

# 6

## Random Sampling and Data Description

---

---

### CHAPTER OUTLINE

- |   |   |
|---|---|
| 6-1 DATA SUMMARY AND DISPLAY                  | 6-5 BOX PLOTS                                   |
| 6-2 RANDOM SAMPLING                           | 6-6 TIME SEQUENCE PLOTS                         |
| 6-3 STEM-AND-LEAF DIAGRAMS                    | 6-7 PROBABILITY PLOTS                           |
| 6-4 FREQUENCY DISTRIBUTIONS<br>AND HISTOGRAMS | 6-8 MORE ABOUT PROBABILITY<br>PLOTING (CD ONLY) |
- 

### LEARNING OBJECTIVES

After careful study of this chapter you should be able to do the following:

1. Compute and interpret the sample mean, sample variance, sample standard deviation, sample median, and sample range
2. Explain the concepts of sample mean, sample variance, population mean, and population variance
3. Construct and interpret visual data displays, including the stem-and-leaf display, the histogram, and the box plot
4. Explain the concept of random sampling
5. Construct and interpret normal probability plots
6. Explain how to use box plots and other data displays to visually compare two or more samples of data
7. Know how to use simple time series plots to visually display the important features of time-oriented data.

### CD MATERIAL

8. Interpret probability plots for distributions other than normal.
- 

Answers for most odd numbered exercises are at the end of the book. Answers to exercises whose numbers are surrounded by a box can be accessed in the e-Text by clicking on the box. Complete worked solutions to certain exercises are also available in the e-Text. These are indicated in the Answers to Selected Exercises section by a box around the exercise number. Exercises are also available for some of the text sections that appear on CD only. These exercises may be found within the e-Text immediately following the section they accompany.

## 6-1 DATA SUMMARY AND DISPLAY

Well-constructed data summaries and displays are essential to good statistical thinking, because they can focus the engineer on important features of the data or provide insight about the type of model that should be used in solving the problem. The computer has become an important tool in the presentation and analysis of data. While many statistical techniques require only a hand-held calculator, much time and effort may be required by this approach, and a computer will perform the tasks much more efficiently.

Most statistical analysis is done using a prewritten library of statistical programs. The user enters the data and then selects the types of analysis and output displays that are of interest. Statistical software packages are available for both mainframe machines and personal computers. We will present examples of output from Minitab (one of the most widely-used PC packages), throughout the book. We will not discuss the hands-on use of Minitab for entering and editing data or using commands. This information is found in the software documentation.

We often find it useful to describe data features **numerically**. For example, we can characterize the location or central tendency in the data by the ordinary arithmetic average or mean. Because we almost always think of our data as a sample, we will refer to the arithmetic mean as the **sample mean**.

### Definition

If the  $n$  observations in a sample are denoted by  $x_1, x_2, \dots, x_n$ , the **sample mean** is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} \quad (6-1)$$

### EXAMPLE 6-1

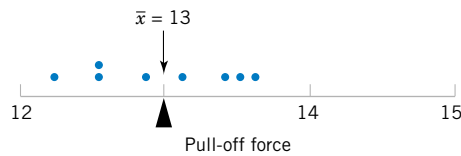
Let's consider the eight observations collected from the prototype engine connectors from Chapter 1. The eight observations are  $x_1 = 12.6$ ,  $x_2 = 12.9$ ,  $x_3 = 13.4$ ,  $x_4 = 12.3$ ,  $x_5 = 13.6$ ,  $x_6 = 13.5$ ,  $x_7 = 12.6$ , and  $x_8 = 13.1$ . The sample mean is

$$\begin{aligned} \bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^8 x_i}{8} = \frac{12.6 + 12.9 + \cdots + 13.1}{8} \\ &= \frac{104}{8} = 13.0 \text{ pounds} \end{aligned}$$

A physical interpretation of the sample mean as a measure of location is shown in the dot diagram of the pull-off force data. See Figure 6-1. Notice that the sample mean  $\bar{x} = 13.0$  can be thought of as a "balance point." That is, if each observation represents 1 pound of mass placed at the point on the  $x$ -axis, a fulcrum located at  $\bar{x}$  would exactly balance this system of weights.

The sample mean is the average value of all the observations in the data set. Usually, these data are a **sample** of observations that have been selected from some larger **population** of observations. Here the population might consist of all the connectors that will be manufactured and sold to customers. Recall that this type of population is called a **conceptual** or

**Figure 6-1** The sample mean as a balance point for a system of weights.



**hypothetical population**, because it does not physically exist. Sometimes there is an actual physical population, such as a lot of silicon wafers produced in a semiconductor factory.

In previous chapters we have introduced the mean of a probability distribution, denoted  $\mu$ . If we think of a probability distribution as a **model** for the population, one way to think of the mean is as the average of all the measurements in the population. For a finite population with  $N$  measurements, the mean is

$$\mu = \frac{\sum_{i=1}^N x_i}{N} \quad (6-2)$$

The sample mean,  $\bar{x}$ , is a reasonable estimate of the population mean,  $\mu$ . Therefore, the engineer designing the connector using a 3/32-inch wall thickness would conclude, on the basis of the data, that an estimate of the mean pull-off force is 13.0 pounds.

Although the sample mean is useful, it does not convey all of the information about a sample of data. The variability or scatter in the data may be described by the **sample variance** or the **sample standard deviation**.

#### Definition

If  $x_1, x_2, \dots, x_n$  is a sample of  $n$  observations, the **sample variance** is

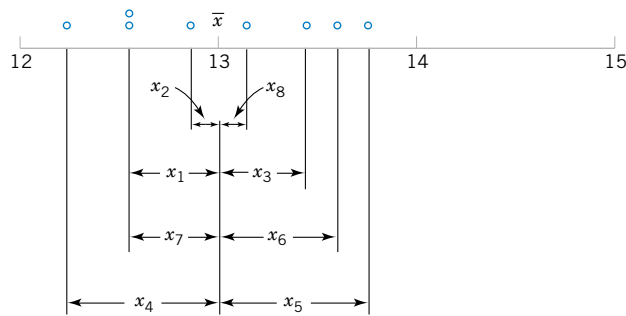
$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (6-3)$$

The **sample standard deviation**,  $s$ , is the positive square root of the sample variance.

The units of measurements for the sample variance are the square of the original units of the variable. Thus, if  $x$  is measured in pounds, the units for the sample variance are (pounds)<sup>2</sup>. The standard deviation has the desirable property of measuring variability in the original units of the variable of interest,  $x$ .

#### How Does the Sample Variance Measure Variability?

To see how the sample variance measures dispersion or variability, refer to Fig. 6-2, which shows the deviations  $x_i - \bar{x}$  for the connector pull-off force data. The greater the amount of variability in the pull-off force data, the larger in absolute magnitude some of the deviations  $x_i - \bar{x}$  will be. Since the deviations  $x_i - \bar{x}$  always sum to zero, we must use a measure of variability that changes the negative deviations to nonnegative quantities. Squaring the deviations is the approach used in the sample variance. Consequently, if  $s^2$  is small, there is relatively little variability in the data, but if  $s^2$  is large, the variability is relatively large.



**Figure 6-2** How the sample variance measures variability through the deviations  $x_i - \bar{x}$ .

### EXAMPLE 6-2

Table 6-1 displays the quantities needed for calculating the sample variance and sample standard deviation for the pull-off force data. These data are plotted in Fig. 6-2. The numerator of  $s^2$  is

$$\sum_{i=1}^8 (x_i - \bar{x})^2 = 1.60$$

so the sample variance is

$$s^2 = \frac{1.60}{8 - 1} = \frac{1.60}{7} = 0.2286 \text{ (pounds)}^2$$

and the sample standard deviation is

$$s = \sqrt{0.2286} = 0.48 \text{ pounds}$$

### Computation of $s^2$

The computation of  $s^2$  requires calculation of  $\bar{x}$ ,  $n$  subtractions, and  $n$  squaring and adding operations. If the original observations or the deviations  $x_i - \bar{x}$  are not integers, the deviations  $x_i - \bar{x}$  may be tedious to work with, and several decimals may have to be carried to ensure

**Table 6-1** Calculation of Terms for the Sample Variance and Sample Standard Deviation

$i$	$x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	12.6	-0.4	0.16
2	12.9	-0.1	0.01
3	13.4	0.4	0.16
4	12.3	-0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	-0.4	0.16
8	13.1	0.1	0.01
	104.0	0.0	1.60

numerical accuracy. A more efficient computational formula for the sample variance is obtained as follows:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{\sum_{i=1}^n (x_i^2 + \bar{x}^2 + 2\bar{x}x_i)}{n - 1} = \frac{\sum_{i=1}^n x_i^2 + n\bar{x}^2 - 2\bar{x} \sum_{i=1}^n x_i}{n - 1}$$

and since  $\bar{x} = (1/n) \sum_{i=1}^n x_i$ , this last equation reduces to

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}{n - 1} \quad (6-4)$$

Note that Equation 6-4 requires squaring each individual  $x_i$ , then squaring the sum of the  $x_i$ , subtracting  $(\sum x_i)^2/n$  from  $\sum x_i^2$ , and finally dividing by  $n - 1$ . Sometimes this is called the shortcut method for calculating  $s^2$  (or  $s$ ).

#### EXAMPLE 6-3

We will calculate the sample variance and standard deviation using the shortcut method, Equation 6-4. The formula gives

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}{n - 1} = \frac{1353.6 - \frac{(104)^2}{8}}{7} = \frac{1.60}{7} = 0.2286 \text{ (pounds)}^2$$

and

$$s = \sqrt{0.2286} = 0.48 \text{ pounds}$$

These results agree exactly with those obtained previously.

Analogous to the sample variance  $s^2$ , the variability in the population is defined by the **population variance** ( $\sigma^2$ ). As in earlier chapters, the positive square root of  $\sigma^2$ , or  $\sigma$ , will denote the **population standard deviation**. When the population is finite and consists of  $N$  values, we may define the population variance as

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \quad (6-5)$$

We observed previously that the sample mean could be used as an estimate of the population mean. Similarly, the sample variance is an estimate of the population variance. In Chapter 7, we will discuss **estimation of parameters** more formally.

Note that the divisor for the sample variance is the sample size minus one ( $n - 1$ ), while for the population variance it is the population size  $N$ . If we knew the true value of the population mean  $\mu$ , we could find the *sample* variance as the average squared deviation of the sample observations about  $\mu$ . In practice, the value of  $\mu$  is almost never known, and so the sum of

the squared deviations about the sample average  $\bar{x}$  must be used instead. However, the observations  $x_i$  tend to be closer to their average,  $\bar{x}$ , than to the population mean,  $\mu$ . Therefore, to compensate for this we use  $n - 1$  as the divisor rather than  $n$ . If we used  $n$  as the divisor in the sample variance, we would obtain a measure of variability that is, on the average, consistently smaller than the true population variance  $\sigma^2$ .

Another way to think about this is to consider the sample variance  $s^2$  as being based on  $n - 1$  **degrees of freedom**. The term *degrees of freedom* results from the fact that the  $n$  deviations  $x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$  always sum to zero, and so specifying the values of any  $n - 1$  of these quantities automatically determines the remaining one. This was illustrated in Table 6-1. Thus, only  $n - 1$  of the  $n$  deviations,  $x_i - \bar{x}$ , are freely determined.

In addition to the sample variance and sample standard deviation, the **sample range**, or the difference between the largest and smallest observations, is a useful measure of variability. The sample range is defined as follows.

### Definition

If the  $n$  observations in a sample are denoted by  $x_1, x_2, \dots, x_n$ , the **sample range** is

$$r = \max(x_i) - \min(x_i) \quad (6-6)$$

For the pull-off force data, the sample range is  $r = 13.6 - 12.3 = 1.3$ . Generally, as the variability in sample data increases, the sample range increases.

The sample range is easy to calculate, but it ignores all of the information in the sample data between the largest and smallest values. For example, the two samples 1, 3, 5, 8, and 9 and 1, 5, 5, 5, and 9, both have the same range ( $r = 8$ ). However, the standard deviation of the first sample is  $s_1 = 3.35$ , while the standard deviation of the second sample is  $s_2 = 2.83$ . The variability is actually less in the second sample.

Sometimes, when the sample size is small, say  $n < 8$  or 10, the information loss associated with the range is not too serious. For example, the range is used widely in statistical quality control where sample sizes of 4 or 5 are fairly common. We will discuss some of these applications in Chapter 16.

## EXERCISES FOR SECTIONS 6-1 AND 6-2

**6-1.** Eight measurements were made on the inside diameter of forged piston rings used in an automobile engine. The data (in millimeters) are 74.001, 74.003, 74.015, 74.000, 74.005, 74.002, 74.005, and 74.004. Calculate the sample mean and sample standard deviation, construct a dot diagram, and comment on the data.

**6-2.** In *Applied Life Data Analysis* (Wiley, 1982), Wayne Nelson presents the breakdown time of an insulating fluid between electrodes at 34 kV. The times, in minutes, are as follows: 0.19, 0.78, 0.96, 1.31, 2.78, 3.16, 4.15, 4.67, 4.85, 6.50, 7.35, 8.01, 8.27, 12.06, 31.75, 32.52, 33.91, 36.71, and 72.89. Calculate the sample mean and sample standard deviation.

**6-3.** The January 1990 issue of *Arizona Trend* contains a supplement describing the 12 “best” golf courses in the state. The yardages (lengths) of these courses are as follows: 6981, 7099, 6930, 6992, 7518, 7100, 6935, 7518, 7013, 6800, 7041,

and 6890. Calculate the sample mean and sample standard deviation. Construct a dot diagram of the data.

