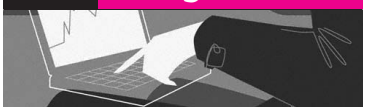
1 Introduction and Descriptive Statistics

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- Distinguish between qualitative and quantitative data.
- Describe nominal, ordinal, interval, and ratio scales of measurement.
- Describe the difference between a population and a sample.
- Calculate and interpret percentiles and quartiles.
- Explain measures of central tendency and how to compute them.
- Create different types of charts that describe data sets.
- Use Excel templates to compute various measures and create charts.

1–1 Using Statistics



It is better to be roughly right than precisely wrong.
–John Maynard Keynes

You all have probably heard the story about Malcolm Forbes, who once got lost floating for miles in one of his famous balloons and finally landed in the middle of a cornfield. He spotted a man coming toward him and asked, "Sir, can you tell me where I am?" The man said, "Certainly, you are in a basket in a field of corn." Forbes said, "You must be a statistician." The man said, "That's amazing, how did you know that?" "Easy," said Forbes, "your information is concise, precise, and absolutely useless!"[1]

The purpose of this book is to convince you that information resulting from a good statistical analysis is always concise, often precise, and never useless! The spirit of statistics is, in fact, very well captured by the quotation above from Keynes. This book should teach you how to be at least roughly right a high percentage of the time. Statistics is a science that helps us make better decisions in business and economics as well as in other fields. Statistics teach us how to summarize data, analyze them, and draw meaningful inferences that then lead to improved decisions. These better decisions we make help us improve the running of a department, a company, or the entire economy.

The word *statistics* is derived from the Italian word *stato,* which means "state," and *statista* refers to a person involved with the affairs of state. Therefore, *statistics* originally meant the collection of facts useful to the *statista.* Statistics in this sense was used in 16th-century Italy and then spread to France, Holland, and Germany. We note, however, that surveys of people and property actually began in ancient times.[2] Today, statistics is not restricted to information about the state but extends to almost every realm of human endeavor. Neither do we restrict ourselves to merely collecting numerical information, called *data.* Our data are summarized, displayed in meaningful ways, and analyzed. Statistical analysis often involves an attempt to generalize from the data. Statistics is a science–the science of information. Information may be *qualitative* or *quantitative.* To illustrate the difference between these two types of information, let's consider an example.

EXAMPLE 1–1

Realtors who help sell condominiums in the Boston area provide prospective buyers with the information given in Table 1–1. Which of the variables in the table are quantitative and which are qualitative?

Solution

The asking price is a *quantitative* variable: it conveys a quantity–the asking price in dollars. The number of rooms is also a quantitative variable. The direction the apartment faces is a *qualitative* variable since it conveys a quality (east, west, north, south). Whether a condominium has a washer and dryer in the unit (yes or no) and whether there is a doorman are also qualitative variables.

[1]From an address by R. Gnanadesikan to the American Statistical Association, reprinted in *American Statistician* 44, no. 2 (May 1990), p. 122.

[2]See Anders Hald, *A History of Probability and Statistics and Their Applications before 1750* (New York: Wiley, 1990), pp. 81–82.

TABLE 1–1 Boston Condominium Data

Asking Price	Number of Bedrooms	Number of Bathrooms	Direction Facing	Washer/Dryer?	Doorman?
$709,000	2	1	E	Y	Y
812,500	2	2	N	N	Y
980,000	3	3	N	Y	Y
830,000	1	2	W	N	N
850,900	2	2	W	Y	N

Source: Boston.condocompany.com, March 2007.

A **quantitative variable** can be described by a number for which arithmetic operations such as averaging make sense. A **qualitative** (or **categorical**) **variable** simply records a quality. If a number is used for distinguishing members of different categories of a qualitative variable, the number assignment is arbitrary.

The field of statistics deals with **measurements**–some quantitative and others qualitative. The measurements are the actual numerical values of a variable. (Qualitative variables could be described by numbers, although such a description might be arbitrary; for example, $N = 1$, $E = 2$, $S = 3$, $W = 4$, $Y = 1$, $N = 0$.)

