduction.)

If (1) the observations are independent, (2) the variance of the errors is constant over these observations, and (3) the linear model postulated is correct, the least squares estimates are the "best linear unbiased estimates": In the class of linear unbiased estimates of the parameters, the least squares estimates have the smallest variance. (Even if these conditions are not satisfied, the least squares technique can be used, although modifications or other methods may provide better estimates.) Note that no assumption has been made concerning the distribution of the random error, and in particular a normal distribution is not assumed. No assumption on this error term will be required until confidence intervals and tests of hypotheses are constructed.

The least squares estimates for the parameters of the linear model $Y = \beta_0 + \beta_1 X + \epsilon$ are

$$b_1 = \frac{\sum(X_m - \overline{X})(Y_m - \overline{Y})}{\sum(X_m - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1\overline{X}$$

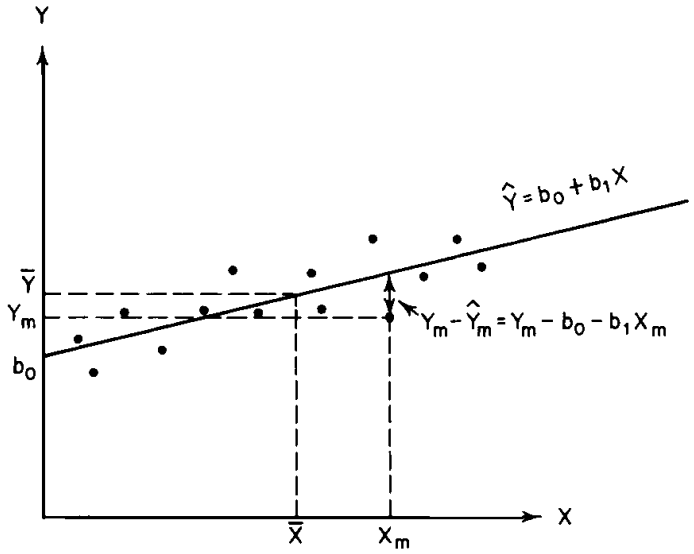


FIGURE 44.28 Least squares.

where $\bar{X} = \Sigma X_m/N$ and $\bar{Y} = \Sigma Y_m/N$ are sample averages. All these summations range from $m = 1$ to $m = N$. (Except where needed, the additional notation will be omitted for typographic simplicity.)

As can be seen, b_1 is related to the sample correlation coefficient

$$r = \frac{\Sigma(X_m - \bar{X})(Y_m - \bar{Y})}{\sqrt{\Sigma(X_m - \bar{X})^2 \Sigma(Y_m - \bar{Y})^2}}$$

by

$$b_1 = r \sqrt{\frac{\Sigma(Y_m - \bar{Y})^2}{\Sigma(X_m - \bar{X})^2}}$$

However, the concept of correlation is meaningful only when both the variables are random, whereas b_1, the least squares estimate of the rate of change of Y per unit change in X, has meaning for both the case of random X and controllable or fixed X.

The sums, sums of squares, and sum of crossproducts for the data ($N = 16$) given in Table 44.38 are

$$\Sigma X_m = 90 + 90 + \ldots + 110 = 1620$$

$$\Sigma Y_m = 41 + 43 + \ldots + 10 = 369$$

$$\Sigma X_m^2 = 8100 + 8100 + \ldots + 12{,}000 = 164{,}900$$

$$\Sigma Y_m^2 = 1681 + 1849 + \ldots + 100 = 10{,}469$$

$$\Sigma X_m Y_m = 3690 + 3870 + \ldots + 1100 = 36{,}170$$

(Note that calculations for regression are very susceptible to both human and numerical error because of their complexity. Hence good software should be used. The numerical examples in this section can be used both to test one's understanding of that software and to test its accuracy.)

The summary statistics are computed using the following computational formulas:

$$\bar{X} = \Sigma X_m/N = 101.25 \qquad \bar{Y} = \Sigma Y_m/N = 23.06$$

$$\sum(X_m - \overline{X})^2 = \sum X_m^2 - \frac{(\sum X_m)^2}{N} = 164{,}900 - \frac{(1{,}620)^2}{16} = 875.00$$

$$\sum(Y_m - \overline{Y})^2 = \sum Y_m^2 - \frac{(\sum Y_m)^2}{N} = 10{,}469 - \frac{(369)^2}{16} = 1958.94$$

$$\sum(X_m - \overline{X})(Y_m - \overline{Y}) = \sum X_m Y_m - \frac{\sum X_m \sum Y_m}{N}$$

$$= 36{,}170 - \frac{(1620)(369)}{16}$$

$$= -1191.25$$

From these results the least squares estimates can be calculated as

$$b_1 = \frac{-1191.25}{875} = -1.3614$$

$$b_0 = 23.06 - (-1.3614)(101.25) = 160.9018$$

and hence the prediction equation is

$$\hat{Y} = 160.90 - 1.3614X$$

The prediction equation is sometimes written in terms of deviations from averages, i.e., $\hat{Y} = \overline{Y} + b_1 (X - \overline{X})$, which for this example becomes

$$\hat{Y} = 23.06 - 1.3614(X - 101.25)$$

Examining the Prediction Equation. After estimating the coefficients of the prediction equation, the equation should be plotted over the data to check for gross calculation errors. Roughly half the data points should be above the line and half below it. In addition, the equation should pass exactly through the point $(\overline{X}, \overline{Y})$.

A number of criteria exist for judging the adequacy of the prediction equation. One common measure of the adequacy of the prediction equation is the proportion of variation R^2 explained. To compute R^2, the sum of the squared deviations of the Y_m about $\overline{Y}$ is partitioned into two parts, the sum of squares due to regression and the residual sum of squares:

$$\begin{aligned}\sum(Y_m - \overline{Y})^2 &= \text{SS(REG)} + \text{SS(RES)} \\ &= \sum(\hat{Y}_m - \overline{Y})^2 + \sum(Y_m - \hat{Y}_m)^2 \\ &= b_1 \sum(X_m - \overline{X})(Y_m - \overline{Y}) + \sum(Y_m - \hat{Y}_m)^2\end{aligned}$$

From this, the proportion of the variation $\sum(Y_m - \overline{Y})^2$ explained by the regression is computed as

$$R^2 = \frac{\text{SS(REG)}}{\sum(Y_m - \overline{Y})^2}$$

$$= \frac{b_1 \sum(X_m - \overline{X})(Y_m - \overline{Y})}{\sum(Y_m - \overline{Y})^2}$$

$$= \frac{(-1.3614)(-1191.25)}{1958.94} = 0.828$$

Thus in this example the prediction equation explains 82.8 percent of the variation of the tool life.

Another interpretation of R^2 (when both the independent and the dependent variables are random) is as the square of the sample multiple correlation coefficient. When there is only one independent variable, this reduces to the square of the sample correlation coefficient r defined earlier.

Although R^2 is a useful measure of the adequacy of the prediction equation, an estimate of the variability of the Y's about the regression equation is usually more important. Either the sample variance s_e^2 or its square root s_e, called the *standard error of the estimate,* can be used. The latter is often preferred because it is measured in the same units as Y. Both of these, as well as other results, can be obtained from the analysis of variance (ANOVA) given in Table 44.40.

The corrected total sum of squares and the regression sum of squares are calculated from the summary statistics and the estimate of the regression coefficient [some authors include the total sum of squares, uncorrected, in the ANOVA table, partitioning it into two parts—the corrected sum of squares and the sum of squares due to $\overline{Y}$ (or b_0) (see Draper and Smith 1981)]. Although the residual sum of squares can be calculated directly, it is more easily obtained as the difference between the corrected total sum of squares and the sum of squares due to regression. Each of these sums of squares has an associated degrees of freedom (see Testing a Hypothesis When the Sample Size Is Fixed in Advance). The corrected total sum of squares has $N - 1$ degrees of freedom, since one degree of freedom is used in estimating the mean. For this *one*-variable model, there is *one* degree of freedom associated with the regression sum of squares, leaving $(N - 1) - 1 = N - 2$ degrees of freedom associated with the residual sum of squares. The mean squares (MS) are calculated by dividing the sum of squares by their associated degrees of freedom. The estimate of the variance of Y about the regression line is $s_e^2 = \text{MS(RES)}$; hence the standard error of the estimate is $s_e = \sqrt{\text{MS(RES)}}$.