**6-4.** An article in the *Journal of Structural Engineering* (Vol. 115, 1989) describes an experiment to test the yield strength of circular tubes with caps welded to the ends. The first yields (in kN) are 96, 96, 102, 102, 102, 104, 104, 108, 126, 126, 128, 128, 140, 156, 160, 160, 164, and 170. Calculate the sample mean and sample standard deviation. Construct a dot diagram of the data.

**6-5.** An article in *Human Factors* (June 1989) presented data on visual accommodation (a function of eye movement) when recognizing a speckle pattern on a high-resolution CRT screen. The data are as follows: 36.45, 67.90, 38.77, 42.18, 26.72, 50.77, 39.30, and 49.71. Calculate the sample mean and sample standard deviation. Construct a dot diagram of the data.

**6-6.** The following data are direct solar intensity measurements (watts/m<sup>2</sup>) on different days at a location in southern Spain: 562, 869, 708, 775, 775, 704, 809, 856, 655, 806, 878, 909, 918, 558, 768, 870, 918, 940, 946, 661, 820, 898, 935, 952, 957, 693, 835, 905, 939, 955, 960, 498, 653, 730, and 753. Calculate the sample mean and sample standard deviation.

**6-7.** The April 22, 1991 issue of *Aviation Week and Space Technology* reports that during Operation Desert Storm, U.S. Air Force F-117A pilots flew 1270 combat sorties for a total of 6905 hours. What is the mean duration of an F-117A mission during this operation? Why is the parameter you have calculated a population mean?

**6-8.** Preventing fatigue crack propagation in aircraft structures is an important element of aircraft safety. An engineering study to investigate fatigue crack in  $n = 9$  cyclically loaded wing boxes reported the following crack lengths (in mm): 2.13, 2.96, 3.02, 1.82, 1.15, 1.37, 2.04, 2.47, 2.60.

- Calculate the sample mean.
- Calculate the sample variance and sample standard deviation.
- Prepare a dot diagram of the data.

**6-9.** Consider the solar intensity data in Exercise 6-6. Prepare a dot diagram of this data. Indicate where the sample mean falls on this diagram. Give a practical interpretation of the sample mean.

**6-10.** Exercise 6-5 describes data from an article in *Human Factors* on visual accommodation from an experiment involving a high-resolution CRT screen.

- Construct a dot diagram of this data.
- Data from a second experiment using a low-resolution screen were also reported in the article. They are 8.85,

35.80, 26.53, 64.63, 9.00, 15.38, 8.14, and 8.24. Prepare a dot diagram for this second sample and compare it to the one for the first sample. What can you conclude about CRT resolution in this situation?

**6-11.** The pH of a solution is measured eight times by one operator using the same instrument. She obtains the following data: 7.15, 7.20, 7.18, 7.19, 7.21, 7.20, 7.16, and 7.18.

- Calculate the sample mean.
- Calculate the sample variance and sample standard deviation.
- What are the major sources of variability in this experiment?

**6-12.** An article in the *Journal of Aircraft* (1988) describes the computation of drag coefficients for the NASA 0012 airfoil. Different computational algorithms were used at  $M_\infty = 0.7$  with the following results (drag coefficients are in units of drag counts; that is, one count is equivalent to a drag coefficient of 0.0001): 79, 100, 74, 83, 81, 85, 82, 80, and 84. Compute the sample mean, sample variance, and sample standard deviation, and construct a dot diagram.

**6-13.** The following data are the joint temperatures of the O-rings (°F) for each test firing or actual launch of the space shuttle rocket motor (from *Presidential Commission on the Space Shuttle Challenger Accident*, Vol. 1, pp. 129–131): 84, 49, 61, 40, 83, 67, 45, 66, 70, 69, 80, 58, 68, 60, 67, 72, 73, 70, 57, 63, 70, 78, 52, 67, 53, 67, 75, 61, 70, 81, 76, 79, 75, 76, 58, 31.

- Compute the sample mean and sample standard deviation.
- Construct a dot diagram of the temperature data.
- Set aside the smallest observation (31°F) and recompute the quantities in part (a). Comment on your findings. How “different” are the other temperatures from this last value?

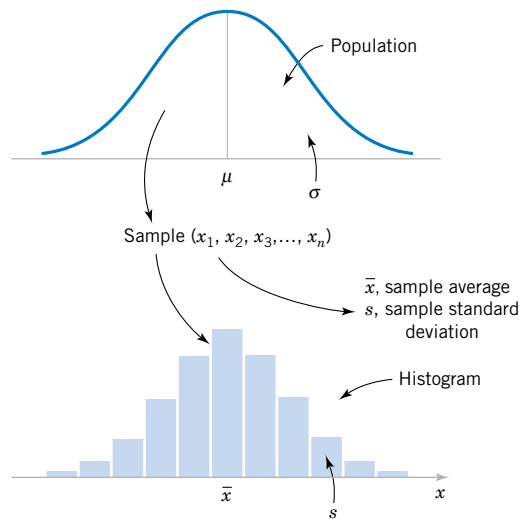
## 6-2 RANDOM SAMPLING

In most statistics problems, we work with a sample of observations selected from the population that we are interested in studying. Figure 6-3 illustrates the relationship between the population and the sample. We have informally discussed these concepts before; however, we now give the formal definitions of some of these terms.

### Definition

A **population** consists of the totality of the observations with which we are concerned.

In any particular problem, the population may be small, large but finite, or infinite. The number of observations in the population is called the **size** of the population. For example, the number of underfilled bottles produced on one day by a soft-drink company is a population of finite size. The observations obtained by measuring the carbon monoxide level every day is a population of infinite size. We often use a **probability distribution** as a **model** for a population. For example, a structural engineer might consider the population of tensile strengths of a



**Figure 6-3** Relationship between a population and a sample.

chassis structural element to be normally distributed with mean  $\mu$  and variance  $\sigma^2$ . We could refer to this as a **normal population** or a normally distributed population.

In most situations, it is impossible or impractical to observe the entire population. For example, we could not test the tensile strength of all the chassis structural elements because it would be too time consuming and expensive. Furthermore, some (perhaps many) of these structural elements do not yet exist at the time a decision is to be made, so to a large extent, we must view the population as **conceptual**. Therefore, we depend on a subset of observations from the population to help make decisions about the population.

### Definition

A **sample** is a subset of observations selected from a population.

For statistical methods to be valid, the sample must be representative of the population. It is often tempting to select the observations that are most convenient as the sample or to exercise judgment in sample selection. These procedures can frequently introduce **bias** into the sample, and as a result the parameter of interest will be consistently underestimated (or overestimated) by such a sample. Furthermore, the behavior of a judgment sample cannot be statistically described. To avoid these difficulties, it is desirable to select a **random sample** as the result of some chance mechanism. Consequently, the selection of a sample is a random experiment and each observation in the sample is the observed value of a random variable. The observations in the population determine the probability distribution of the random variable.

To define a random sample, let  $X$  be a random variable that represents the result of one selection of an observation from the population. Let  $f(x)$  denote the probability density function of  $X$ . Suppose that each observation in the sample is obtained independently, under unchanging conditions. That is, the observations for the sample are obtained by observing  $X$  independently under unchanging conditions, say,  $n$  times. Let  $X_i$  denote the random variable that represents the  $i$ th replicate. Then,  $X_1, X_2, \dots, X_n$  is a random sample and the numerical values obtained are denoted as  $x_1, x_2, \dots, x_n$ . The random variables in a random sample are independent with the same probability distribution  $f(x)$  because of the identical conditions under which each observation is obtained. That is, the marginal probability density function of  $X_1, X_2, \dots, X_n$  is



$f(x_1), f(x_2), \dots, f(x_n)$ , respectively, and by independence the joint probability density function of the random sample is  $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = f(x_1)f(x_2) \dots f(x_n)$ .

### Definition

The random variables  $X_1, X_2, \dots, X_n$  are a random sample of size  $n$  if (a) the  $X_i$ 's are independent random variables, and (b) every  $X_i$  has the same probability distribution.

To illustrate this definition, suppose that we are investigating the effective service life of an electronic component used in a cardiac pacemaker and that component life is normally distributed. Then we would expect each of the observations on component life  $X_1, X_2, \dots, X_n$  in a random sample of  $n$  components to be independent random variables with exactly the same normal distribution. After the data are collected, the numerical values of the observed lifetimes are denoted as  $x_1, x_2, \dots, x_n$ .

The primary purpose in taking a random sample is to obtain information about the unknown population parameters. Suppose, for example, that we wish to reach a conclusion about the proportion of people in the United States who prefer a particular brand of soft drink. Let  $p$  represent the unknown value of this proportion. It is impractical to question every individual in the population to determine the true value of  $p$ . In order to make an inference regarding the true proportion  $p$ , a more reasonable procedure would be to select a random sample (of an appropriate size) and use the observed proportion  $\hat{p}$  of people in this sample favoring the brand of soft drink.

The sample proportion,  $\hat{p}$  is computed by dividing the number of individuals in the sample who prefer the brand of soft drink by the total sample size  $n$ . Thus,  $\hat{p}$  is a function of the observed values in the random sample. Since many random samples are possible from a population, the value of  $\hat{p}$  will vary from sample to sample. That is,  $\hat{p}$  is a random variable. Such a random variable is called a **statistic**.

### Definition

A **statistic** is any function of the observations in a random sample.

We have encountered statistics before. For example, if  $X, X_2, \dots, X_n$  is a random sample of size  $n$ , the **sample mean**  $\bar{X}$ , the **sample variance**  $S^2$ , and the **sample standard deviation**  $S$  are statistics.

Although numerical summary statistics are very useful, **graphical displays** of sample data are a very powerful and extremely useful way to visually examine the data. We now present a few of the techniques that are most relevant to engineering applications of probability and statistics.

## 6-3 STEM-AND-LEAF DIAGRAMS

The dot diagram is a useful data display for small samples, up to (say) about 20 observations. However, when the number of observations is moderately large, other graphical displays may be more useful.

For example, consider the data in Table 6-2. These data are the compressive strengths in pounds per square inch (psi) of 80 specimens of a new aluminum-lithium alloy undergoing evaluation as a possible material for aircraft structural elements. The data were recorded in the order of testing, and in this format they do not convey much information about compressive strength. Questions such as “What percent of the specimens fail below 120 psi?” are not easy to answer.

**Table 6-2** Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Because there are many observations, constructing a dot diagram of these data would be relatively inefficient; more effective displays are available for large data sets.

A **stem-and-leaf diagram** is a good way to obtain an informative visual display of a data set  $x_1, x_2, \dots, x_n$ , where each number  $x_i$  consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps.

**Steps for  
Constructing a Stem-  
and-Leaf Diagram**

- (1) Divide each number  $x_i$  into two parts: a **stem**, consisting of one or more of the leading digits and a **leaf**, consisting of the remaining digit.
- (2) List the stem values in a vertical column.
- (3) Record the leaf for each observation beside its stem.
- (4) Write the units for stems and leaves on the display.

To illustrate, if the data consist of percent defective information between 0 and 100 on lots of semiconductor wafers, we can divide the value 76 into the stem 7 and the leaf 6. In general, we should choose relatively few stems in comparison with the number of observations. It is usually best to choose between 5 and 20 stems.

**EXAMPLE 6-4**

To illustrate the construction of a stem-and-leaf diagram, consider the alloy compressive strength data in Table 6-2. We will select as stem values the numbers 7, 8, 9,  $\dots$ , 24. The resulting stem-and-leaf diagram is presented in Fig. 6-4. The last column in the diagram is a frequency count of the number of leaves associated with each stem. Inspection of this display immediately reveals that most of the compressive strengths lie between 110 and 200 psi and that a central value is somewhere between 150 and 160 psi. Furthermore, the strengths are distributed approximately symmetrically about the central value. The stem-and-leaf diagram enables us to determine quickly some important features of the data that were not immediately obvious in the original display in Table 6-2.

In some data sets, it may be desirable to provide more classes or stems. One way to do this would be to modify the original stems as follows: Divide the stem 5 (say) into two new stems, 5L and 5U. The stem 5L has leaves 0, 1, 2, 3, and 4, and stem 5U has leaves 5, 6, 7, 8, and 9. This will double the number of original stems. We could increase the number of original stems by four by defining five new stems: 5z with leaves 0 and 1, 5t (for twos and three) with leaves 2 and 3, 5f (for fours and fives) with leaves 4 and 5, 5s (for six and seven) with leaves 6 and 7, and 5e with leaves 8 and 9.

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

**Figure 6-4** Stem-and-leaf diagram for the compressive strength data in Table 6-2.

Stem : Tens and hundreds digits (psi); Leaf: Ones digits (psi)

#### EXAMPLE 6-5

Figure 6-5 illustrates the stem-and-leaf diagram for 25 observations on batch yields from a chemical process. In Fig. 6-5(a) we have used 6, 7, 8, and 9 as the stems. This results in too few stems, and the stem-and-leaf diagram does not provide much information about the data. In Fig. 6-5(b) we have divided each stem into two parts, resulting in a display that more

Stem	Leaf	Stem	Leaf	Stem	Leaf
6	1 3 4 5 5 6	6L	1 3 4	6z	1
7	0 1 1 3 5 7 8 8 9	6U	5 5 6	6t	3
8	1 3 4 4 7 8 8	7L	0 1 1 3	6f	4 5 5
9	2 3 5	7U	5 7 8 8 9	6s	6
(a)		8L	1 3 4 4	6e	
		8U	7 8 8	7z	0 1 1
		9L	2 3	7t	3
		9U	5	7f	5
		(b)		7s	7
				7e	8 8 9
				8z	1
				8t	3
				8f	4 4
				8s	7
				8e	8 8
				9z	
				9t	2 3
				9f	5
				9s	
				9e	
				(c)	

**Figure 6-5** Stem-and-leaf displays for Example 6-5. Stem: Tens digits. Leaf: Ones digits.

**Character Stem-and-Leaf Display**

Stem-and-leaf of Strength

N = 80 Leaf Unit = 1.0

1	7	6
2	8	7
3	9	7
5	10	1 5
8	11	0 5 8
11	12	0 1 3
17	13	1 3 3 4 5 5
25	14	1 2 3 5 6 8 9 9
37	15	0 0 1 3 4 4 6 7 8 8 8 8
(10)	16	0 0 0 3 3 5 7 7 8 9
33	17	0 1 1 2 4 4 5 6 6 8
23	18	0 0 1 1 3 4 6
16	19	0 3 4 6 9 9
10	20	0 1 7 8
6	21	8
5	22	1 8 9
2	23	7
1	24	5

**Figure 6-6** A stem-and-leaf diagram from Minitab.

adequately displays the data. Figure 6-5(c) illustrates a stem-and-leaf display with each stem divided into five parts. There are too many stems in this plot, resulting in a display that does not tell us much about the shape of the data.