The four generally used **scales of measurement** are listed here from weakest to strongest.

Nominal Scale. In the **nominal scale** of measurement, numbers are used simply as labels for groups or classes. If our data set consists of blue, green, and red items, we may designate blue as 1, green as 2, and red as 3. In this case, the numbers 1, 2, and 3 stand only for the category to which a data point belongs. "Nominal" stands for "name" of category. The nominal scale of measurement is used for qualitative rather than quantitative data: blue, green, red; male, female; professional classification; geographic classification; and so on.

Ordinal Scale. In the **ordinal scale** of measurement, data elements may be ordered according to their relative size or quality. Four products ranked by a consumer may be ranked as 1, 2, 3, and 4, where 4 is the best and 1 is the worst. In this scale of measurement we do not know how much better one product is than others, only that it is better.

Interval Scale. In the **interval scale** of measurement the value of zero is assigned arbitrarily and therefore we cannot take ratios of two measurements. But *we can take ratios of intervals.* A good example is how we measure time of day, which is in an interval scale. We cannot say 10:00 A.M. is twice as long as 5:00 A.M. But we can say that the interval between 0:00 A.M. (midnight) and 10:00 A.M., which is a duration of 10 hours, is twice as long as the interval between 0:00 A.M. and 5:00 A.M., which is a duration of 5 hours. This is because 0:00 A.M. does not mean absence of any time. Another example is temperature. When we say 0°F, we do not mean zero heat. A temperature of 100°F is not twice as hot as 50°F.

Ratio Scale. If two measurements are in **ratio scale,** then we can take ratios of those measurements. The zero in this scale is an absolute zero. Money, for example, is measured in a ratio scale. A sum of $100 is twice as large as $50. A sum of $0 means absence of any money and is thus an absolute zero. We have already seen that measurement of duration (but not time of day) is in a ratio scale. In general, the interval between two interval scale measurements will be in ratio scale. Other examples of the ratio scale are measurements of weight, volume, area, or length.

Samples and Populations

In statistics we make a distinction between two concepts: a population and a sample.

The **population** consists of the set of all measurements in which the investigator is interested. The population is also called the **universe**.

A **sample** is a subset of measurements selected from the population. Sampling from the population is often done randomly, such that every possible sample of n elements will have an equal chance of being selected. A sample selected in this way is called a **simple random sample**, or just a **random sample**. A random sample allows chance to determine its elements.

For example, Farmer Jane owns 1,264 sheep. These sheep constitute her entire *population* of sheep. If 15 sheep are selected to be sheared, then these 15 represent a *sample* from Jane's population of sheep. Further, if the 15 sheep were selected at *random* from Jane's population of 1,264 sheep, then they would constitute a *random sample* of sheep.

The definitions of *sample* and *population* are relative to what we want to consider. If Jane's sheep are all we care about, then they constitute a population. If, however, we are interested in all the sheep in the county, then all Jane's 1,264 sheep are a sample of that larger population (although this sample would not be random).

The distinction between a sample and a population is very important in statistics.

Data and Data Collection

A set of measurements obtained on some variable is called a **data set.** For example, heart rate measurements for 10 patients may constitute a data set. The variable we're interested in is heart rate, and the scale of measurement here is a ratio scale. (A heart that beats 80 times per minute is twice as fast as a heart that beats 40 times per minute.) Our actual observations of the patients' heart rates, the data set, might be 60, 70, 64, 55, 70, 80, 70, 74, 51, 80.

Data are collected by various methods. Sometimes our data set consists of the entire population we're interested in. If we have the actual point spread for five football games, and if we are interested only in these five games, then our data set of five measurements is the entire population of interest. (In this case, our data are on a ratio scale. Why? Suppose the data set for the five games told only whether the home or visiting team won. What would be our measurement scale in this case?)

In other situations data may constitute a sample from some population. If the data are to be used to draw some conclusions about the larger population they were drawn from, then we must collect the data with great care. A conclusion drawn about a population based on the information in a sample from the population is called a **statistical inference.** Statistical inference is an important topic of this book. To ensure the accuracy of statistical inference, data must be drawn randomly from the population of interest, and we must make sure that every segment of the population is adequately and proportionally represented in the sample.