From the mean squares an F statistic can be calculated as

$$F_{\text{CALC}} = \frac{\text{MS(REG)}}{\text{MS(RES)}}$$

If (1) the ϵ's in the original model are normally distributed with a common variance, (2) the observations are independent, and (3) the postulated linear model is correct, then the regression can be tested for significance, i.e., the statistical hypothesis

$$H_0: \quad \beta_1 = 0$$

can be tested against the alternative hypothesis

$$H_1: \quad \beta_1 \neq 0$$

by comparing F_{CALC} with the tabulated F_{TAB} at an appropriate level of significance α. If $F_{\text{CALC}} > F_{\text{TAB}}$, we conclude that the regression is significant and that the prediction equation is a better predictor of Y than $\overline{Y}$. Although it is difficult to check the assumptions stated above, the test is not extremely sen-

TABLE 44.40 ANOVA Table (Linear Model)

Source	Sum of squares	Degrees of freedom	Mean square
1. Due to regression (b_1)	$b_1\Sigma(X_m - \overline{X})(Y_m - \overline{Y})$	1	MS (REG) = SS(REG)/1
2. Residual	$\Sigma(Y_m - \hat{Y}_m)^2$†	$N - 2$	MS (RES) = SS(RES)/($N - 2$)
3. Total corrected for the mean	$\Sigma(Y_m - \overline{Y})^2$	$N - 1$	

†Obtained by subtracting (1) from (3).

sitive to departures in the distribution of ϵ from normality if the number of observations is relatively large. If the X's are random, this test must be interpreted in a conditional sense, i.e., given the values of the X's.

For the example, the analysis of variance (ANOVA) table is given in Table 44.41. (See Section 26, under Completely Randomized Design: One Factor, k Levels.) The regression is significant at an $\alpha = 0.01$ level ($F_{TAB} = 8.86$) and $s_e = \sqrt{24.08} = 4.91$.

It is important to note that even when the regression is significant, the unexplained variability can still be large, and the prediction equation may not be of any value.

Residuals, Outliers, Confidence and Prediction Bands, Extrapolation; Lack of Fit—Replicated Observations. If it is feasible to replicate, i.e., take more than one observation of Y at one or more values of X, the adequacy of the model also can be tested. (An estimate of the pure error sometimes may be available from sources outside the immediate experiment.) In this case, the SS(RES) can be partitioned into two parts—that due to pure error, SS(PE), and that due to lack of fit, SS(LF).

Suppose that there are N_m readings $Y_{m1}, Y_{m2},\ldots, Y_{mN_m}$ at x_m, where $m = 1, 2,\ldots, k$. The contribution to the sum of squares due to pure error for X_m is

$$\sum_{j=1}^{Nm} (\hat{Y}_{mj} - \overline{Y}_m)^2 = \sum_{j=1}^{Nm} Y_{mj}^{\ 2} - \frac{\left(\sum_{j=1}^{Nm} Y_{mj}\right)^2}{N_m}$$

and the associated degrees of freedom is $N_m - 1$. The SS(PE) is just the sum of these k contributions, and the associated degrees of freedom (DF) is

$$\sum_{m=1}^{k} (N_m - 1) = \sum_{m=1}^{k} N_m - k$$

For the example given in Table 44.38,

X_m		SS(PE)	DF
90	$41^2 + 43^2 + 35^2 + 32^2 - (151)^2/4 =$	78.75	3
100	$22^2 + 35^2 + 29^2 + 18^2 - (104)^2/4 =$	170.00	3
105	$21^2 + 13^2 + 18^2 + 20^2 - (72)^2/4 =$	38.00	3
110	$15^2 + 16^2 + 6^2 + 10^2 - (42)^2/4 =$	41.00	3
	Total	327.75	12

The SS(LF) is found by subtraction as

$$\text{SS(LF)} = \text{SS(RES)} - \text{SS(PE)} = 337.14 - 327.75 = 9.39$$

TABLE 44.41 ANOVA Example

Source	Sum of squares	Degrees of freedom	Mean square	F_{CALC}
1. Due to regression	1621.80	1	1621.80	67.35
2. Residual	337.14	14	24.08	
3. Total corrected for the mean	1958.94	15		

and the lack of fit degrees of freedom is obtained in a similar manner as $14 - 12 = 2$. The mean squares are then found by dividing the sum of squares by the appropriate degrees of freedom and

$$F_{\text{CALC}} = \frac{\text{MS(LF)}}{\text{MS(PE)}}$$

If the F_{CALC} is greater than the tabled F, the lack of fit is "significant," and a better or more complete model is needed (e.g., $Y = \beta_0 + \beta_1 X + \beta_2 X^2$). Plots of the residuals $(Y_m - \hat{Y}_m)$ versus X_m are particularly helpful in suggesting alternative models. Some examples are given in Figure 44.29. [Daniel and Wood (1971, pp. 19–24) present graphs of a number of nonlinear functions and give transformations that "linearize" them.] In each case the model $Y = \beta_0 + \beta_1 X$ was postulated and the plots are of the resulting residuals.

If F_{CALC} is less than the tabled F, the model is accepted. This does not mean that other variables should not be considered in the model, but only that the form of X in the model is adequate.

The calculations for our example are summarized in Table 44.42. The lack of fit is judged not significant at an α level of 0.05 ($F_{\text{TAB}} = 3.89$). Hence the postulated model is accepted, and the residual mean square is used as the estimate of the variance.

If replication is not possible, e.g., X is random rather than controllable, the Y values corresponding to X values that are close together can be used to obtain an estimate of the variability and hence judge the lack of fit. [See pp. 123–125 of Daniel and Wood (1971).]

Confidence Intervals. Both R^2 and s_e^2 provide measures of the reliability or adequacy of a prediction equation. Confidence intervals provide another measure of the reliability of the various

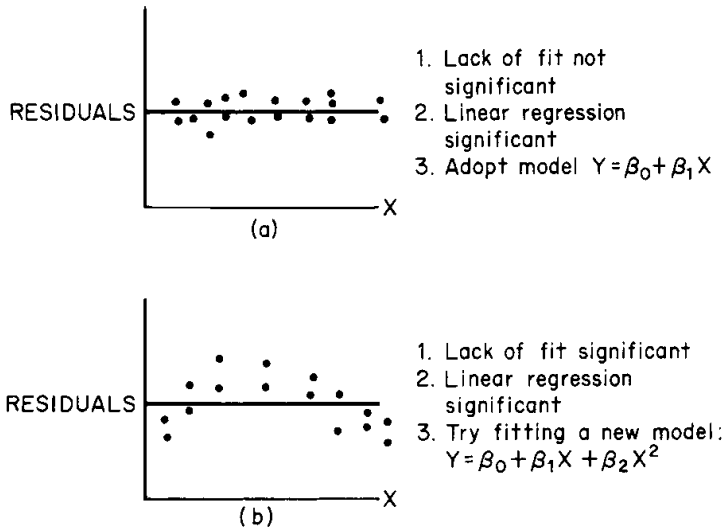


FIGURE 44.29 Residuals and lack of fit.