Figure 6-6 shows a stem-and-leaf display of the compressive strength data in Table 6-2 produced by Minitab. The software uses the same stems as in Fig. 6-4. Note also that the computer orders the leaves from smallest to largest on each stem. This form of the plot is usually called an **ordered stem-and-leaf diagram**. This is not usually done when the plot is constructed manually because it can be time consuming. The computer adds a column to the left of the stems that provides a count of the observations at and above each stem in the upper half of the display and a count of the observations at and below each stem in the lower half of the display. At the middle stem of 16, the column indicates the number of observations at this stem.

The ordered stem-and-leaf display makes it relatively easy to find data features such as percentiles, quartiles, and the median. The sample **median** is a measure of central tendency that divides the data into two equal parts, half below the median and half above. If the number of observations is even, the median is halfway between the two central values. From Fig. 6-6 we find the 40th and 41st values of strength as 160 and 163, so the median is  $(160 + 163)/2 = 161.5$ . If the number of observations is odd, the median is the central value. The sample **mode** is the most frequently occurring data value. Figure 6-6 indicates that the mode is 158; this value occurs four times, and no other value occurs as frequently in the sample.

We can also divide data into more than two parts. When an ordered set of data is divided into four equal parts, the division points are called **quartiles**. The *first* or *lower quartile*,  $q_1$ , is a value that has approximately 25% of the observations below it and approximately 75% of the observations above. The *second quartile*,  $q_2$ , has approximately 50% of the observations below its value. The second quartile is exactly equal to the median. The *third* or *upper quartile*,  $q_3$ , has approximately 75% of the observations below its value. As in the case of the median, the quartiles may not be unique. The compressive strength data in Fig. 6-6 contains  $n = 80$  observations. Minitab software calculates the first and third quartiles as the  $(n + 1)/4$

**Table 6-3** Summary Statistics for the Compressive Strength Data from Minitab

Variable	N	Mean	Median	StDev	SE Mean
	80	162.66	161.50	33.77	3.78
	Min	Max	Q1	Q3	
	76.00	245.00	143.50	181.00	

and  $3(n + 1)/4$  ordered observations and interpolates as needed. For example,  $(80 + 1)/4 = 20.25$  and  $3(80 + 1)/4 = 60.75$ . Therefore, Minitab interpolates between the 20th and 21st ordered observation to obtain  $q_1 = 143.50$  and between the 60th and 61st observation to obtain  $q_3 = 181.00$ . In general, the 100 $k$ th **percentile** is a data value such that approximately 100 $k$ % of the observations are at or below this value and approximately  $100(1 - k)$ % of them are above it. Finally, we may use the **interquartile range**, defined as  $IQR = q_3 - q_1$ , as a measure of variability. The interquartile range is less sensitive to the extreme values in the sample than is the ordinary sample range.

Many statistics software packages provide data summaries that include these quantities. The output obtained for the compressive strength data in Table 6-2 from Minitab is shown in Table 6-3.

### EXERCISES FOR SECTION 6-3

**6-14.** An article in *Technometrics* (Vol. 19, 1977, p. 425) presents the following data on the motor fuel octane ratings of several blends of gasoline:

88.5	98.8	89.6	92.2	92.7	88.4	87.5	90.9
94.7	88.3	90.4	83.4	87.9	92.6	87.8	89.9
84.3	90.4	91.6	91.0	93.0	93.7	88.3	91.8
90.1	91.2	90.7	88.2	94.4	96.5	89.2	89.7
89.0	90.6	88.6	88.5	90.4	84.3	92.3	92.2
89.8	92.2	88.3	93.3	91.2	93.2	88.9	
91.6	87.7	94.2	87.4	86.7	88.6	89.8	
90.3	91.1	85.3	91.1	94.2	88.7	92.7	
90.0	86.7	90.1	90.5	90.8	92.7	93.3	
91.5	93.4	89.3	100.3	90.1	89.3	86.7	
89.9	96.1	91.1	87.6	91.8	91.0	91.0	

Construct a stem-and-leaf display for these data.

**6-15.** The following data are the numbers of cycles to failure of aluminum test coupons subjected to repeated alternating stress at 21,000 psi, 18 cycles per second:

1115	865	1015	885	1594	1000	1416	1501
1310	2130	845	1223	2023	1820	1560	1238
1540	1421	1674	375	1315	1940	1055	990
1502	1109	1016	2265	1269	1120	1764	1468
1258	1481	1102	1910	1260	910	1330	1512
1315	1567	1605	1018	1888	1730	1608	1750
1085	1883	706	1452	1782	1102	1535	1642
798	1203	2215	1890	1522	1578	1781	
1020	1270	785	2100	1792	758	1750	

(a) Construct a stem-and-leaf display for these data.

(b) Does it appear likely that a coupon will “survive” beyond 2000 cycles? Justify your answer.

**6-16.** The percentage of cotton in material used to manufacture men’s shirts follows. Construct a stem-and-leaf display for the data.

34.2	37.8	33.6	32.6	33.8	35.8	34.7	34.6
33.1	36.6	34.7	33.1	34.2	37.6	33.6	33.6
34.5	35.4	35.0	34.6	33.4	37.3	32.5	34.1
35.6	34.6	35.4	35.9	34.7	34.6	34.1	34.7
36.3	33.8	36.2	34.7	34.6	35.5	35.1	35.7
35.1	37.1	36.8	33.6	35.2	32.8	36.8	36.8
34.7	34.0	35.1	32.9	35.0	32.1	37.9	34.3
33.6	34.1	35.3	33.5	34.9	34.5	36.4	32.7

**6-17.** The following data represent the yield on 90 consecutive batches of ceramic substrate to which a metal coating has been applied by a vapor-deposition process. Construct a stem-and-leaf display for these data.

94.1	86.1	95.3	84.9	88.8	84.6	94.4	84.1
93.2	90.4	94.1	78.3	86.4	83.6	96.1	83.7
90.6	89.1	97.8	89.6	85.1	85.4	98.0	82.9
91.4	87.3	93.1	90.3	84.0	89.7	85.4	87.3
88.2	84.1	86.4	93.1	93.7	87.6	86.6	86.4
86.1	90.1	87.6	94.6	87.7	85.1	91.7	84.5
95.1	95.2	94.1	96.3	90.6	89.6	87.5	
90.0	86.1	92.1	94.7	89.4	90.0	84.2	
92.4	94.3	96.4	91.1	88.6	90.1	85.1	
87.3	93.2	88.2	92.4	84.1	94.3	90.5	
86.6	86.7	86.4	90.6	82.6	97.3	95.6	
91.2	83.0	85.0	89.1	83.1	96.8	88.3	



**6-18.** Find the median and the quartiles for the motor fuel octane data in Exercise 6-14.



**6-19.** Find the median and the quartiles for the failure data in Exercise 6-15.



**6-20.** Find the median, mode, and sample average of the data in Exercise 6-16. Explain how these three measures of location describe different features in the data.



**6-21.** Find the median and the quartiles for the yield data in Exercise 6-19.

**6-22.** The female students in an undergraduate engineering core course at ASU self-reported their heights to the nearest inch. The data are

62 64 66 67 65 68 61 65 67 65 64 63 67  
68 64 66 68 69 65 67 62 66 68 67 66 65  
69 65 70 65 67 68 65 63 64 67 67

- Calculate the sample mean and standard deviation of height.
- Construct a stem-and-leaf diagram for the height data and comment on any important features that you notice.
- What is the median height of this group of female engineering students?

**6-23.** The shear strengths of 100 spot welds in a titanium alloy follow. Construct a stem-and-leaf diagram for the weld strength data and comment on any important features that you notice.

5408 5431 5475 5442 5376 5388 5459 5422 5416 5435  
5420 5429 5401 5446 5487 5416 5382 5357 5388 5457  
5407 5469 5416 5377 5454 5375 5409 5459 5445 5429  
5463 5408 5481 5453 5422 5354 5421 5406 5444 5466  
5399 5391 5477 5447 5329 5473 5423 5441 5412 5384  
5445 5436 5454 5453 5428 5418 5465 5427 5421 5396  
5381 5425 5388 5388 5378 5481 5387 5440 5482 5406  
5401 5411 5399 5431 5440 5413 5406 5342 5452 5420  
5458 5485 5431 5416 5431 5390 5399 5435 5387 5462  
5383 5401 5407 5385 5440 5422 5448 5366 5430 5418

**6-24.** An important quality characteristic of water is the concentration of suspended solid material. Following are 60 measurements on suspended solids from a certain lake. Construct a stem-and-leaf diagram for this data and comment on any important features that you notice. Compute the sample mean, sample standard deviation, and the sample median.

42.4 65.7 29.8 58.7 52.1 55.8 57.0 68.7 67.3 67.3  
54.3 54.0 73.1 81.3 59.9 56.9 62.2 69.9 66.9 59.0  
56.3 43.3 57.4 45.3 80.1 49.7 42.8 42.4 59.6 65.8  
61.4 64.0 64.2 72.6 72.5 46.1 53.1 56.1 67.2 70.7  
42.6 77.4 54.7 57.1 77.3 39.3 76.4 59.3 51.1 73.8  
61.4 73.1 77.3 48.5 89.8 50.7 52.0 59.6 66.1 31.6

**6-25.** The United States Golf Association tests golf balls to ensure that they conform to the rules of golf. Balls are tested for weight, diameter, roundness, and overall distance. The overall

distance test is conducted by hitting balls with a driver swung by a mechanical device nicknamed “Iron Byron” after the legendary great Byron Nelson, whose swing the machine is said to emulate. Following are 100 distances (in yards) achieved by a particular brand of golf ball in the overall distance test. Construct a stem-and-leaf diagram for this data and comment on any important features that you notice. Compute the sample mean, sample standard deviation, and the sample median.

261.3 259.4 265.7 270.6 274.2 261.4 254.5 283.7  
258.1 270.5 255.1 268.9 267.4 253.6 234.3 263.2  
254.2 270.7 233.7 263.5 244.5 251.8 259.5 257.5  
257.7 272.6 253.7 262.2 252.0 280.3 274.9 233.7  
237.9 274.0 264.5 244.8 264.0 268.3 272.1 260.2  
255.8 260.7 245.5 279.6 237.8 278.5 273.3 263.7  
241.4 260.6 280.3 272.7 261.0 260.0 279.3 252.1  
244.3 272.2 248.3 278.7 236.0 271.2 279.8 245.6  
241.2 251.1 267.0 273.4 247.7 254.8 272.8 270.5  
254.4 232.1 271.5 242.9 273.6 256.1 251.6  
256.8 273.0 240.8 276.6 264.5 264.5 226.8  
255.3 266.6 250.2 255.8 285.3 255.4 240.5  
255.0 273.2 251.4 276.1 277.8 266.8 268.5

**6-26.** A semiconductor manufacturer produces devices used as central processing units in personal computers. The speed of the device (in megahertz) is important because it determines the price that the manufacturer can charge for the devices. The following table contains measurements on 120 devices. Construct a stem-and-leaf diagram for this data and comment on any important features that you notice. Compute the sample mean, sample standard deviation, and the sample median. What percentage of the devices has a speed exceeding 700 megahertz?

680 669 719 699 670 710 722 663 658 634 720 690  
677 669 700 718 690 681 702 696 692 690 694 660  
649 675 701 721 683 735 688 763 672 698 659 704  
681 679 691 683 705 746 706 649 668 672 690 724  
652 720 660 695 701 724 668 698 668 660 680 739  
717 727 653 637 660 693 679 682 724 642 704 695  
704 652 664 702 661 720 695 670 656 718 660 648  
683 723 710 680 684 705 681 748 697 703 660 722  
662 644 683 695 678 674 656 667 683 691 680 685  
681 715 665 676 665 675 655 659 720 675 697 663

**6-27.** A group of wine enthusiasts taste-tested a pinot noir wine from Oregon. The evaluation was to grade the wine on a 0 to 100 point scale. The results follow:

94 90 92 91 91 86 89 91 91 90  
90 93 87 90 91 92 89 86 89 90  
88 95 91 88 89 92 87 89 95 92  
85 91 85 89 88 84 85 90 90 83

- Construct a stem-and-leaf diagram for this data and comment on any important features that you notice.

- (b) Compute the sample mean, sample standard deviation, and the sample median.
- (c) A wine rated above 90 is considered truly exceptional. What proportion of the taste-tasters considered this particular pinot noir truly exceptional?

**6-28.** In their book *Introduction to Linear Regression Analysis* (3rd edition, Wiley, 2001) Montgomery, Peck, and Vining present measurements on  $\text{NbOCl}_3$  concentration from a tube-flow reactor experiment. The data, in gram – mole per liter  $\times 10^{-3}$ , are as follows:

450 450 473 507 457 452 453 1215 1256  
1145 1085 1066 1111 1364 1254 1396 1575 1617  
1733 2753 3186 3227 3469 1911 2588 2635 2725

- (a) Construct a stem-and-leaf diagram for this data and comment on any important features that you notice.

- (b) Compute the sample mean, sample standard deviation, and the sample median.

**6-29. A Comparative Stem-and-Leaf Diagram.** In Exercise 6-22, we presented height data that was self-reported by female undergraduate engineering students in a core course at ASU. In the same class, the male students self-reported their heights as follows:

69 67 69 70 65 68 69 70 71 69 66 67 69 75 68 67 68  
69 70 71 72 68 69 69 70 71 68 72 69 69 68 69 73 70  
73 68 69 71 67 68 65 68 68 69 70 74 71 69 70 69

- (a) Construct a comparative stem-and-leaf diagram by listing the stems in the center of the display and then placing the female leaves on the left and the male leaves on the right.
- (b) Comment on any important features that you notice in this display.

## 6-4 FREQUENCY DISTRIBUTIONS AND HISTOGRAMS

A **frequency distribution** is a more compact summary of data than a stem-and-leaf diagram. To construct a frequency distribution, we must divide the range of the data into intervals, which are usually called **class intervals**, **cells**, or **bins**. If possible, the bins should be of equal width in order to enhance the visual information in the frequency distribution. Some judgment must be used in selecting the number of bins so that a reasonable display can be developed. The number of bins depends on the number of observations and the amount of scatter or dispersion in the data. A frequency distribution that uses either too few or too many bins will not be informative. We usually find that between 5 and 20 bins is satisfactory in most cases and that the number of bins should increase with  $n$ . Choosing the **number of bins** approximately equal to the square root of the number of observations often works well in practice.

A frequency distribution for the comprehensive strength data in Table 6-2 is shown in Table 6-4. Since the data set contains 80 observations, and since  $\sqrt{80} \approx 9$ , we suspect that about eight to nine bins will provide a satisfactory frequency distribution. The largest and smallest data values are 245 and 76, respectively, so the bins must cover a range of at least  $245 - 76 = 169$  units on the psi scale. If we want the lower limit for the first bin to begin slightly below the smallest data value and the upper limit for the last bin to be slightly above the largest data value, we might start the frequency distribution at 70 and end it at 250. This is an interval or range of 180 psi units. Nine bins, each of width 20 psi, give a reasonable frequency distribution, so the frequency distribution in Table 6-4 is based on nine bins.

The second row of Table 6-4 contains a **relative frequency distribution**. The relative frequencies are found by dividing the observed frequency in each bin by the total number of

**Table 6-4** Frequency Distribution for the Compressive Strength Data in Table 6-2

Class	$70 \leq x < 90$	$90 \leq x < 110$	$110 \leq x < 130$	$130 \leq x < 150$	$150 \leq x < 170$	$170 \leq x < 190$	$190 \leq x < 210$	$210 \leq x < 230$	$230 \leq x < 250$
Frequency	2	3	6	14	22	17	10	4	2
Relative frequency	0.0250	0.0375	0.0750	0.1750	0.2750	0.2125	0.1250	0.0500	0.0250
Cumulative relative frequency	0.0250	0.0625	0.1375	0.3125	0.5875	0.8000	0.9250	0.9750	1.0000



observations. The last row of Table 6-4 expresses the relative frequencies on a cumulative basis. Frequency distributions are often easier to interpret than tables of data. For example, from Table 6-4 it is very easy to see that most of the specimens have compressive strengths between 130 and 190 psi and that 97.5 percent of the specimens fail below 230 psi.

The **histogram** is a visual display of the frequency distribution. The stages for constructing a histogram follow.

**Constructing a  
Histogram (Equal  
Bin Widths)**

- (1) Label the bin (class interval) boundaries on a horizontal scale.
- (2) Mark and label the vertical scale with the frequencies or the relative frequencies.
- (3) Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.

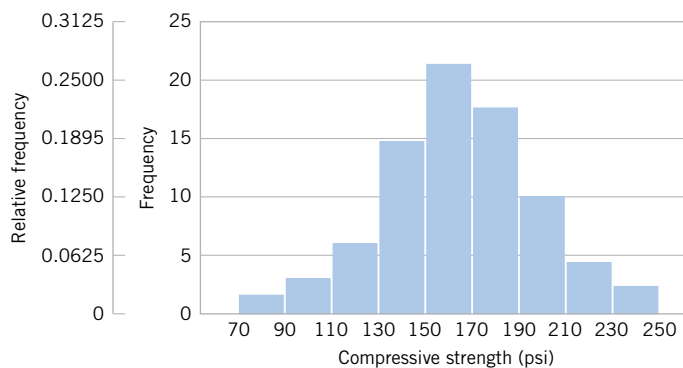
Figure 6-7 is the histogram for the compression strength data. The histogram, like the **stem-and-leaf diagram**, provides a visual impression of the shape of the distribution of the measurements and information about the central tendency and scatter or dispersion in the data. Notice the symmetric, bell-shaped distribution of the strength measurements in Fig. 6-7. This display often gives insight about possible choices of probability distribution to use as a model for the population. For example, here we would likely conclude that the **normal distribution** is a reasonable model for the population of compression strength measurements.

Sometimes a histogram with **unequal bin widths** will be employed. For example, if the data have several extreme observations or outliers, using a few equal-width bins will result in nearly all observations falling in just a few of the bins. Using many equal-width bins will result in many bins with zero frequency. A better choice is to use shorter intervals in the region where most of the data falls and a few wide intervals near the extreme observations. When the bins are of unequal width, the rectangle's **area** (not its height) should be proportional to the bin frequency. This implies that the rectangle height should be

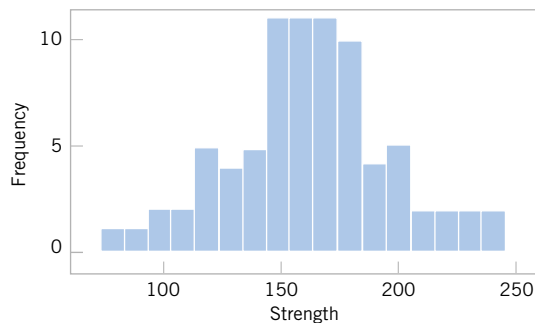
$$\text{Rectangle height} = \frac{\text{bin frequency}}{\text{bin width}}$$

In passing from either the original data or stem-and-leaf diagram to a frequency distribution or histogram, we have lost some information because we no longer have the individual observations. However, this information loss is often small compared with the conciseness and ease of interpretation gained in using the frequency distribution and histogram.

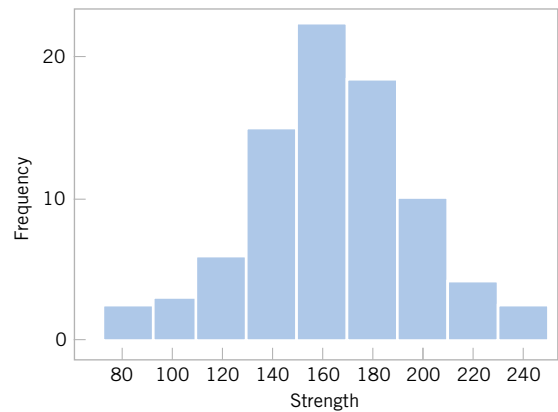
**Figure 6-7** Histogram of compressive strength for 80 aluminum-lithium alloy specimens.







**Figure 6-8** A histogram of the compressive strength data from Minitab with 17 bins.

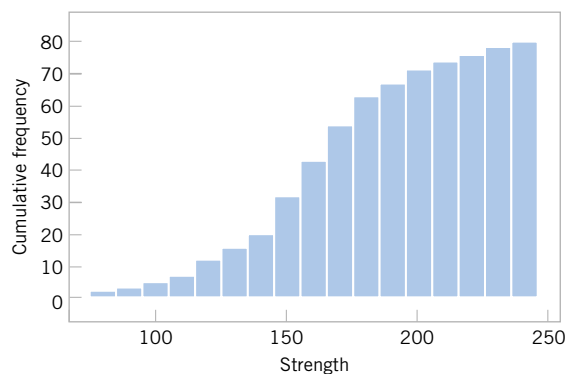


**Figure 6-9** A histogram of the compressive strength data from Minitab with nine bins.

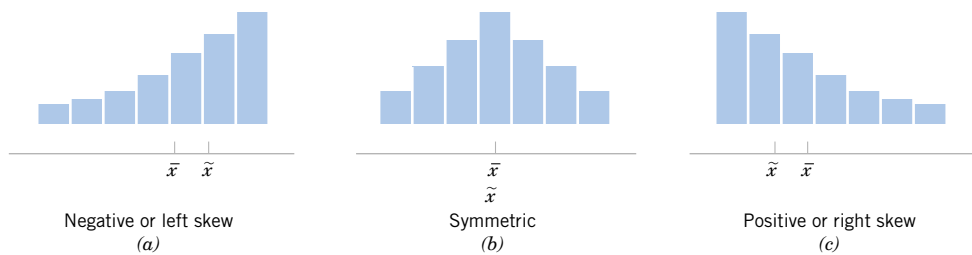
Figure 6-8 shows a histogram of the compressive strength data from Minitab. The “default” settings were used in this histogram, leading to 17 bins. We have noted that histograms may be relatively sensitive to the number of bins and their width. For small data sets, histograms may change dramatically in appearance if the number and/or width of the bins changes. Histograms are more stable for larger data sets, preferably of size 75 to 100 or more. Figure 6-9 shows the Minitab histogram for the compressive strength data with nine bins. This is similar to the original histogram shown in Fig. 6-7. Since the number of observations is moderately large ( $n = 80$ ), the choice of the number of bins is not especially important, and both Figs. 6-8 and 6-9 convey similar information.

Figure 6-10 shows a variation of the histogram available in Minitab, the **cumulative frequency plot**. In this plot, the height of each bar is the total number of observations that are less than or equal to the upper limit of the bin. Cumulative distributions are also useful in data interpretation; for example, we can read directly from Fig. 6-10 that there are approximately 70 observations less than or equal to 200 psi.

When the sample size is large, the histogram can provide a reasonably reliable indicator of the general **shape** of the distribution or population of measurements from which the sample was drawn. Figure 6-11 presents three cases. The median is denoted as  $\tilde{x}$ . Generally, if the data are symmetric, as in Fig. 6-11(b), the mean and median coincide. If, in addition, the data have only one mode (we say the data are *unimodal*), the mean, median, and mode all coincide. If the data are *skewed* (asymmetric, with a long tail to one side), as in Fig. 6-11(a) and (c), the mean, median, and mode do not coincide. Usually, we find that  $\text{mode} < \text{median} < \text{mean}$  if the



**Figure 6-10** A cumulative distribution plot of the compressive strength data from Minitab.



**Figure 6-11**  
Histograms for symmetric and skewed distributions.

distribution is skewed to the right, whereas  $\text{mode} > \text{median} > \text{mean}$  if the distribution is skewed to the left.

Frequency distributions and histograms can also be used with qualitative or categorical data. In some applications there will be a natural ordering of the categories (such as freshman, sophomore, junior, and senior), whereas in others the order of the categories will be arbitrary (such as male and female). When using categorical data, the bins should have equal width.

**EXAMPLE 6-6** Figure 6-12 presents the production of transport aircraft by the Boeing Company in 1985. Notice that the 737 was the most popular model, followed by the 757, 747, 767, and 707.

A chart of occurrences by category (in which the categories are ordered by the number of occurrences) is sometimes referred to as a **Pareto chart**. See Exercise 6-41.

In this section we have concentrated on descriptive methods for the situation in which each observation in a data set is a single number or belongs to one category. In many cases, we work with data in which each observation consists of several measurements. For example, in a gasoline mileage study, each observation might consist of a measurement of miles per gallon, the size of the engine in the vehicle, engine horsepower, vehicle weight, and vehicle length. This is an example of **multivariate data**. In later chapters, we will discuss analyzing this type of data.

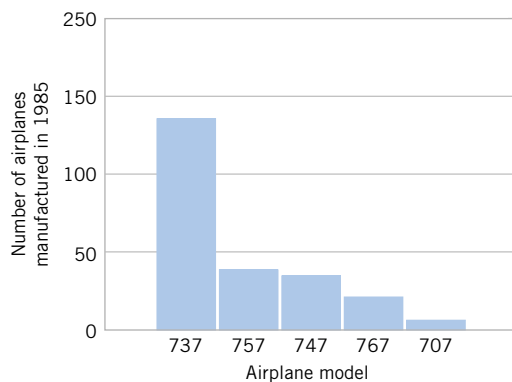
### EXERCISES FOR SECTION 6-4

- 6-30.** Construct a frequency distribution and histogram for the motor fuel octane data from Exercise 6-14. Use eight bins.






**6-31.** Construct a frequency distribution and histogram using the failure data from Exercise 6-15.

**6-32.** Construct a frequency distribution and histogram for the cotton content data in Exercise 6-16.

**6-33.** Construct a frequency distribution and histogram for the yield data in Exercise 6-17.



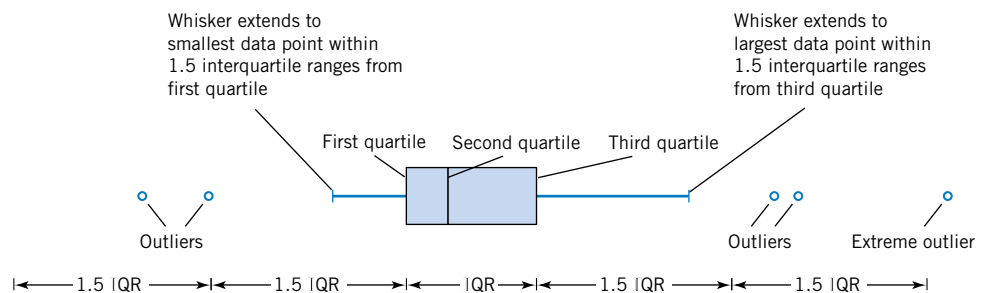
**Figure 6-12**  
Airplane production in 1985. (Source: Boeing Company.)