TABLE 44.42 ANOVA Lack of Fit Example

Source	Sum of squares	Degrees of freedom	Mean square	F_{CALC}
Lack of fit	9.39*	2	4.695	0.172
Pure error	327.75	12	27.3125	

*Obtained by subtracting SS(PE) and SS(RES).

estimates. All these confidence intervals are based on the square root of the residual mean square. A $(1 - \alpha)$ two-sided confidence interval for the slope β_1 is given by

$$b_1 \pm \frac{ts_e}{\sqrt{\sum(X_m - \overline{X})^2}}$$

where the value t is obtained from Appendix II, Table G, with $N-2$ degrees of freedom. The term in the denominator plays the role that $n^{1/2}$ plays in confidence intervals on population means. For the example, the 0.95 confidence interval on β_1 is

$$-1.36 \pm \frac{(2.145)(4.91)}{\sqrt{875}} = -1.36 \pm 0.356$$

The term $s_e/\sqrt{\sum(X_m - \overline{X})^2}$ is often called the *standard error of the regression coefficient.*

In addition to the confidence interval on β_1, more importantly, *confidence intervals* also can be constructed *for the mean of Y at a given value of X.* The $(1 - \alpha)$ confidence interval on the mean of Y at X (or equivalently on $\beta_0 + \beta_1 X$) is

$$b_0 + b_1 X \pm ts_e \sqrt{\frac{1}{N} + \frac{(X - \overline{X})^2}{\sum(X_m - \overline{X})^2}}$$

where X is the value at which the confidence interval is being constructed and t again has $N - 2$ degrees of freedom. (By letting $X=0$, a confidence level for β_0 is obtained.)

In addition to the assumptions previously stated, these confidence intervals also require (1) that the independent variable is fixed rather than random and (2) that the errors are normally distributed. However, if the X's are random, confidence intervals can still be calculated, but they must be interpreted in a conditional sense. Confidence intervals are not sensitive to departures from normality if the sample size is reasonably large. This is not the case for the following interval, which is very sensitive to the normality assumption. *Least absolute value* (LAV) and *Chebyshey estimation* are two possible alternatives to least squares estimation, which are less sensitive to model departures than is least squares. For sources of efficient computer algorithms, with a detailed numerical example, see Dielman and Pfaffenberger (1984).

In addition to a confidence interval on the expected value of Y at a given X, there may be a need for *an interval estimate for a future individual observation on Y at X.* [A more complete discussion of confidence intervals is given in Draper and Smith (1981). See Daniel and Wood (1971) for a confidence interval that simultaneously includes the whole line. See Dudewicz (1976, p. 427) for a plot of the interval for all X, called a *prediction band,* and its uses.] In this case the interval also must take into account the variability of Y about $\beta_0+\beta_1 X$, and the result is

$$b_0 + b_1 X \pm ts_e \sqrt{1 + \frac{1}{N} + \frac{(X - \overline{X})^2}{\sum(X_m - \overline{X})^2}}$$

where t has $N - 2$ degrees of freedom. Computations of these intervals for various values of X are given in Table 44.43.

Multiple Regression. Although there are many problems involving single predictor variables, more often there are many predictor variables. A generalization of the least squares technique, previously discussed, can be used to estimate the coefficients of the multivariable prediction equation. This problem is called *multiple regression.*

The General Model. For a problem with k predictor variables, the model can be written as

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon$$

TABLE 44.43 Computation of Confidence Limits

X (1)	$b_0 + b_1X$ (2)	$\frac{(X - \overline{X})^2}{\Sigma(\overline{X}_m - \overline{X})^2}$ (3)	$\sqrt{\frac{1}{N} + (3)}$ (4)	$ts \times (4)$ (5)	90% confidence limits on the mean of Y at X		$\sqrt{1 + \frac{1}{N} + (3)}$ (8)	$ts \times (8)$ (9)	90% prediction limits on Y at X	
					Lower (2) − (5) (6)	Upper (2) + (5) (7)			Lower (2) − (9) (10)	Upper (2) + (9) (11)
90	38.37	0.14464	0.4551	3.94	34.43	42.31	1.0987	9.50	28.87	47.87
100	24.76	0.00179	0.2535	2.20	22.56	26.96	1.0316	8.92	15.84	33.68
($\overline{X}$) 101.5	($\overline{Y}$) 23.06	0.0	0.2550	2.17	20.89	25.23	1.0308	8.91	14.15	31.97
110	11.15	0.08750	0.3870	3.36	7.79	14.51	1.0724	9.27	1.88	20.42

Note: Numbers in parentheses are column numbers. $t_{14,0.95} = 1.761$.

where the β's are unknown parameters and ϵ is the random error. These variables may be transformations of the original data. For example, in predicting gasoline yields from data on the specific gravity and vapor pressure of crude oil, Y may be the log of the gasoline yield, X_1 the crude oil specific gravity, X_2 the crude oil vapor pressure, and X_3 the product of the crude oil specific gravity with its vapor pressure.

The general model includes polynomial models in one or more variables such as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \epsilon$$

[which is called the *full quadratic model* of Y on X_1 and X_2 and is of great use in designed experiments (see Dudewicz and Karian 1985)]. This is still a linear model, since the term *linear model* means that the model is linear in the β's. [See Draper and Smith (1981) for a discussion of models which are nonlinear.]

Estimating the Prediction Equation. The objective now is to find the least squares estimates (b_0, b_1,…, b_k) of the unknown parameters (β_0, β_1,…, β_k) and obtain a prediction equation

$$\hat{Y} = b_0 + b_1 X_1 + \ldots + b_k X_k$$

where $\hat{Y}$ is the predicted value of Y for the given values of X_1,…, X_k. Letting $x_i = X_i - \overline{X}_i$ and using the fact that

$$b_0 = \overline{Y} - b_1 \overline{X}_1 - \ldots - b_k \overline{X}_k$$

this prediction equation can be expressed in the alternative form

$$\hat{Y} = \overline{Y} + b_1 x_1 + \ldots + b_k x_k$$

To simplify the formulas, the observations also can be expressed as deviations from their sample averages; i.e., for the mth observation $x_{im} = X_{im} - \overline{X}_i$ and $y_m = Y_m - \overline{Y}$. Then the least squares estimates of the $k + 1$ parameters of the multivariable linear model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$ can be obtained by solving the set of $k + 1$ linear equations:

$$\begin{aligned}
b_1 \Sigma x_1^2 m + b_2 \Sigma x_{1m} x_{2m} + \ldots + b_k \Sigma x_{1m} x_{km} &= \Sigma x_{1m} y_m \\
b_1 \Sigma x_{1m} x_{2m} + b_2 \Sigma x_{2m}^2 + \ldots + b_k \Sigma x_{2m} x_{km} &= \Sigma x_{2m} y_m \\
\vdots \qquad\qquad\qquad\qquad\qquad & \qquad \vdots \\
b_1 \Sigma x_{1m} x_{km} + b_2 \Sigma x_{2m} x_{km} + \ldots + b_k \Sigma x_{km}^2 &= \Sigma x_{km} y_m \\
b_0 = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2 - \ldots - b_k \overline{X}_k &
\end{aligned}$$

(All the above summations are on m and range from 1 to N.)

Solving these "reduced normal" equations simultaneously can be tedious, is error prone, and involves matrix algebra. Many reference works emphasize how to perform these calculations accurately. To the user of modern accurate statistical software, these calculations are of no direct importance: That user can trust that they are being done accurately and concentrate on statistical aspects of model adequacy, interpretation, and use.

Examining the Prediction Equation. After obtaining (b_0, b_1,…, b_k), an ANOVA table can be constructed and the adequacy of the prediction equation evaluated by a number of criteria. The ANOVA table, which is a generalization of that derived for the single predictor variable, is given in Table 44.44. The third row in Table 44.44 is the same as in Table 44.40. Note that the expressions in the first row reduce to those in the first row of Table 44.40 when k=1.

Since there are k variables in the model, the sum of squares due to regression has k degrees of freedom associated with it. In addition, since k coefficients and one intercept have been estimated, the residual sum of squares has $N - (k + 1) = N - k - 1$ degrees of freedom associated with it. The F

TABLE 44.44 ANOVA Table (Linear Model)

Source	Sum of squares	Degrees of freedom	Mean square
1. Due to regression	$b'a$†	k	MS(REG) = SS(REG)/k
2. Residual	$\Sigma(\overline{Y}_m - \hat{Y}_m)^2$‡	$N - k - 1$	MS(RES) = SS(RES)/$(N - k - 1)$
3. Total corrected for the mean	$\Sigma(Y_m - \overline{Y}_m)^2$	$N - 1$	

†The sum of squares due to regression can be written as $b'a = b_1\Sigma x_{1m}y_m + b_2\Sigma x_{2m}y_m + \cdots + b_k\Sigma x_{km}y_m$. [Note that b' is the transpose of the b vector, i.e., $b' = (b_1\ b_2 \cdots b_k)$. Those not familiar with vector notation can regard $b'a$ as a shorthand for the sum given in this footnote. Since computations are usually done with computer software, details of vector notation are not given and will not be needed by most readers.]