-  **6-34.** Construct a frequency distribution and histogram with 16 bins for the motor fuel octane data in Exercise 6-14. Compare its shape with that of the histogram with eight bins from Exercise 6-30. Do both histograms display similar information?
- 6-35.** Construct a histogram for the female student height data in Exercise 6-22.
-  **6-36.** Construct a histogram with 10 bins for the spot weld shear strength data in Exercise 6-23. Comment on the shape of the histogram. Does it convey the same information as the stem-and-leaf display?
-  **6-37.** Construct a histogram for the water quality data in Exercise 6-24. Comment on the shape of the histogram. Does it convey the same information as the stem-and-leaf display?
-  **6-38.** Construct a histogram with 10 bins for the overall distance data in Exercise 6-25. Comment on the shape of the histogram. Does it convey the same information as the stem-and-leaf display?
-  **6-39.** Construct a histogram for the semiconductor speed data in Exercise 6-26. Comment on the shape of the histogram. Does it convey the same information as the stem-and-leaf display?
- 6-40.** Construct a histogram for the pinot noir wine rating data in Exercise 6-27. Comment on the shape of the histogram. Does it convey the same information as the stem-and-leaf display?
- 6-41. The Pareto Chart.** An important variation of a histogram for categorical data is the Pareto chart. This chart is widely used in quality improvement efforts, and the categories usually represent different types of defects, failure modes, or product/process problems. The categories are ordered so that the category with the largest frequency is on the left, followed by the category with the second largest frequency and so forth. These charts are named after the Italian economist V. Pareto, and they usually exhibit “Pareto’s law”; that is, most of the defects can be accounted for by only a few categories. Suppose that the following information on structural defects in automobile doors is obtained: dents, 4; pits, 4; parts assembled out of sequence, 6; parts undertrimmed, 21; missing holes/slots, 8; parts not lubricated, 5; parts out of contour, 30; and parts not deburred, 3. Construct and interpret a Pareto chart.

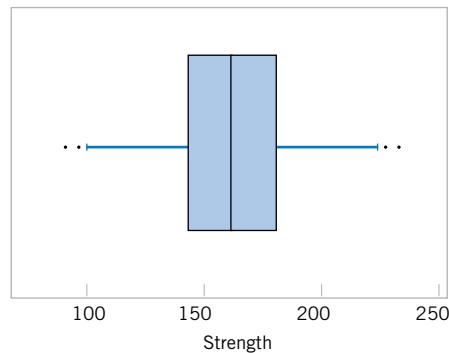
## 6-5 BOX PLOTS

The stem-and-leaf display and the histogram provide general visual impressions about a data set, while numerical quantities such as  $\bar{x}$  or  $s$  provide information about only one feature of the data. The **box plot** is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry, and identification of unusual observations or outliers.

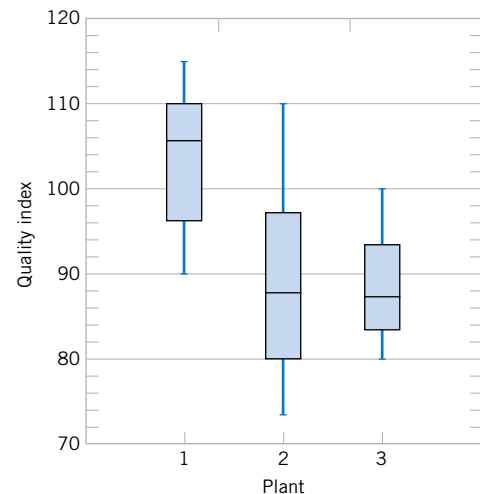
A box plot displays the three quartiles, the minimum, and the maximum of the data on a rectangular box, aligned either horizontally or vertically. The box encloses the interquartile range with the left (or lower) edge at the first quartile,  $q_1$ , and the right (or upper) edge at the third quartile,  $q_3$ . A line is drawn through the box at the second quartile (which is the 50th percentile or the median),  $q_2 = \bar{x}$ . A line, or **whisker**, extends from each end of the box. The lower whisker is a line from the first quartile to the smallest data point within 1.5 interquartile ranges from the first quartile. The upper whisker is a line from the third quartile to the largest data point within 1.5 interquartile ranges from the third quartile. Data farther from the box than the whiskers are plotted as individual points. A point beyond a whisker, but less than 3 interquartile ranges from the box edge, is called an **outlier**. A point more than 3 interquartile ranges from the box edge is called an **extreme outlier**. See Fig. 6-13. Occasionally, different symbols, such as open and filled circles, are used to identify the two types of outliers. Sometimes box plots are called *box-and-whisker plots*.



**Figure 6-13** Description of a box plot.



**Figure 6-14** Box plot for compressive strength data in Table 6-2.



**Figure 6-15** Comparative box plots of a quality index at three plants.

Figure 6-14 presents the box plot from Minitab for the alloy compressive strength data shown in Table 6-2. This box plot indicates that the distribution of compressive strengths is fairly symmetric around the central value, because the left and right whiskers and the lengths of the left and right boxes around the median are about the same. There are also two mild outliers on either end of the data.

Box plots are very useful in graphical comparisons among data sets, because they have high visual impact and are easy to understand. For example, Fig. 6-15 shows the comparative box plots for a manufacturing quality index on semiconductor devices at three manufacturing plants. Inspection of this display reveals that there is too much variability at plant 2 and that plants 2 and 3 need to raise their quality index performance.

## EXERCISES FOR SECTION 6-5

**6-42.** Exercise 6-13 presented the joint temperatures of the O-rings ( $^{\circ}\text{F}$ ) for each test firing or actual launch of the space shuttle rocket motor. In that exercise you were asked to find the sample mean and sample standard deviation of temperature.

- Find the upper and lower quartiles of temperature.
- Find the median.
- Set aside the smallest observation ( $31^{\circ}\text{F}$ ) and recompute the quantities in parts (a) and (b). Comment on your findings. How “different” are the other temperatures from this smallest value?
- Construct a box plot of the data and comment on the possible presence of outliers.

**6-43.** An article in the *Transactions of the Institution of Chemical Engineers* (Vol. 34, 1956, pp. 280–293) reported data from an experiment investigating the effect of several

process variables on the vapor phase oxidation of naphthalene. A sample of the percentage mole conversion of naphthalene to maleic anhydride follows: 4.2, 4.7, 4.7, 5.0, 3.8, 3.6, 3.0, 5.1, 3.1, 3.8, 4.8, 4.0, 5.2, 4.3, 2.8, 2.0, 2.8, 3.3, 4.8, 5.0.

- Calculate the sample mean.
- Calculate the sample variance and sample standard deviation.
- Construct a box plot of the data.

**6-44.** The “cold start ignition time” of an automobile engine is being investigated by a gasoline manufacturer. The following times (in seconds) were obtained for a test vehicle: 1.75, 1.92, 2.62, 2.35, 3.09, 3.15, 2.53, 1.91.

- Calculate the sample mean and sample standard deviation.
- Construct a box plot of the data.

**6-45.** The nine measurements that follow are furnace temperatures recorded on successive batches in a semiconductor

manufacturing process (units are °F): 953, 950, 948, 955, 951, 949, 957, 954, 955.

- Calculate the sample mean, sample variance, and standard deviation.
- Find the median. How much could the largest temperature measurement increase without changing the median value?
- Construct a box plot of the data.

**6-46.** Exercise 6-12 presents drag coefficients for the NASA 0012 airfoil. You were asked to calculate the sample mean, sample variance, and sample standard deviation of those coefficients.

- Find the upper and lower quartiles of the drag coefficients.
- Construct a box plot of the data.
- Set aside the largest observation (100) and rework parts a and b. Comment on your findings.

**6-47.** The following data are the temperatures of effluent at discharge from a sewage treatment facility on consecutive days:

43	47	51	48	52	50	46	49
45	52	46	51	44	49	46	51
49	45	44	50	48	50	49	50

- Calculate the sample mean and median.
- Calculate the sample variance and sample standard deviation.
- Construct a box plot of the data and comment on the information in this display.

**6-48.** Reconsider the golf course yardage data in Exercise 6-3. Construct a box plot of the yardages and write an interpretation of the plot.

**6-49.** Reconsider the motor fuel octane rating data in Exercise 6-14. Construct a box plot of the yardages and write an interpretation of the plot. How does the box plot compare in interpretive value to the original stem-and-leaf diagram in Exercise 6-14?

**6-50.** Reconsider the spot weld shear strength data in Exercise 6-23. Construct a box plot of the strengths and write an interpretation of the plot. How does the box plot compare in interpretive value to the original stem-and-leaf diagram in Exercise 6-23?

**6-51.** Reconsider the female engineering student height data in Exercise 6-22. Construct a box plot of the heights and write an interpretation of the plot. How does the box plot compare in interpretive value to the original stem-and-leaf diagram in Exercise 6-22?

**6-52.** Reconsider the water quality data in Exercise 6-24. Construct a box plot of the concentrations and write an interpretation of the plot. How does the box plot compare in interpretive value to the original stem-and-leaf diagram in Exercise 6-24?

**6-53.** Reconsider the golf ball overall distance data in Exercise 6-25. Construct a box plot of the yardage distance and write an interpretation of the plot. How does the box plot compare in interpretive value to the original stem-and-leaf diagram in Exercise 6-25?

**6-54.** Reconsider the wine rating data in Exercise 6-27. Construct a box plot of the wine ratings and write an interpretation of the plot. How does the box plot compare in interpretive value to the original stem-and-leaf diagram in Exercise 6-27?

**6-55.** Use the data on heights of female and male engineering students from Exercises 6-22 and 6-29 to construct comparative box plots. Write an interpretation of the information that you see in these plots.

**6-56.** In Exercise 6-44, data was presented on the cold start ignition time of a particular gasoline used in a test vehicle. A second formulation of the gasoline was tested in the same vehicle, with the following times (in seconds): 1.83, 1.99, 3.13, 3.29, 2.65, 2.87, 3.40, 2.46, 1.89, and 3.35. Use this new data along with the cold start times reported in Exercise 6-44 to construct comparative box plots. Write an interpretation of the information that you see in these plots.

## 6-6 TIME SEQUENCE PLOTS

The graphical displays that we have considered thus far such as histograms, stem-and-leaf plots, and box plots are very useful visual methods for showing the variability in data. However, we noted in Section 1-2.2 that time is an important factor that contributes to variability in data, and those graphical methods do not take this into account. A **time series** or **time sequence** is a data set in which the observations are recorded in the order in which they occur. A **time series plot** is a graph in which the vertical axis denotes the observed value of the variable (say  $x$ ) and the horizontal axis denotes the time (which could be minutes, days, years, etc.) When measurements are plotted as a time series, we often see trends, cycles, or other broad features of the data that could not be seen otherwise.

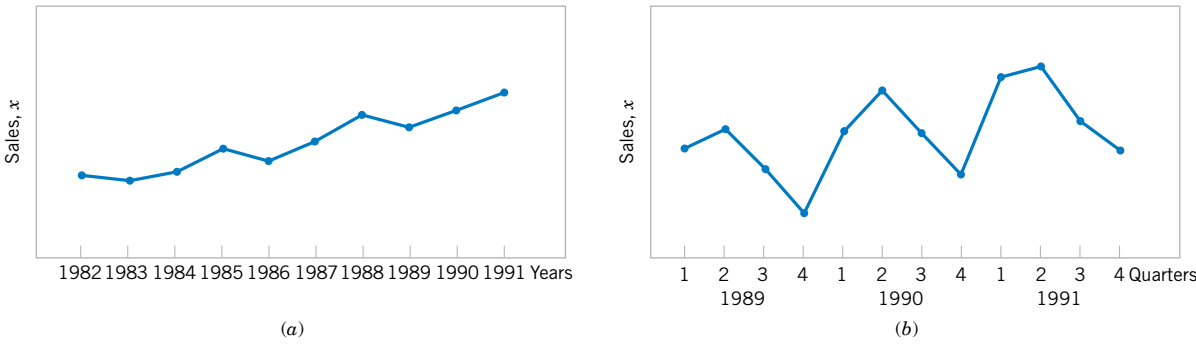


Figure 6-16 Company sales by year (a) and by quarter (b).

For example, consider Fig. 6-16(a), which presents a time series plot of the annual sales of a company for the last 10 years. The general impression from this display is that sales show an upward **trend**. There is some variability about this trend, with some years' sales increasing over those of the last year and some years' sales decreasing. Figure 6-16(b) shows the last three years of sales reported by quarter. This plot clearly shows that the annual sales in this business exhibit a **cyclic** variability by quarter, with the first- and second-quarter sales being generally greater than sales during the third and fourth quarters.

Sometimes it can be very helpful to combine a time series plot with some of the other graphical displays that we have considered previously. J. Stuart Hunter (*The American Statistician*, Vol. 42, 1988, p. 54) has suggested combining the stem-and-leaf plot with a time series plot to form a **digidot plot**.

Figure 6-17 shows a digidot plot for the observations on compressive strength from Table 6-2, assuming that these observations are recorded in the order in which they

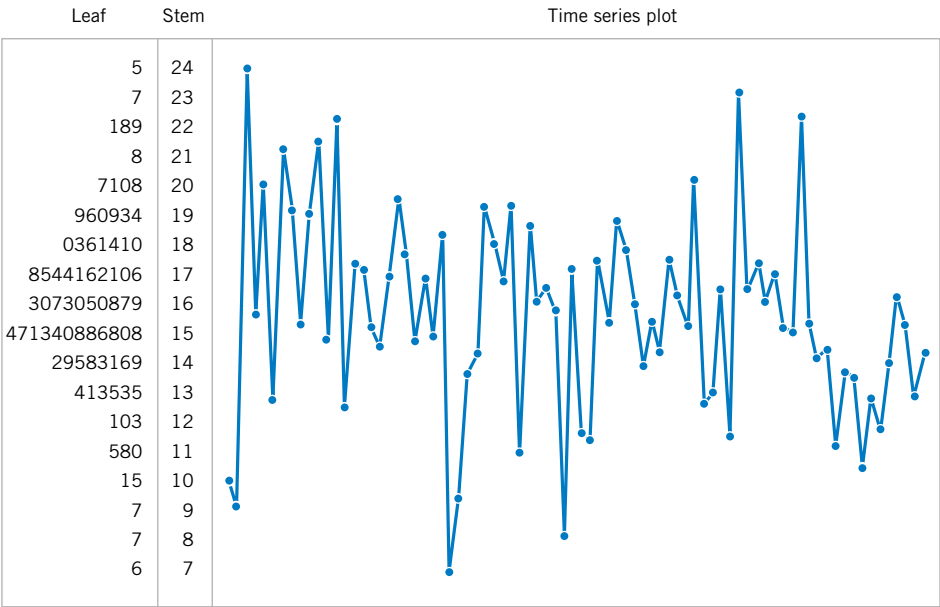
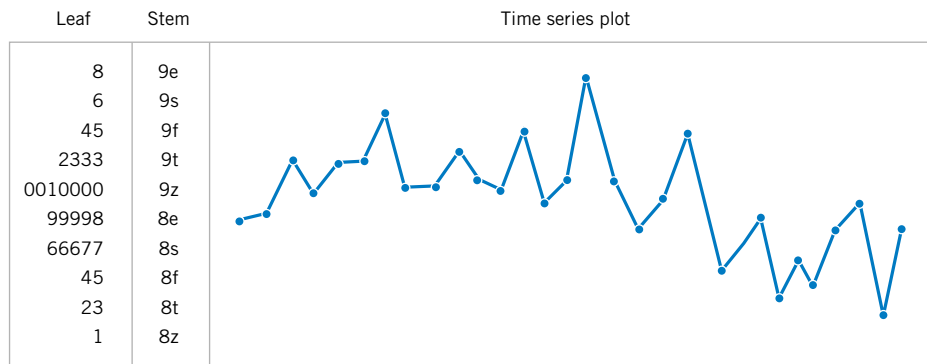


Figure 6-17 A digidot plot of the compressive strength data in Table 6-2.

**Figure 6-18** A digidot plot of chemical process concentration readings, observed hourly.



occurred. This plot effectively displays the overall variability in the compressive strength data and simultaneously shows the variability in these measurements over time. The general impression is that compressive strength varies around the mean value of 162.67, and there is no strong obvious pattern in this variability over time.

The digidot plot in Fig. 6-18 tells a different story. This plot summarizes 30 observations on concentration of the output product from a chemical process, where the observations are recorded at one-hour time intervals. This plot indicates that during the first 20 hours of operation this process produced concentrations generally above 85 grams per liter, but that following sample 20, something may have occurred in the process that results in lower concentrations. If this variability in output product concentration can be reduced, operation of this process can be improved.

## EXERCISES FOR SECTION 6-6

**6-57.** The College of Engineering and Applied Science at Arizona State University had a VAX computer system. Response times for 20 consecutive jobs were recorded and are as follows: (read across)

5.3	10.1	5.9	12.2	11.2	12.4	9.2
5.0	5.8	7.2	8.5	7.3	3.9	10.5
9.5	6.2	10.0	4.7	6.4	8.1	

Construct and interpret a time series plot of these data.

**6-58.** The following data are the viscosity measurements for a chemical product observed hourly (read down, then left to right).

47.9	48.6	48.0	48.1	43.0	43.2
47.9	48.8	47.5	48.0	42.9	43.6
48.6	48.1	48.6	48.3	43.6	43.2
48.0	48.3	48.0	43.2	43.3	43.5
48.4	47.2	47.9	43.0	43.0	43.0
48.1	48.9	48.3	43.5	42.8	
48.0	48.6	48.5	43.1	43.1	

(a) Construct and interpret either a digidot plot or a separate stem-and-leaf and time series plot of these data.

(b) Specifications on product viscosity are at  $48 \pm 2$ . What conclusions can you make about process performance?

**6-59.** The pull-off force for a connector is measured in a laboratory test. Data for 40 test specimens follow (read down, then left to right).

241	203	201	251	236	190
258	195	195	238	245	175
237	249	255	210	209	178
210	220	245	198	212	175
194	194	235	199	185	190
225	245	220	183	187	
248	209	249	213	218	

(a) Construct a time series plot of the data.

(b) Construct and interpret either a digidot plot or a stem-and-leaf plot of the data.

**6-60.** In their book *Time Series Analysis, Forecasting, and Control* (Prentice Hall, 1994), G. E. P. Box, G. M. Jenkins, and G. C. Reinsel present chemical process concentration readings made every two hours. Some of these data follow (read down, then left to right).

**Table 6-5** United Kingdom Passenger Airline Miles Flown

Month	1964	1965	1966	1967	1968	1969	1970
Jan.	7.269	8.350	8.186	8.334	8.639	9.491	10.840
Feb.	6.775	7.829	7.444	7.899	8.772	8.919	10.436
Mar.	7.819	8.829	8.484	9.994	10.894	11.607	13.589
Apr.	8.371	9.948	9.864	10.078	10.455	8.852	13.402
May	9.069	10.638	10.252	10.801	11.179	12.537	13.103
June	10.248	11.253	12.282	12.953	10.588	14.759	14.933
July	11.030	11.424	11.637	12.222	10.794	13.667	14.147
Aug.	10.882	11.391	11.577	12.246	12.770	13.731	14.057
Sept.	10.333	10.665	12.417	13.281	13.812	15.110	16.234
Oct.	9.109	9.396	9.637	10.366	10.857	12.185	12.389
Nov.	7.685	7.775	8.094	8.730	9.290	10.645	11.594
Dec.	7.682	7.933	9.280	9.614	10.925	12.161	12.772

17.0	16.7	17.1	17.5	17.6	41	10	16	8	62	94
16.6	17.4	17.4	18.1	17.5	21	8	7	13	98	96
16.3	17.2	17.4	17.5	16.5	16	2	4	57	124	77
16.1	17.4	17.5	17.4	17.8	6	0	2	122	96	59
17.1	17.4	17.4	17.4	17.3	4	1	8	138	66	44
16.9	17.0	17.6	17.1	17.3	7	5	17	103	64	47
16.8	17.3	17.4	17.6	17.1	14	12	36	86	54	30
17.4	17.2	17.3	17.7	17.4	34	14	50	63	39	16
17.1	17.4	17.0	17.4	16.9	45	35	62	37	21	7
17.0	16.8	17.8	17.8	17.3	43	46	67	24	7	37
					48	41	71	11	4	74
					42	30	48	15	23	
					28	24	28	40	55	

**6-61.** Construct and interpret either a digidot plot or a stem-and-leaf plot of these data. The 100 annual Wolfer sunspot numbers from 1770 to 1869 follow. (For an interesting analysis and interpretation of these numbers, see the book by Box, Jenkins, and Reinsel referenced in Exercise 6-60. Their analysis requires some advanced knowledge of statistics and statistical model building.) (read down, then left to right)

- (a) Construct a time series plot of these data.  
 (b) Construct and interpret either a digidot plot or a stem-and-leaf plot of these data.

101	31	154	38	83	90
82	7	125	23	132	67
66	20	85	10	131	60
35	92	68	24	118	47

**6-62.** In their book *Forecasting and Time Series Analysis*, 2nd edition (McGraw-Hill, 1990), D. C. Montgomery, L. A. Johnson, and J. S. Gardiner analyze the data in Table 6-5, which are the monthly total passenger airline miles flown in the United Kingdom, 1964–1970 (in millions of miles).

- (a) Draw a time series plot of the data and comment on any features of the data that are apparent.  
 (b) Construct and interpret either a digidot plot or a stem-and-leaf plot of these data.

## 6-7 PROBABILITY PLOTS

How do we know if a particular probability distribution is a reasonable model for data? Sometimes, this is an important question because many of the statistical techniques presented in subsequent chapters are based on an assumption that the population distribution is of a specific type. Thus, we can think of determining whether data come from a specific



probability distribution as **verifying assumptions**. In other cases, the form of the distribution can give insight into the underlying physical mechanism generating the data. For example, in reliability engineering, verifying that time-to-failure data come from an exponential distribution identifies the **failure mechanism** in the sense that the failure rate is constant with respect to time.

Some of the visual displays we have used earlier, such as the histogram, can provide insight about the form of the underlying distribution. However, histograms are usually not really reliable indicators of the distribution form unless the sample size is very large. **Probability plotting** is a graphical method for determining whether sample data conform to a hypothesized distribution based on a subjective visual examination of the data. The general procedure is very simple and can be performed quickly. It is also more reliable than the histogram for small to moderate size samples. Probability plotting typically uses special graph paper, known as **probability paper**, that has been designed for the hypothesized distribution. Probability paper is widely available for the normal, lognormal, Weibull, and various chi-square and gamma distributions. We focus primarily on normal probability plots because many statistical techniques are appropriate only when the population is (at least approximately) normal.

To construct a probability plot, the observations in the sample are first ranked from smallest to largest. That is, the sample  $x_1, x_2, \dots, x_n$  is arranged as  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ , where  $x_{(1)}$  is the smallest observation,  $x_{(2)}$  is the second smallest observation, and so forth, with  $x_{(n)}$  the largest. The ordered observations  $x_{(j)}$  are then plotted against their observed cumulative frequency  $(j - 0.5)/n$  on the appropriate probability paper. If the hypothesized distribution adequately describes the data, the plotted points will fall approximately along a straight line; if the plotted points deviate significantly from a straight line, the hypothesized model is not appropriate. Usually, the determination of whether or not the data plot as a straight line is subjective. The procedure is illustrated in the following example.

#### EXAMPLE 6-7

Ten observations on the effective service life in minutes of batteries used in a portable personal computer are as follows: 176, 191, 214, 220, 205, 192, 201, 190, 183, 185. We hypothesize that battery life is adequately modeled by a normal distribution. To use probability plotting to investigate this hypothesis, first arrange the observations in ascending order and calculate their cumulative frequencies  $(j - 0.5)/10$  as shown in Table 6-6.

The pairs of values  $x_{(j)}$  and  $(j - 0.5)/10$  are now plotted on normal probability paper. This plot is shown in Fig. 6-19. Most normal probability paper plots  $100(j - 0.5)/n$  on the left vertical scale and  $100[1 - (j - 0.5)/n]$  on the right vertical scale, with the variable value plotted on the horizontal scale. A straight line, chosen subjectively, has been drawn through the plotted points. In drawing the straight line, you should be influenced more by the points near the middle of the plot than by the extreme points. A good rule of thumb is to draw the line approximately between the 25th and 75th percentile points. This is how the line in Fig. 6-19 was determined. In assessing the “closeness” of the points to the straight line, imagine a “fat pencil” lying along the line. If all the points are covered by this imaginary pencil, a normal distribution adequately describes the data. Since the points in Fig. 6-19 would pass the “fat pencil” test, we conclude that the normal distribution is an appropriate model.

A **normal probability plot** can also be constructed on ordinary graph paper by plotting the standardized normal scores  $z_j$  against  $x_{(j)}$ , where the standardized normal scores satisfy

$$\frac{j - 0.5}{n} = P(Z \leq z_j) = \Phi(z_j)$$

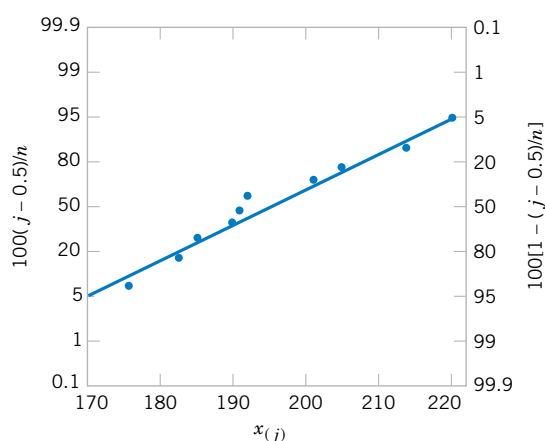


Figure 6-19 Normal probability plot for battery life.

Table 6-6 Calculation for Constructing a Normal Probability Plot

$j$	$x_{(j)}$	$(j - 0.5)/10$	$z_j$
1	176	0.05	-1.64
2	183	0.15	-1.04
3	185	0.25	-0.67
4	190	0.35	-0.39
5	191	0.45	-0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64

For example, if  $(j - 0.5)/n = 0.05$ ,  $\Phi(z_j) = 0.05$  implies that  $z_j = -1.64$ . To illustrate, consider the data from Example 6-4. In the last column of Table 6-6 we show the standardized normal scores. Figure 6-20 presents the plot of  $z_j$  versus  $x_{(j)}$ . This normal probability plot is equivalent to the one in Fig. 6-19.

We have constructed our probability plots with the probability scale (or the  $z$ -scale) on the vertical axis. Some computer packages “flip” the axis and put the probability scale on the horizontal axis.