‡Obtained by subtracting (1) from (3).

statistic is calculated as before—i.e., F = MS(REG)/MS(RES)—and $s_e = \sqrt{\text{SS(RES)}/(N-\text{k}-1)}$. In this case, the F statistic can be used to test the statistical hypothesis

$$\text{H}_0: \quad \beta_i = 0 \qquad (i = 1, 2, \ldots, k)$$

against the alternative statistical hypothesis

$$\text{H}_1: \quad \text{Some } \beta_i \neq 0 \qquad (i = 1, 2, \ldots, \text{k})$$

The only change from simple linear regression is in the degrees of freedom used to look up the F_{TAB} (k, $n - k - 1$ versus 1, $n - 2$).

The proportion of variation explained by the equation (R^2) can be obtained from the ANOVA table. Note that R^2 does not depend on the number of variables in the equation, but s_e^2 does. In fact, if a new variable is added to the model (and the least squares estimates and ANOVA table are recomputed), the value of R^2 *cannot decrease.* However, s_e^2 can either increase or decrease, since it depends on the residual degrees of freedom in addition to the residual sum of squares, which decreases by one when a new variable is added.

Confidence Intervals. Confidence intervals for individual β's can be developed as

$$b_i \pm ts_e\sqrt{c_{ii}}$$

where t has $(N - k - 1)$ degrees of freedom, and c_{ii} is defined below. However, since the b_i (i = 1,…, k) have a joint distribution and are in general not uncorrelated, care must be taken in the interpretation of sets of these confidence intervals [see, for example, Draper and Smith (1981)].

More usefully, a confidence interval on the regression equation at a point $x = (x_1, \ldots, x_k)$, where $x_i = X_i - \overline{X}_i$ is given by

$$\overline{Y} + b_1x_1 + \ldots + b_kx_k \pm ts_e\left(\frac{1}{N} + x'Cx\right)^{1/2}$$

where t has $N - k - 1$ degrees of freedom, and x'Cx is a quadratic form that takes into account the covariances and variances of the b's. (Here C is as defined in the next paragraph.)

In a similar manner, an interval for a future Y at X is given by

$$\overline{Y} + b_1x_1 + \ldots + b_kx_k \pm ts_e\left(1 + \frac{1}{N} + x'Cx\right)^{1/2}$$

where t has $N - k - 1$ degrees of freedom. (Here $C = S^{-1}$, where S is the $k \times$ k matrix whose entry in row i and column j is $\sum x_{im}X_{jm}$. In practice, as we will see, these matrix calculations are done by

the computer, and the user need not bother with them—or even understand the concept of a *matrix* and its *inverse.*)

An example: The methods discussed can now be illustrated by an example furnished by Mason E. Wescott (example 5 from *Mimeo Notes,* Mason E. Wescott, Rochester Institute of Technology, Rochester, NY) with $k = 2$ predictor variables. The problem is to relate the green strength (flexural strength before baking) of electric circuit breaker arc chutes to the hydraulic pressure used in forming them and the acid concentration. The data are given in Table 44.45, with hydraulic pressure and green strength given in units of 10 lb/in^2 and the acid concentration given as a percentage of the nominal rate for 20 observations. Two-variable plots of the data are given in Figures 44.30, 44.31, and 44.32. Summary statistics including sums, sums of squares, and crossproducts, both raw and corrected, as well as the sample means, are given in Table 44.46.

The estimates are

$$b_2 = 4.162940$$

$$b_1 = 1.571779$$

$$b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 = 16.27475$$

The C matrix is

$$c_{22} = 0.00048916$$

$$c_{12}(=c_{21}) = -0.000058099$$

$$c_{11} = 0.00029787$$

TABLE 44.45 Arc Chute Data

Green strength Y in units of 10 lb/in^2	Hydraulic pressure X_1 in units of 10 lb/in^2	Acid concentration X_2, as % of nominal rate
665	110	116
618	119	104
620	138	94
578	130	86
682	143	110
594	133	87
722	147	114
700	142	106
681	125	107
695	135	106
664	152	98
548	118	86
620	155	87
595	128	96
740	146	120
670	132	108
640	130	104
590	112	91
570	113	92
640	120	100

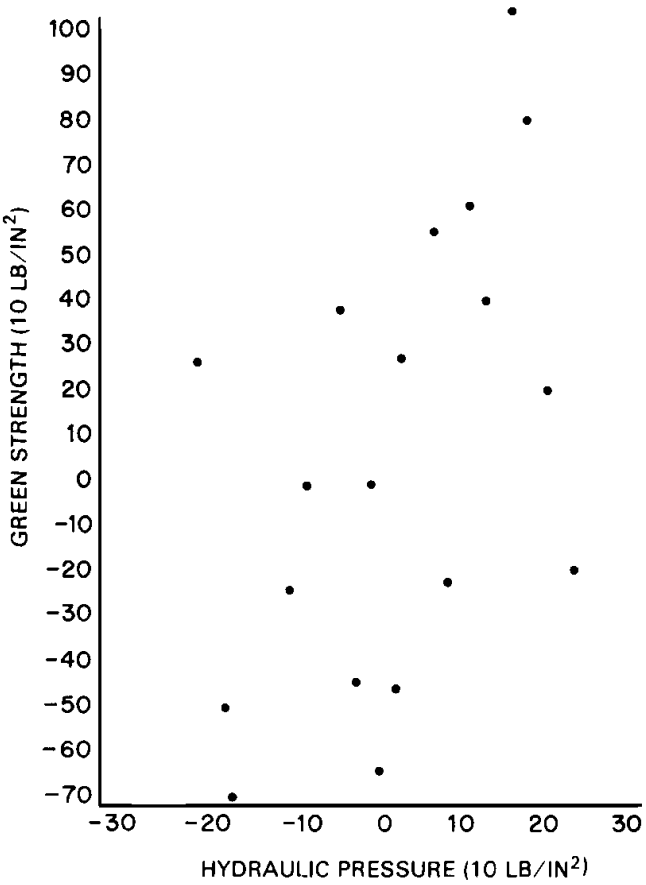


FIGURE 44.30 Green strength versus hydraulic pressure (deviations from averages).

and hence

$$C = \begin{vmatrix} 297.869 & -58.099 \\ -58.099 & 489.160 \end{vmatrix} \times 10^{-6}$$

The prediction equations can be calculated as

$$\hat{Y} = 16.277 + 1.572X_1 + 4.163X_2$$

$$\hat{Y} = 641.600 + 1.572(X_1 - 131.400) + 4.163(X_2 - 100.600)$$

(The latter form will be used in the remaining computations.)

The ANOVA table (Table 44.47) for the example follows directly from the summary statistics of Table 44.46 and from Table 44.44. The residual mean square error is 228.0, and hence the standard deviation s_e is 15.100.