The normal probability plot can be useful in identifying distributions that are symmetric but that have tails that are “heavier” or “lighter” than the normal. They can also be useful in identifying skewed distributions. When a sample is selected from a light-tailed distribution (such as the uniform distribution), the smallest and largest observations will not be as extreme as would be expected in a sample from a normal distribution. Thus if we consider the straight line drawn through the observations at the center of the normal probability plot, observations on the left side will tend to fall below the line, whereas observations on the right side will tend to fall above the line. This will produce an S-shaped normal probability plot such as shown in

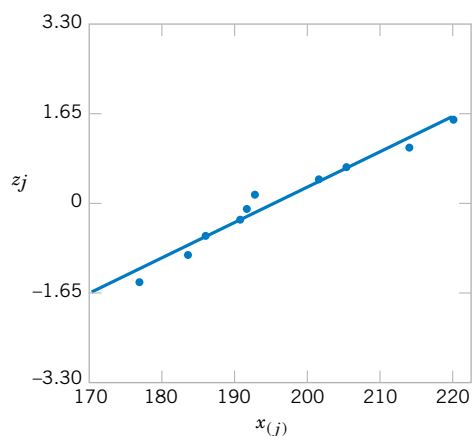
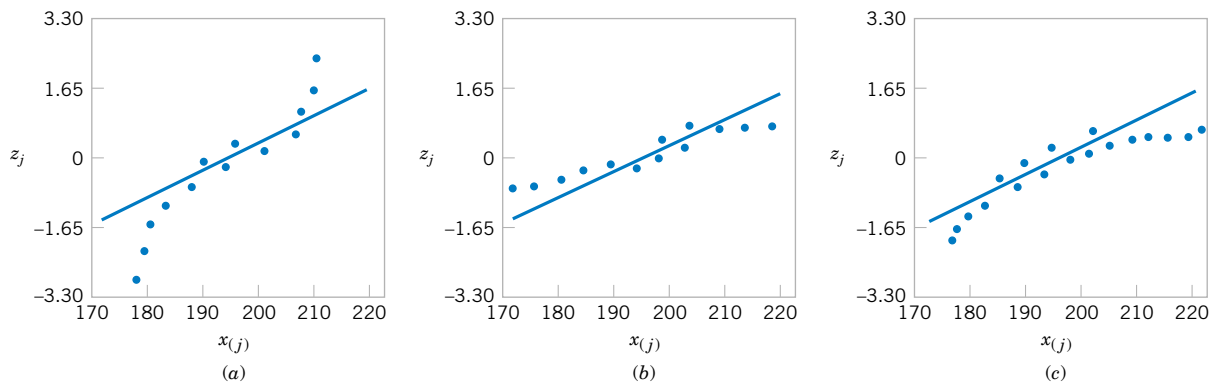


Figure 6-20 Normal probability plot obtained from standardized normal scores.











**Figure 6-21** Normal probability plots indicating a nonnormal distribution. (a) Light-tailed distribution. (b) Heavy-tailed distribution. (c) A distribution with positive (or right) skew.

Fig. 6-21(a). A heavy-tailed distribution will result in data that also produces an S-shaped normal probability plot, but now the observations on the left will be above the straight line and the observations on the right will lie below the line. See Fig. 6-19(b). A positively skewed distribution will tend to produce a pattern such as shown in Fig. 6-19(c), where points on both ends of the plot tend to fall below the line, giving a curved shape to the plot. This occurs because both the smallest and the largest observations from this type of distribution are larger than expected in a sample from a normal distribution.

Even when the underlying population is exactly normal, the sample data will not plot exactly on a straight line. Some judgment and experience are required to evaluate the plot. Generally, if the sample size is  $n < 30$ , there can be a lot of deviation from linearity in normal plots, so in these cases only a very severe departure from linearity should be interpreted as a strong indication of nonnormality. As  $n$  increases, the linear pattern will tend to become stronger, and the normal probability plot will be easier to interpret and more reliable as an indicator of the form of the distribution.

### EXERCISES FOR SECTION 6-7

-  **6-63.** Construct a normal probability plot of the piston ring diameter data in Exercise 6-1. Does it seem reasonable to assume that piston ring diameter is normally distributed?
-  **6-64.** Construct a normal probability plot of the insulating fluid breakdown time data in Exercise 6-2. Does it seem reasonable to assume that breakdown time is normally distributed?
-  **6-65.** Construct a normal probability plot of the visual accommodation data in Exercise 6-5. Does it seem reasonable to assume that visual accommodation is normally distributed?
-  **6-66.** Construct a normal probability plot of the O-ring joint temperature data in Exercise 6-13. Does it seem reasonable to assume that O-ring joint temperature is normally distributed? Discuss any interesting features that you see on the plot.
-  **6-67.** Construct a normal probability plot of the octane rating data in Exercise 6-14. Does it seem reasonable to assume that octane rating is normally distributed?
-  **6-68.** Construct a normal probability plot of the cycles to failure data in Exercise 6-15. Does it seem reasonable to assume that cycles to failure is normally distributed?
-  **6-69.** Construct a normal probability plot of the wine quality rating data in Exercise 6-27. Does it seem reasonable to assume that this variable is normally distributed?
-  **6-70.** Construct a normal probability plot of the suspended solids concentration data in Exercise 6-24. Does it seem reasonable to assume that the concentration of suspended solids in water from this particular lake is normally distributed?

**6-71.** Construct two normal probability plots for the height data in Exercises 6-22 and 6-29. Plot the data for female and male students on the same axes. Does height seem to be normally distributed for either group of students? If both populations have the same variance, the two normal probability plots should have identical slopes. What conclusions would you draw about the heights of the two groups of students from visual examination of the normal probability plots?

**6-72.** It is possible to obtain a “quick and dirty” estimate of the mean of a normal distribution from the fiftieth percentile value on a normal probability plot. Provide an argument why this is so. It is also possible to obtain an estimate of the standard deviation of a normal distribution by subtracting the sixty-fourth percentile value from the fiftieth percentile value. Provide an argument why this is so.

## 6-8 MORE ABOUT PROBABILITY PLOTTING (CD ONLY)

### Supplemental Exercises

**6-73.** The concentration of a solution is measured six times by one operator using the same instrument. She obtains the following data: 63.2, 67.1, 65.8, 64.0, 65.1, and 65.3 (grams per liter).

- Calculate the sample mean. Suppose that the desirable value for this solution has been specified to be 65.0 grams per liter. Do you think that the sample mean value computed here is close enough to the target value to accept the solution as conforming to target? Explain your reasoning.
- Calculate the sample variance and sample standard deviation.
- Suppose that in measuring the concentration, the operator must set up an apparatus and use a reagent material. What do you think the major sources of variability are in this experiment? Why is it desirable to have a small variance of these measurements?

**6-74.** A sample of six resistors yielded the following resistances (ohms):  $x_1 = 45$ ,  $x_2 = 38$ ,  $x_3 = 47$ ,  $x_4 = 41$ ,  $x_5 = 35$ , and  $x_6 = 43$ .

- Compute the sample variance and sample standard deviation.
- Subtract 35 from each of the original resistance measurements and compute  $s^2$  and  $s$ . Compare your results with those obtained in part (a) and explain your findings.
- If the resistances were 450, 380, 470, 410, 350, and 430 ohms, could you use the results of previous parts of this problem to find  $s^2$  and  $s$ ?

**6-75.** Consider the following two samples:

Sample 1: 10, 9, 8, 7, 8, 6, 10, 6

Sample 2: 10, 6, 10, 6, 8, 10, 8, 6

- Calculate the sample range for both samples. Would you conclude that both samples exhibit the same variability? Explain.
- Calculate the sample standard deviations for both samples. Do these quantities indicate that both samples have the same variability? Explain.
- Write a short statement contrasting the sample range versus the sample standard deviation as a measure of variability.

**6-76.** An article in *Quality Engineering* (Vol. 4, 1992, pp. 487–495) presents viscosity data from a batch chemical process. A sample of these data follows:

13.3	14.3	14.9	15.2	15.8	14.2	16.0	14.0
14.5	16.1	13.7	15.2	13.7	16.9	14.9	14.4
15.3	13.1	15.2	15.9	15.1	14.9	13.6	13.7
15.3	15.5	14.5	16.5	13.4	15.2	15.3	13.8
14.3	12.6	15.3	14.8	14.1	14.4	14.3	15.6
14.8	14.6	15.6	15.1	14.8	15.2	15.6	14.5
15.2	14.3	15.8	17.0	14.3	14.6	16.1	12.8
14.5	15.4	13.3	14.9	14.3	16.4	13.9	16.1
14.6	15.2	14.1	14.8	16.4	14.2	15.2	16.6
14.1	16.8	15.4	14.0	16.9	15.7	14.4	15.6

- Reading down and left to right, draw a time series plot of all the data and comment on any features of the data that are revealed by this plot.
- Consider the notion that the first 40 observations were generated from a specific process, whereas the last 40 observations were generated from a different process. Does the plot indicate that the two processes generate similar results?
- Compute the sample mean and sample variance of the first 40 observations; then compute these values for the second 40 observations. Do these quantities indicate that both processes yield the same mean level? The same variability? Explain.

**6-77.** Reconsider the data from Exercise 6-76. Prepare comparative box plots for two groups of observations: the first 40 and the last 40. Comment on the information in the box plots.

**6-78.** The data shown in Table 6-7 are monthly champagne sales in France (1962–1969) in thousands of bottles.

- Construct a time series plot of the data and comment on any features of the data that are revealed by this plot.
- Speculate on how you would use a graphical procedure to forecast monthly champagne sales for the year 1970.

**6-79.** A manufacturer of coil springs is interested in implementing a quality control system to monitor his production process. As part of this quality system, it is decided to record the number of nonconforming coil springs in each production

**Table 6-7** Champagne Sales in France

Month	1962	1963	1964	1965	1966	1967	1968	1969
Jan.	2.851	2.541	3.113	5.375	3.633	4.016	2.639	3.934
Feb.	2.672	2.475	3.006	3.088	4.292	3.957	2.899	3.162
Mar.	2.755	3.031	4.047	3.718	4.154	4.510	3.370	4.286
Apr.	2.721	3.266	3.523	4.514	4.121	4.276	3.740	4.676
May	2.946	3.776	3.937	4.520	4.647	4.968	2.927	5.010
June	3.036	3.230	3.986	4.539	4.753	4.677	3.986	4.874
July	2.282	3.028	3.260	3.663	3.965	3.523	4.217	4.633
Aug.	2.212	1.759	1.573	1.643	1.723	1.821	1.738	1.659
Sept.	2.922	3.595	3.528	4.739	5.048	5.222	5.221	5.591
Oct.	4.301	4.474	5.211	5.428	6.922	6.873	6.424	6.981
Nov.	5.764	6.838	7.614	8.314	9.858	10.803	9.842	9.851
Dec.	7.132	8.357	9.254	10.651	11.331	13.916	13.076	12.670

batch of size 50. During 40 days of production, 40 batches of data were collected as follows:

Read data across.

9	12	6	9	7	14	12	4	6	7
8	5	9	7	8	11	3	6	7	7
11	4	4	8	7	5	6	4	5	8
19	19	18	12	11	17	15	17	13	13

- Construct a stem-and-leaf plot of the data.
- Find the sample average and standard deviation.
- Construct a time series plot of the data. Is there evidence that there was an increase or decrease in the average number of nonconforming springs made during the 40 days? Explain.

**6-80.** A communication channel is being monitored by recording the number of errors in a string of 1000 bits. Data for 20 of these strings follow:

Read data across.

3	1	0	1	3	2	4	1	3	1
1	1	2	3	3	2	0	2	0	1

- Construct a stem-and-leaf plot of the data.
- Find the sample average and standard deviation.
- Construct a time series plot of the data. Is there evidence that there was an increase or decrease in the number of errors in a string? Explain.

**6-81.** Reconsider the data in Exercise 6-76. Construct normal probability plots for two groups of the data: the first 40 and the last 40 observations. Construct both plots on the same axes. What tentative conclusions can you draw?

**6-82.** Construct a normal probability plot of the effluent discharge temperature data from Exercise 6-47. Based on the plot, what tentative conclusions can you draw?

**6-83.** Construct normal probability plots of the cold start ignition time data presented in Exercises 6-44 and 6-56.

Construct a separate plot for each gasoline formulation, but arrange the plots on the same axes. What tentative conclusions can you draw?

**6-84. Transformations.** In some data sets, a transformation by some mathematical function applied to the original data, such as  $\sqrt{y}$  or  $\log y$ , can result in data that are simpler to work with statistically than the original data. To illustrate the effect of a transformation, consider the following data, which represent cycles to failure for a yarn product: 675, 3650, 175, 1150, 290, 2000, 100, 375.

- Construct a normal probability plot and comment on the shape of the data distribution.
- Transform the data using logarithms; that is, let  $y^*$  (new value) =  $\log y$  (old value). Construct a normal probability plot of the transformed data and comment on the effect of the transformation.

**6-85.** In 1879, A. A. Michelson made 100 determinations of the velocity of light in air using a modification of a method proposed by the French physicist Foucault. He made the measurements in five trials of 20 measurements each. The observations (in kilometers per second) follow. Each value has 299,000 subtracted from it.

#### Trial 1

850	900	930	950	980
1000	930	760	1000	960
740	1070	850	980	880
980	650	810	1000	960

#### Trial 2

960	960	880	850	900
830	810	880	800	760
940	940	800	880	840
790	880	830	790	800

**Trial 3**

880	880	720	620	970
880	850	840	850	840
880	860	720	860	950
910	870	840	840	840

**Trial 4**

890	810	800	760	750
910	890	880	840	850
810	820	770	740	760
920	860	720	850	780

**Trial 5**

890	780	760	790	820
870	810	810	950	810
840	810	810	810	850
870	740	940	800	870

The currently accepted true velocity of light in a vacuum is 299,792.5 kilometers per second. Stigler (1977, *The Annals of Statistics*) reports that the “true” value for comparison to these measurements is 734.5. Construct comparative box plots of these measurements. Does it seem that all five trials are con-

sistent with respect to the variability of the measurements? Are all five trials centered on the same value? How does each group of trials compare to the true value? Could there have been “startup” effects in the experiment that Michelson performed? Could there have been bias in the measuring instrument?

**6-86.** In 1789, Henry Cavendish estimated the density of the earth by using a torsion balance. His 29 measurements follow, expressed as a multiple of the density of water.

5.50	5.30	5.47	5.10	5.29	5.65
5.55	5.61	5.75	5.63	5.27	5.44
5.57	5.36	4.88	5.86	5.34	5.39
5.34	5.53	5.29	4.07	5.85	5.46
5.42	5.79	5.62	5.58	5.26	



- Calculate the sample mean, sample standard deviation, and median of the Cavendish density data.
- Construct a normal probability plot of the data. Comment on the plot. Does there seem to be a “low” outlier in the data?
- Would the sample median be a better estimate of the density of the earth than the sample mean? Why?

### MIND-EXPANDING EXERCISES

**6-87.** Consider the airfoil data in Exercise 6-12. Subtract 30 from each value and then multiply the resulting quantities by 10. Now compute  $s^2$  for the new data. How is this quantity related to  $s^2$  for the *original* data? Explain why.

**6-88.** Consider the quantity  $\sum_{i=1}^n (x_i - a)^2$ . For what value of  $a$  is this quantity minimized?

**6-89** Using the results of Exercise 6-87, which of the two quantities  $\sum_{i=1}^n (x_i - \bar{x})^2$  and  $\sum_{i=1}^n (x_i - \mu)^2$  will be smaller, provided that  $\bar{x} \neq \mu$ ?

**6-90. Coding the Data.** Let  $y_i = a + bx_i$ ,  $i = 1, 2, \dots, n$ , where  $a$  and  $b$  are nonzero constants. Find the relationship between  $\bar{x}$  and  $\bar{y}$ , and between  $s_x$  and  $s_y$ .

**6-91.** A sample of temperature measurements in a furnace yielded a sample average ( $^{\circ}\text{F}$ ) of 835.00 and a sample standard deviation of 10.5. Using the results from Exercise 6-90, what are the sample average and sample standard deviations expressed in  $^{\circ}\text{C}$ ?

**6-92.** Consider the sample  $x_1, x_2, \dots, x_n$  with sample mean  $\bar{x}$  and sample standard deviation  $s$ . Let  $z_i = (x_i - \bar{x})/s$ ,  $i = 1, 2, \dots, n$ . What are the values of the sample mean and sample standard deviation of the  $z_i$ ?

**6-93.** An experiment to investigate the survival time in hours of an electronic component consists of placing the parts in a test cell and running them for 100 hours under elevated temperature conditions. (This is called an “accelerated” life test.) Eight components were tested with the following resulting failure times:

75, 63, 100<sup>+</sup>, 36, 51, 45, 80, 90

The observation 100<sup>+</sup> indicates that the unit still functioned at 100 hours. Is there any meaningful measure of location that can be calculated for these data? What is its numerical value?

**6-94.** Suppose that we have a sample  $x_1, x_2, \dots, x_n$  and we have calculated  $\bar{x}_n$  and  $s_n^2$  for the sample. Now an  $(n + 1)$ st observation becomes available. Let  $\bar{x}_{n+1}$  and  $s_{n+1}^2$  be the sample mean and sample variance for the sample using all  $n + 1$  observations.

(a) Show how  $\bar{x}_{n+1}$  can be computed using  $\bar{x}_n$  and  $x_{n+1}$ .

(b) Show that  $ns_{n+1}^2 = (n - 1)s_n^2 + \frac{n(x_{n+1} - \bar{x}_n)^2}{n + 1}$

(c) Use the results of parts (a) and (b) to calculate the new sample average and standard deviation for the data of Exercise 6-22, when the new observation is  $x_{38} = 64$ .

**6-95. The Trimmed Mean.** Suppose that the data are arranged in increasing order,  $T\%$  of the observations are removed from each end and the sample mean of the remaining numbers is calculated. The resulting quantity is called a *trimmed mean*. The trimmed mean generally lies between the sample mean  $\bar{x}$  and the sample median  $\bar{x}$ . Why?

(a) Calculate the 10% trimmed mean for the yield data in Exercise 6-17.

(b) Calculate the 20% trimmed mean for the yield data in Exercise 6-17 and compare it with the quantity found in part (a).

(c) Compare the values calculated in parts (a) and (b) with the sample mean and median for the yield data. Is there much difference in these quantities? Why?

**6-96. The Trimmed Mean.** Suppose that the sample size  $n$  is such that the quantity  $nT/100$  is not an integer. Develop a procedure for obtaining a trimmed mean in this case.

### IMPORTANT TERMS AND CONCEPTS

In the E-book, click on any term or concept below to go to that subject.

Box plot

Frequency distribution and histogram

Median, quartiles and percentiles

Normal probability plot

Population mean

Population standard deviation

Population variance

Random sample

Sample mean

Sample standard deviation

Sample variance

Stem-and-leaf diagram

Time series plots

### CD MATERIAL

Exponential probability plot

Goodness of fit

Weibull probability plot



## 6-8 MORE ABOUT PROBABILITY PLOTTING (CD ONLY)

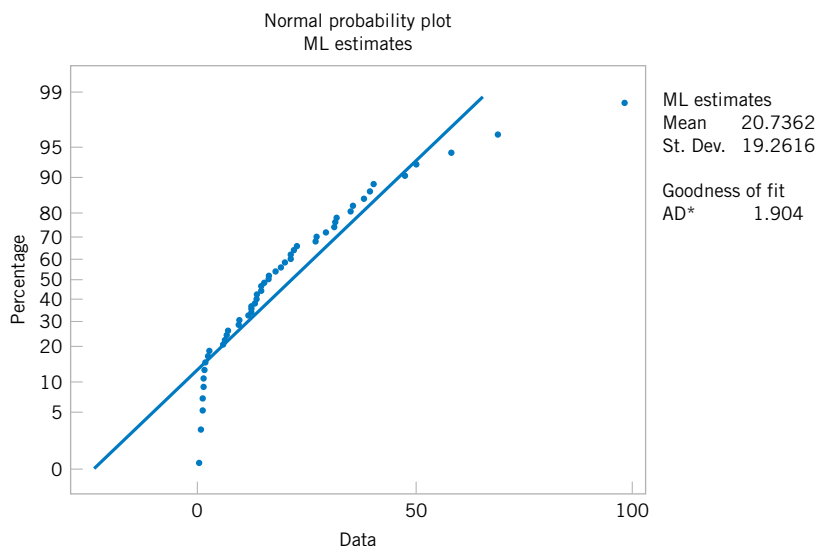
Probability plots are extremely useful and are often the first technique used in an effort to determine which probability distribution is likely to provide a reasonable model for the data.

We give a simple illustration of how a normal probability plot can be useful in distinguishing between normal and nonnormal data. Table S6-1 contains 50 observations generated at random from an exponential distribution with mean 20 (or  $\lambda = 0.05$ ). These data were generated using the random number generation capability in Minitab. Figure S6-1 presents a normal probability plot of these data, constructed using Minitab. The observations do not even approximately lie along a straight line, giving a clear indication that the data do not follow a normal distribution. The strong curvature at both ends of the plot suggests that the data come from a distribution with right or positive skew. Compare Fig. S6-1 with Fig. 6-19c.

Minitab also provides estimates of the mean and standard deviation of the distribution using the **method of maximum likelihood** (abbreviated ML on the graph in Figure S6-1). We will discuss maximum likelihood estimation in Chapter 7. For the normal distribution, this is the familiar sample mean and sample standard deviation that we first presented in Chapter 1. Minitab also presents a quantitative measure of how well the data are described by a normal distribution. This goodness-of-fit measure is called the Anderson-Darling statistic (abbreviated AD on the Minitab probability plot). The Anderson-Darling statistic is based on the probability integral transformation

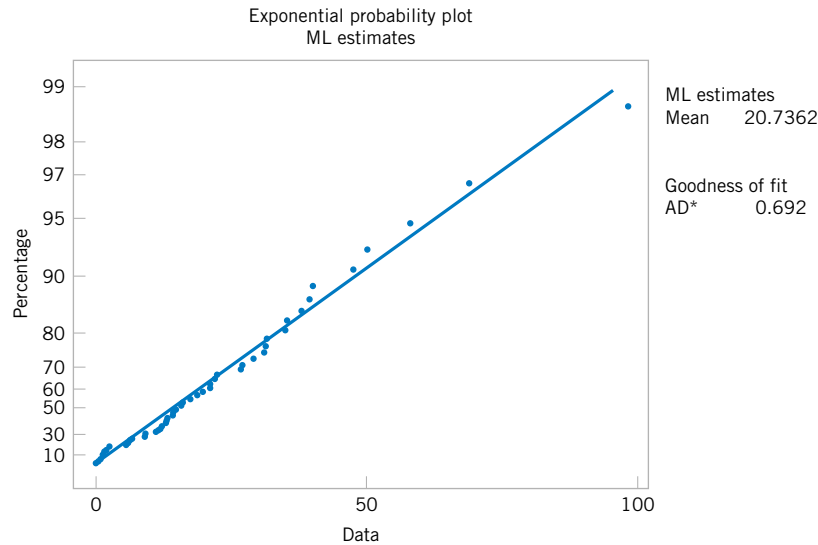
$$F(x) = \int_{-\infty}^x f(u) du$$

that can be used to convert the data to a uniform distribution if the hypothesized distribution is correct. Thus, if  $x_1, x_2, \dots, x_n$  are independent and identically distributed random variables whose cumulative distribution function is  $F(x)$ , then  $F(x_1), F(x_2), \dots, F(x_n)$  are independent uniform (0, 1) random variables. The Anderson-Darling statistic essentially compares how close the  $F(x_1), F(x_2), \dots, F(x_n)$  values are to values from a uniform (0, 1) distribution. For



**Figure S6-1.** Normal probability plot (from Minitab) of the data from Table S6-1.



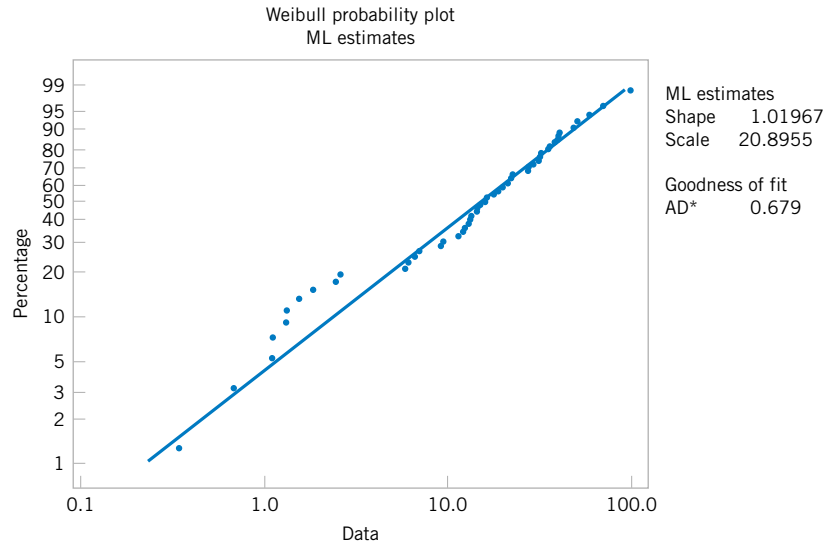


**Figure S6-2.** Exponential probability plot (from Minitab) of the data from Table S6-1.

this reason, the Anderson-Darling test is sometimes called a “distance” test. The test is upper-tailed; that is, if the computed value exceeds a critical value, the hypothesis of normality is rejected. The 5% critical value of the Anderson-Darling statistic is 0.752 and the 1% value is 1.035. Because the Anderson-Darling statistic in Figure S6-1 is 1.904, and this exceeds the 1% critical value, we conclude that the assumption of normality would be inappropriate.

Minitab can construct several other types of probability plots. An exponential probability plot of the data in Table S6-1 is shown in Figure S6-2. Notice that the data lies very close to the straight line in this plot, implying that the exponential is a good model for the data. Minitab also provides an estimate of the mean of the exponential distribution. This estimate is just the sample mean.

Figure S6-3 is a Weibull probability plot of the data from Table S6-1, constructed using Minitab. The data lies approximately along a straight line, suggesting that the Weibull distribution is also a reasonable model for the data. Notice that Minitab provides maximum



**Figure S6-3.** Weibull probability plot (from Minitab) of the data from Table S6-1.

**Table S6-1** 50 Observations Drawn at Random from an Exponential Distribution with Mean 20 ( $\lambda = 0.05$ )

1.2934	14.1968	13.2798	16.1154	14.8891	31.6489
2.4330	13.1818	5.7511	39.9558	97.8874	21.1057
15.8770	9.0942	31.3655	1.3104	12.0008	11.2846
21.9606	21.1336	26.8364	9.3134	31.0346	29.0222
47.4481	0.3389	1.0999	19.8350	1.5191	2.5623
34.8720	39.2494	12.1621	18.8295	35.3307	
27.0908	0.6731	14.2467	49.9397	6.0479	
17.5253	68.7876	57.8919	1.0882	22.4244	
6.8290	37.9023	1.8219	6.4967	12.8239	

likelihood estimates of the shape parameter  $\beta$  and the scale parameter  $\delta$ . The shape parameter estimate is very close to unity, and we know that a Weibull distribution with  $\beta = 1$  is the exponential distribution.