The 95 percent confidence intervals on β_1 and β_2 are obtained as

$$1.572 \pm 2.110 \times 15.100 \times \sqrt{297.869} \times 10^{-3}$$

$$= 1.572 \pm 0.550 = 1.02 \text{ and } 2.12$$

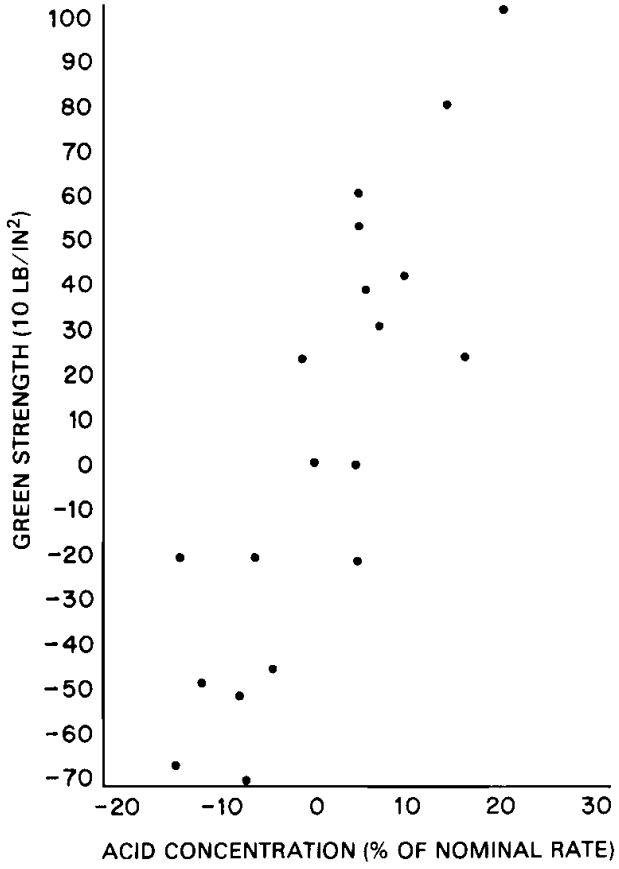


FIGURE 44.31 Green strength versus acid concentration (deviations from averages).

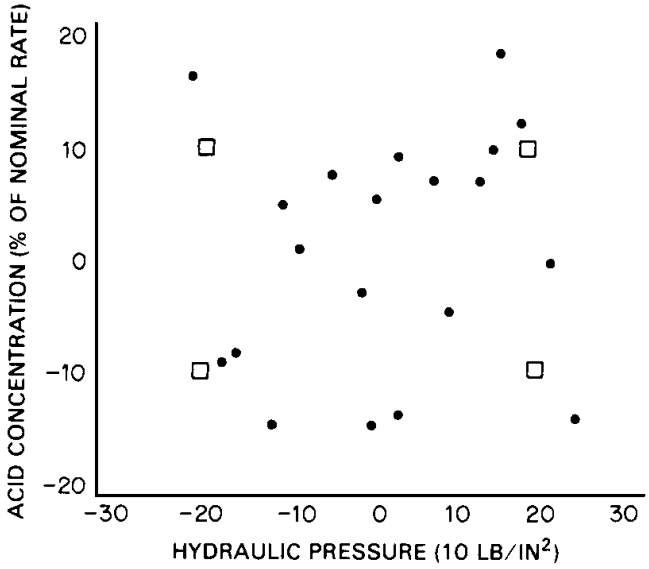


FIGURE 44.32 Acid concentration versus hydraulic pressure (deviations from averages). Rectangles indicate points at which confidence intervals have been calculated.

TABLE 44.46 Summary Statistics

$N = 20$		
$\Sigma Y_m = 12{,}832$	$\Sigma X_{1m} = 2628$	$\Sigma X_{2m} = 2012$
$\Sigma Y_m^2 = 8{,}286{,}988$	$\Sigma X_{1m}^2 = 348{,}756$	$\Sigma X_{2m}^2 = 204{,}500$
$\Sigma X_{1m} Y_m = 1{,}693{,}226$	$\Sigma X_{2m} Y_m = 1{,}300{,}253$	$\Sigma X_{1m} X_{2m} = 264{,}785$
$\overline{Y} = 641.6$	$\overline{X}_1 = 131.4$	$\overline{X}_2 = 100.6$
	S	a
$s_{11} = 3436.8$	$s_{12} = 408.2$*	$a_1 = 7{,}101.2$†
$s_{21} = 408.2$	$s_{22} = 2092.8$	$a_2 = 9{,}353.8$

*$s_{12} = \Sigma x_{1m} x_{2m} = \Sigma X_{1m} X_{2m} - (\Sigma X_{1m} \Sigma X_{2m}/N) = 264{,}785 - [(2628)(2012)/20] = 408.2$.
†$a_1 = \Sigma x_{1m} y_m = \Sigma X_{1m} Y_m - (\Sigma X_{1m} \Sigma Y_m/N) = 1{,}693{,}226 - [(2628)(12{,}832)/20] = 7101.2$.

TABLE 44.47 ANOVA Example

Source	Sum of squares	Degrees of freedom	Mean square
1. Due to regression	50,100.8*	2	25,050.40
2. Residual	3,876.0†	17	228.0
3. Total corrected for the mean	53,976.8	19	

*(1.571779)(7101.2) + (4.16294)(9353.8) = 50,100.825.
†By subtraction.

and

$$4.163 \pm 2.110 \times 15.0997 \times \sqrt{489.16} \times 10^{-3}$$

$$= 4.163 \pm 0.705 = 3.46 \text{ and } 4.87$$

Confidence intervals for $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ and for a future observation of Y, at five combinations of X_1 and X_2, are given in Table 44.48. (The five points at which the confidence intervals are computed are also indicated in Figure 44.32.)

An additional excellent example is provided in Golden and Wasil (1992, pp. 227–245), where 37,000 observations gathered from 34 stations in the Chesapeake Bay are used to develop 10 regression models for salinity dynamics. Data quality, model building, model results, and model validation are all discussed.

Computer Programs. Because of the widespread popularity of regression, almost every computer facility has at least one and most have many regression programs. These programs may have been locally written, or they may have been obtained from other sources. In the latter case, the program usually has been modified in some manner so that it can be run on the local computer system and satisfy the needs of the local users.

Studies by Longley (1967) and Wampler (1970) indicate that the user should *strongly* prefer software such as that of SAS and BMDP. One should not presume that the program has been checked just because a sample data problem is given in the program manual. Unfortunately, a number of the algorithms used in these programs are often taken directly from desk calculator instructions. These algorithms are often not good and can produce numerically inaccurate results, even in

TABLE 44.48 Computation of Confidence Intervals

X_1 (1)	X_2 (2)	$\hat{Y}$ (3)	$(x'Cx)$ (4)	$\sqrt{\frac{1}{N} + (4)}$ (5)	$ts \times (5)$† (6)	95% confidence limits on the mean of Y at X_1, X_2		$\sqrt{1 + \frac{1}{N} + (4)}$ (9)	$ts \times (9)$ (10)	95% prediction limits on Y at X_1, X_2	
						Lower (3) − (6) (7)	Upper (3) + (6) (8)			Lower (3) − (10) (11)	Upper (3) + (10) (12)
111.4	90.6	568.53	0.1457*	0.4424	14.10	554.43	582.63	1.0935	34.84	533.69	603.37
111.4	110.6	651.79	0.1922	0.4921	15.68	636.11	667.47	1.1145	35.51	616.28	687.30
131.4	100.6	641.60	0.0	0.2236	7.12	634.48	648.72	1.0247	32.65	608.95	674.25
151.4	90.6	631.41	0.1922	0.4921	15.68	615.73	647.09	1.1145	35.51	595.90	666.92
151.4	110.6	714.67	0.1457	0.4424	14.10	700.57	728.77	1.0935	34.85	679.82	749.52

Note: Numbers in parentheses are column numbers.

*Since $x' = (111.4 - 131.4, 90.6 - 100.6) = (-20, -10)$, then

$$x'Cx = (-20, -10)\begin{bmatrix} 297.869 & -58.099 \\ -58.099 & 498.160 \end{bmatrix} \times 10^{-6}\begin{bmatrix} -20 \\ -10 \end{bmatrix} = (-5{,}376.39, -3{,}819.62) \times 10^{-6}\begin{bmatrix} -20 \\ -10 \end{bmatrix} = 0.1457.$$

†t_{TAB} with 17 degrees of freedom is 2.110, $s = 15.100$.

double precision. For example, Longley found that many of the programs computed the squared deviations of a variable about its mean by the computational formula $\sum X_m^2 - (\sum X_m)^2/N$, rather than by $\sum(X_m - \overline{X})^2$. Since these quantities are the base for most of the regression calculations, numerical errors may be present in all the results.

Since the user may not be able (or simply does not want to invest the time) to check a program, a few simple checks for this purpose are given below. These can be used to gain some idea of the limitations of a program; however, I again strongly recommend against "rolling your own" when SAS, BMDP, and (perhaps) other excellent software is available at reasonable prices and is kept up to date with advances in statistics.

1. Add the residuals about the regression line. They should sum to zero, within rounding error.
2. If the residuals sum to zero, make additional runs of the problem after adding 10, 100, 1000, 10,000, etc. to each variable. The coefficients will begin to change at a point where round-off error occurs.
3. As a check on the accuracy of the inversion routine, run a problem with two variables X_1 and X_2. Then make another run with the same response variable but two new independent variables $X_1^* = X_1 + X_2$ and $X_2^* = X_1 - X_2$. The following results should hold: $b_1 = b_1^* + b_2^*$ and $b_2 = b_1^* - b_2^*$.

Longley (1967) gives some additional checks, and Wampler (1970) lists results on many regression programs.

A write-up is usually available with regression programs; it should include a complete discussion of the *input* required and an explanation of the *output* and *options* available, as well as a complete statement of the calculation formulas and a sample problem. (Although documenting a program is a difficult task, poorly written documentation is often a warning of a poorly or improperly written computer program.)

While the input formats of regression programs vary, most programs have an option that allows the user to specify a variety of transformations of the data, such as logs, powers, and crossproducts. Typical regression outputs include ANOVA tables, residuals plots, and other statistics, in addition to the estimates of the regression coefficients. Although often omitted, an *echo-check,* i.e., a printout of the original and transformed data, is essential. Often "strange" regression results can be traced to a misplaced decimal point in an observation, the wrong variables being read in, or incorrect use of the transformation option. (If this is not available in a program, ask the computer center to modify the program so that it is automatically printed out unless the user deletes it.)

R^2 and C_p Criteria for Model Choice. In many situations there are a large number of possible variables for a model, and the problem is to select the "vital few" from these "useful many," instead of obtaining the complete regression equation. There are many reasons for not using all the variables. For example, a subset of variables can provide a better prediction equation than the full set, even though the full set has a higher R, since the full set also will include more variability. In addition, equations with fewer variables are easier to understand and hence more likely to gain acceptance and be used.

Unless the data come from a properly designed experiment, there is no simple test for significant variables. Since there are $2^k - 1$ possible prediction equations to evaluate, for large k it is obvious that a brute-force approach is not feasible (e.g., if $k = 20$, then $2^k - 1 = 1{,}048{,}575$).

Stepwise regression is a heuristic technique for avoiding this computational problem. It begins by selecting the single independent variable that is the "best" predictor in the sense that it maximizes R^2. Then it adds variables to the equation in a sequential manner, in order of importance. At each step the variable added is the one that increases the regression sum of squares (and hence R^2) or equivalently reduces the residual sum of squares by the largest amount. This procedure not only selects variables but deletes variables previously selected, if at some point they no longer appear important.

Stepwise regression does not guarantee that the "best" set of variables will be included in the final equation. However, it does provide an efficient method for reducing the number of variables k to a manageable size; e.g., if $k = 100$, then stepwise regression can be used to select the best 25 or fewer for more exhaustive study. Stepwise regression programs are widely available; see Draper and Smith (1981).

In addition to stepwise regression, numerous other techniques have been developed. One that seems to have great potential was developed by Hocking and Leslie (1967) and improved by LaMotte and Hocking (1970). Their algorithm finds the "best" subset of variables of size 1, 2,…, *k,* where *k* is the total number of variables submitted. Although the computations require more time than the stepwise procedure, this procedure guarantees the best subset in the R^2 sense and in addition gives a number of the "contending" subsets. (Of course, if the number of variables is small, we could compute all regressions.) These "best" regression algorithms are available in BMDP and can handle *k* up to 20 or 25 without computer time problems.

The set of possible variables should be selected on the basis of preliminary investigations of the factors that influence the response variables. The indiscriminate use of regression analysis to "find" relationships, where no facts suggest the existence of a relationship, often leads to nonsensical results. Unfortunately, this is usually discovered after the prediction equation fails miserably in predicting future observations.

If a large amount of data is available, one portion of it can be used for selecting variables and estimating coefficients, saving the remainder of the data for testing the derived equations. In any case, the equation should be periodically reviewed as new data become available.

A powerful tool in modern regression analysis is the C_p statistic [see, e.g., Dudewicz and Karian (1985, pp. 236, 413)]. While R^2 measures the goodness of the regression equation in predicting the data points in the data set one is using to develop a model, C_p estimates the variance of future predictions made using the model. While adding a variable will increase the R^2 (even if that variable is totally unrelated to what we are trying to predict), C_p typically decreases as variables are added to the model, then increases. Thus, searching for the minimal C_p statistic in all possible regressions is a reasonable approach—though if a large gain in R^2 can be obtained with a modest increase in C_p, it should be taken.

REFERENCES

ASQC (1996). "QA/QC Software Directory." *Quality Progress,* April, pp. 31–59.

Barr, Donald R. (1969). "Using Confidence Intervals to Test Hypotheses." *Journal of Quality Technology,* vol. 1, no. 4, pp. 256–258.

Bennett, Carl A., and Franklin, Norman L. (1954). *Statistical Analysis in Chemistry and the Chemical Industry.* John Wiley & Sons, New York, chap. 11.

Bingham, R. S., Jr. (1962). "Tolerance Limits and Process Capability Studies." *Industrial Quality Control,* vol. XIX, no. 1, pp. 36–40.

Choksi, Suresh (1971). "Computer Can Optimize Manufacturing Tolerances for an Economical Design." *ASQC Technical Conference Transactions.* Milwaukee, pp. 323–330.

Cohen, A. Clifford, and Whitten, Betty Jones (1981). "Estimation of Lognormal Distributions." *American Journal of Mathematical and Management Sciences,* vol. 1, no. 2, pp. 139–153.

Daniel, Cuthbert (1977). Personal communication, August.

——— and Wood, Fred S. (1971). *Fitting Equations to Data.* John Wiley & Sons, New York.

Dielman, Terry E., and Pfaffenberger, Roger C. (1984). "Computational Algorithms for Calculating Least Absolute Value and Chebyshev Estimates for Multiple Regression." *American Journal of Mathematical and Management Sciences,* vol. 4, nos. 1 and 2, pp. 169–197.

Dixon, W. J. (1983). *BMDP Statistical Software, 1983 Printing with Additions.* University of California Press, Berkeley, CA.

——— and Massey, Frank J., Jr. (1969). *Introduction to Statistical Analysis,* 3rd ed. McGraw-Hill, New York.

Dodge, H. F., Kinsburg, B. J., and Kruger, M. K. (1953). "The L3 Coaxial System—Quality Control Requirements." *Bell System Technical Journal,* vol. 32, pp. 943–1005.

Draper, N. R., and Smith, H. (1981). *Applied Regression Analysis,* 2nd ed. John Wiley & Sons, New York.

Dudewicz, Edward J. (1976). *Introduction to Statistics and Probability.* Holt, Rinehart and Winston, New York. (Currently published by American Sciences Press, Columbus, OH. Address: P.O. Box 21161, Columbus 43221-0161.)

——— (1983). "Heteroscedasticity." In N. L. Johnson, S. Kotz, and C. B. Read (eds.), *Encyclopedia of Statistical Sciences,* vol. 3. John Wiley & Sons, New York, pp. 611–619.

——— (1992). "The Generalized Bootstrap." In Jöckel, K.-H., Rothe, G., and Sendler, W. (eds.), *Bootstrapping and Related Techniques,* vol. 376 of Lecture Notes in Economics and Mathematical Systems. Springer-Verlag, Berlin, pp. 31–37.

——— (1993). *Solutions in Statistics and Probability.* 2nd ed. American Sciences Press, Columbus, OH.

——— (1995). "The Heteroscedastic Method: Fifty+ Years of Progress 1945–2000." In Hayakawa, T., Aoshima, M., and Shimizu, K. (eds.), *MSI-2000: Multivariate Statistical Analysis,* vol. I. American Sciences Press, Columbus, OH, pp. 179–197.

——— (ed.) (1997). *Modern Digital Simulation Methodology, III: Advances in Theory, Application, and Design—Electric Power Systems, Spare Parts Inventory, Purchase Interval and Incidence Modeling, Automobile Insurance Bonus-Malus Systems, Genetic Algorithms—DNA Sequence Assembly, Education, & Water Resources Case Studies.* American Sciences Press, Columbus, OH.

——— and Dalal, Siddhartha R. (1983). "Multiple-Comparisons with a Control when Variances Are Unknown and Unequal." *American Journal of Mathematical and Management Sciences,* vol. 3, no. 4, pp. 275–295.

——— and Karian, Zaven A. (1985). *Modern Design and Analysis of Discrete-Event Computer Simulations.* IEEE Computer Society Press, Los Angeles, CA. (Address is P.O. Box 80452, Worldway Postal Center, Los Angeles 90080; Order No. CN597.)

——— and ——— (1996). "The EGLD (Extended Generalized Lambda Distribution) System for Fitting Distributions to Data with Moments: II. Tables." In Dudewicz, E. J. (ed.), *Modern Digital Simulation Methodology, II: Univariate and Bivariate Distribution Fitting, Bootstrap Methods, and Applications to CensusPES and CensusPlus of the U.S. Bureau of the Census, Bootstrap Sample Size, and Biology and Environment Case Studies.* American Sciences Press, Columbus, OH, pp. 271–332.

——— and Koo, J. O. (1982). *The Complete Categorized Guide to Statistical Selection and Ranking Procedures.* American Sciences Press, Columbus, OH.

——— and Mishra, S. N. (1988). *Modern Mathematical Statistics.* John Wiley & Sons, New York.

——— and Ralley, Thomas G. (1981). *The Handbook of Random Number Generation and Testing with TESTRAND Computer Code.* American Sciences Press, Columbus, OH.

———, Karian, Zaven A., and Marshall, Rudolph James, III (1985). "Random Number Generation on Microcomputers." *Proceedings of the Conference on Simulation on Microcomputers.* The Society for Computer Simulation (Simulation Councils, Inc.), La Jolla, CA. (Address: P.O. Box 2228, La Jolla 92038.)

Duncan, Acheson J. (1974). *Quality Control and Industrial Statistics,* 4th ed. Richard D. Irwin, Homewood, IL.

Edissonov, I. (1994). "The New ARSTI Optimization Method: Adaptive Random Search with Translating Intervals." *American Journal of Mathematical and Management Sciences,* vol. 14, pp. 143–166.

Epstein, Benjamin (1960). "Estimation from Life Test Data." *Technometrics,* vol. 2, no. 4, pp. 447–454.

Evans, Ralph A. (1966). "Problems in Probability." *Proceedings of the Annual Symposium on Reliability.* IEEE, New York, pp. 347–353.

Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics,* vol. 7, pp. 179–188.

Gibbons, Jean Dickinson (1997). *Nonparametric Methods for Quantitative Analysis,* 3rd ed. American Sciences Press, Columbus, OH.

Golden, Bruce L. (ed.) (1993). *Vehicle Routing 2000: Advances in Time-Windows, Optimality, Fast Bounds, and Multi-Depot Routing.* American Sciences Press, Columbus, OH.

——— and Eiselt, H. A. (eds.) (1992). *Location Modeling in Practice: Applications (Site Location, Oil Field Generators, Emergency Facilities, Postal Boxes), Theory, and History.* American Sciences Press, Columbus, OH.

——— and Wasil, Edward A. (eds.) (1992). *Fisheries: Control and Management via Application of Management Sciences (with a Comprehensive Annotated Bibliography of Applications).* American Sciences Press, Columbus, OH.

———, Assad, Arjang A., and Zanakis, Stelios H. (eds.) (1984). *Statistics and Optimization: The Interface.* American Sciences Press, Columbus, OH.

Govindarajulu, Z. (1981). *The Sequential Statistical Analysis of Hypothesis Testing: Point and Interval Estimation, and Decision Theory.* American Sciences Press, Columbus, OH.

Grant, E. L., and Leavenworth, R. S. (1980). *Statistical Quality Control,* 5th ed., McGraw-Hill, New York.

Gray, H. L., and Lewis, Truman (1967). "A Confidence Interval for the Availability Ratio." *Technometrics,* vol. 9, no. 3, pp. 465–471.

Grubbs, Frank E. (1969). "Procedures for Detecting Outlying Observations in Samples." *Technometrics,* vol. 11, no. 1, pp. 1–21.

Gutt, J. D., and Gruska, G. F. (1977). "Variation Simulation." *ASQC Technical Conference Transactions.* Milwaukee, pp. 557–563.

Hadley, G. (1967). *Introduction to Probability and Statistical Decision Theory.* Holden-Day, San Francisco.

Hahn, G. J. (1970*a*). "Additional Factors for Calculating Prediction Intervals for Samples from a Normal Distribution." *Journal of the American Statistical Association,* vol. 65, no. 332, pp. 1668–1676.

——— (1970*b*). "Statistical Intervals for a Normal Population: 1. Tables, Examples, and Applications; II. Formulas, Assumptions, Some Derivations." *Journal of Quality Technology,* vol. 2, no. 3, pp. 115–125; vol. 2, no. 4, pp. 195–206.

——— (1971). "How Abnormal is Normality?" *Journal of Quality Technology,* vol. 3, no. 1, pp. 18–22.

——— (1972). "The Absolute Sample Size Is What Counts." *Quality Progress,* vol. V, no. 5, pp. 18–19.

——— and Shapiro, S. S. (1967). *Statistical Models in Engineering.* John Wiley & Sons, New York.

Harter, H. Leon (1983). *The Chronological Annotated Bibliography of Order Statistics,* vol. II: *1950–1959.* American Sciences Press, Columbus, OH.

——— and Dubey, Satya D. (1967). *Theory and Tables for Tests of Hypotheses Concerning the Mean and the Variance of a Weibull Population.* Document AD 653593, Clearinghouse for Federal Scientific and Technical Information, U.S. Department of Commerce, Washington.

Hocking, R. R., and Leslie, R. N. (1967). "Selection of the Best Subset in Regression Analysis." *Technometrics,* vol. 9, no. 4, pp. 531–540.

Hoerl, Arthur E., and Kennard, Robert W. (1981). "Ridge Regression—1980. Advances, Algorithms, and Applications." *American Journal of Mathematical and Management Sciences,* vol. 1, no. 1, pp. 5–83.

Huh, M.-H., and Olkin, I. (1995). "Asymptotic Aspects of Ordinary Ridge Regression." In Hayakawa, T., Aoshima, M., and Shimizu, K. (eds.), *MSI-2000: Multivariate Statistical Analysis,* vol. I. American Sciences Press, Columbus, OH, pp. 239–254.

Iman, Ronald L. (1982). "Graphs for Use with the Lillefors Test for Normal and Exponential Distributions." *The American Statistician,* vol. 36, no. 2, pp. 109–112.

Ireson, W. Grant (ed.) (1982). *Reliability Handbook.* McGraw-Hill, New York.

Johnson, Norman L., and Kotz, Samuel (1969). *Discrete Distributions.* Houghton Mifflin, Boston.

——— and Leone, Fred C. (1964). *Statistics and Experimental Design,* vol. I. John Wiley & Sons, New York.

Juran, J. M., and Gryna, Frank M., Jr. (1980). *Quality Planning and Analysis,* 2nd ed. McGraw-Hill, New York.

Karian, Z., and Dudewicz, E. J. (1991). *Modern Statistical, Systems, and GPSS Simulation.* Computer Science Press/W.H. Freeman and Company, Book Publishing Division of Scientific American, New York. Currently published by American Sciences Press, Columbus, OH.

———, Dudewicz, E. J., and McDonald, P. (1996). "The Extended Generalized Lambda Distribution System for Fitting Distributions to Data: History, Completion of Theory, Tables, Applications, the `Final Word' on Moment Fits." *Communications in Statistics: Simulation and Computation,* vol. 25, pp. 611–642.

——— and Dudewicz, Edward J. (1999). *Fitting Statistical Distributions to Data: The Generalized Lambda Distribution (GLD) and the Generalized Bootstrap (GB) Methods.* CRC Press, Boca Raton, FL.

Kececioglu, D., and Cormier, D. (1964). "Designing a Specified Reliability into a Component." *Proceedings of the Third Annual Aerospace Reliability and Maintainability Conference.* Society of Automotive Engineers, Washington, p. 546.

King, James R. (1971). *Probability Charts for Decision Making.* The Industrial Press, New York.

Kirkpatrick, R. L. (1970). "Confidence Limits on a Percent Defective Characterized by Two Specification Limits." *Journal of Quality Technology,* vol. 2, no. 3, pp. 150–155.

Kramer, C. Y., and Jensen, D. R. (1969, 1970). "Fundamentals of Multivariate Analysis." *Journal of Quality Technology,* vol. 1, no. 2, pp. 120–133, no. 3, pp. 189–204, no. 4, pp. 264–276; vol. 2, no. 1, pp. 32–40.

LaMotte, L. R., and Hocking, R. R. (1970). "Computational Efficiency in the Selection of Regression Variables." *Technometrics,* vol. 12, no. 1, pp. 83–93.

Lenz, H.-J., and Rendtel, U. (1984). "Performance Evaluation of the MIL-STD-105 D, Skip-Lot Sampling Plans, and Bayesian Single Sampling Plans." In Lenz, H.-J., Wetherill, G. B., and Wilrich, P.-Th. (eds.), *Frontiers in Statistical Quality Control 2.* Physica-Verlag, Rudolf Liebing GmbH & Co., Würzburg, Germany, pp. 92–106.

Levy, George C., and Dumoulin, Charles L. (1984). *The LAB ONE NMRI Spectroscopic Data Analysis Software System: Revision 2.00 User's Manual.* Department of Chemistry, Syracuse University, Syracuse, NY (Address: Dept. of Chemistry, Browne Hall, Syracuse University, Syracuse 13244).

Lieberman, G. J., and Owen, D. B. (1961). *Tables of the Hypergeometric Probability Distribution.* Stanford University Press, Stanford, CA.

Lloyd, David K., and Lipow, Myron (1982). *Reliability: Management, Methods and Mathematics,* 2nd ed. Prentice-Hall, Englewood Cliffs, NJ.

Longley, J. W. (1967). "An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User." *Journal of the American Statistical Association,* vol. 62, no. 319, pp. 819–841.

Lusser, R. (1958). *Reliability Through Safety Margins.* United States Army Ordnance Missile Command, Redstone Arsenal, AL.

Martin Marietta Corp. (1966). *Reliability for the Engineer,* book 5: *Testing for Reliability.* Martin Marietta, Orlando, FL, pp. 29–31.

Moult, John F. (1963). "Critical Agents in Bearing Fatigue Testing." *Lubrication Engineering,* December, pp. 503–511.

Mouradian, G. (1966). "Tolerance Limits for Assemblies and Engineering Relationships." *ASQC Technical Conference Transactions.* Milwaukee, pp. 598–606.

Mulekar, Madhuri S., and Mishra, Satya N. (1997). "Inference on Measures of Niche Overlap for Heteroscedastic Normal Populations." In Hayakawa, T., Aoshima, M., and Shimizu, K. (eds.), *MSI-2000: Multivariate Statistical Analysis,* vol. III. American Sciences Press, Columbus, OH, pp.163-185.

Natrella, Mary G. (1963). *Experimental Statistics, National Bureau of Standards Handbook 91.* U.S. Government Printing Office, Washington, DC.

Nelson, Wayne (1979). *How to Analyze Data with Simple Plots.* Volume 1 of Dudewicz, E. J. (ed.). *The ASQC Basic References in Quality Control: Statistical Techniques.* American Society for Quality Control, Milwaukee.

——— (1982). *Applied Life Data Analysis.* John Wiley & Sons, New York.

Oliver, Larry R., and Springer, Melvin D. (1972). "A General Set of Bayesian Attribute Acceptance Sampling Plans." *Technical Papers of the 1972 Conference, American Institute of Industrial Engineers,* pp. 443–455.

Owen, D. B., and Frawley, W. H. (1971). "Factors for Tolerance Limits Which Control Both Tails of the Normal Distribution." *Journal of Quality Technology,* vol. 3, no. 2, pp. 69–79. See also Frawley, W. H., Kapadia, C. H., Rao, J. N. K., and Owen, D. B. (1971). "Tolerance Limits Based on Range and Mean Range." *Technometrics,* vol. 13, no. 3, pp. 651–656.

Peters, J. (1970). "Tolerancing the Components of an Assembly for Minimum Cost." ASME Paper 70-Prod.-9. *Transactions of the ASME Journal of Engineering for Industry.* American Society of Mechanical Engineers, New York.

Pinsky, Paul D. (1983). *Statpro: The Statistics and Graphics Database Workstation, Statistics User's Guide.* Wadsworth Electronic Publishing Company, Boston (Address: Statler Office Building, 20 Park Plaza, Boston 02116).

Ramberg, John S., Dudewicz, Edward J., Tadikamalla, Pandu R., and Mykytka, Edward F. (1979). "A Probability Distribution and its Uses in Fitting Data." *Technometrics,* vol. 21, no. 2, pp. 201–214.

Ranganathan, J., and Kale, B. K. (1983). "Outlier-Resistant Tolerance Intervals for Exponential Distributions." *American Journal of Mathematical and Management Sciences,* vol. 3, no. 1, pp. 5–25.

Rich, Barrett G., Smith, O. A., and Korte, Lee (1967). *Experience with a Formal Reliability Program.* SAE Paper 670731, Farm, Construction and Industrial Machinery Meeting, Society of Automotive Engineers, Warrendale, PA.

Romeu, J. L., and Ozturk, A. (1996). "A New Graphical Test for Multivariate Normality." In Hayakawa, T., Aoshima, M., and Shimizu, K. (eds.), *MSI-2000: Multivariate Statistical Analysis,* vol. II. American Sciences Press, Columbus, OH, pp. 5–48.

Rutherford, John R. (1971). "A Logic Structure for Experimental Development Programs." *Chemical Technology,* March, pp. 159–164.

SAS Institute, Inc. (1982). *SAS User's Guide: Basics,* 1982 Edition. SAS Institute, Cary, NC (Address: P.O. Box 8000, Cary 27511).

Schmid, C. F. (1983). *Statistical Graphics, Design Principles and Practices.* John Wiley & Sons, New York.

Sheesley, J. H. (1977). "Tests for Outlying Observations." *Journal of Quality Technology,* vol. 9, no. 1, pp. 38–41.

Siotani, Minoru, Hayakawa, T., and Fujikoshi, Y. (1985). *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook.* American Sciences Press, Columbus, OH.

Sun, L., and Müller-Schwarze, D. (1996). "Statistical Resampling Methods in Biology: A Case Study of Beaver Dispersal Patterns." *American Journal of Mathematical and Management Sciences,* vol. 16, pp. 463–502.

Thoman, D. R., Bain, L. J., and Antle, C. E. (1970). "Maximum Likelihood Estimation. Exact Confidence Intervals for Reliability and Tolerance Limits in the Weibull Distribution." *Technometrics,* vol. 12, no. 2, pp. 363–371.

Volk, William (1956). "Industrial Statistics." *Chemical Engineering,* March, pp. 165–190.

Wampler, R. H. (1970). "On the Accuracy of Least Squares Computer Programs." *Journal of the American Statistical Association,* vol. 65, no. 330, pp. 549–565.

Wardrop, R. L. (1995). "Simpson's Paradox and the Hot Hand in Basketball." *The American Statistician,* vol. 49, no. 1, pp. 24–28.

Wescott, M. E. *Mimeo Notes.* Rochester Institute of Technology, Rochester, NY.

Western Electric Company, Inc. (1982). *Statistical Quality Control Handbook.* American Society for Quality Control, Milwaukee.

Williamson, E., and Bretherton, M. H. (1963). *Tables of the Negative Binomial Probability Distribution.* John Wiley & Sons, New York.

Zacks, Shelemyahu (1983). *Workshop on Statistical Methods of Reliability Analysis for Engineers.* Center for Statistics, Quality Control, and Design, State University of New York, Binghamton, NY.