Statistical inference may be based on data collected in surveys or experiments, which must be carefully constructed. For example, when we want to obtain information from people, we may use a mailed questionnaire or a telephone interview as a convenient instrument. In such surveys, however, we want to minimize any **nonresponse bias.** This is the biasing of the results that occurs when we disregard the fact that some people will simply not respond to the survey. The bias distorts the findings, because the people who do not respond may belong more to one segment of the population than to another. In social research some questions may be sensitive—for example, "Have you ever been arrested?" This may easily result in a nonresponse bias, because people who have indeed been arrested may be less likely to answer the question (unless they can be perfectly certain of remaining anonymous). Surveys

conducted by popular magazines often suffer from nonresponse bias, especially when their questions are provocative. What makes good magazine reading often makes bad statistics. An article in the *New York Times* reported on a survey about Jewish life in America. The survey was conducted by calling people at home on a Saturday–thus strongly biasing the results since Orthodox Jews do not answer the phone on Saturday.[3]

Suppose we want to measure the speed performance or gas mileage of an automobile. Here the data will come from experimentation. In this case we want to make sure that a variety of road conditions, weather conditions, and other factors are represented. Pharmaceutical testing is also an example where data may come from experimentation. Drugs are usually tested against a placebo as well as against no treatment at all. When an experiment is designed to test the effectiveness of a sleeping pill, the variable of interest may be the time, in minutes, that elapses between taking the pill and falling asleep.

In experiments, as in surveys, it is important to **randomize** if inferences are indeed to be drawn. People should be randomly chosen as subjects for the experiment if an inference is to be drawn to the entire population. Randomization should also be used in assigning people to the three groups: pill, no pill, or placebo. Such a design will minimize potential biasing of the results.

In other situations data may come from published sources, such as statistical abstracts of various kinds or government publications. The published unemployment rate over a number of months is one example. Here, data are "given" to us without our having any control over how they are obtained. Again, caution must be exercised. The unemployment rate over a given period is not a random sample of any *future* unemployment rates, and making statistical inferences in such cases may be complex and difficult. If, however, we are interested only in the period we have data for, then our data do constitute an entire population, which may be described. In any case, however, we must also be careful to note any missing data or incomplete observations.

In this chapter, we will concentrate on the processing, summarization, and display of data—the first step in statistical analysis. In the next chapter, we will explore the theory of probability, the connection between the random sample and the population. Later chapters build on the concepts of probability and develop a system that allows us to draw a logical, consistent inference from our sample to the underlying population.

Why worry about inference and about a population? Why not just look at our data and interpret them? Mere inspection of the data will suffice when interest centers on the particular observations you have. If, however, you want to draw meaningful conclusions with implications extending beyond your limited data, statistical inference is the way to do it.

In marketing research, we are often interested in the relationship between advertising and sales. A data set of randomly chosen sales and advertising figures for a given firm may be of some interest in itself, but the information in it is much more useful if it leads to implications about the underlying process–the relationship between the firm's level of advertising and the resulting level of sales. An understanding of the true relationship between advertising and sales–the relationship in the population of advertising and sales possibilities for the firm–would allow us to predict sales for any level of advertising and thus to set advertising at a level that maximizes profits.

A pharmaceutical manufacturer interested in marketing a new drug may be required by the Food and Drug Administration to prove that the drug does not cause serious side effects. The results of tests of the drug on a random sample of people may then be used in a statistical inference about the entire population of people who may use the drug if it is introduced.

[3]Laurie Goodstein, "Survey Finds Slight Rise in Jews Intermarrying," *The New York Times,* September 11, 2003, p. A13.

A bank may be interested in assessing the popularity of a particular model of automatic teller machines. The machines may be tried on a randomly chosen group of bank customers. The conclusions of the study could then be generalized by statistical inference to the entire population of the bank's customers.

A quality control engineer at a plant making disk drives for computers needs to make sure that no more than 3% of the drives produced are defective. The engineer may routinely collect random samples of drives and check their quality. Based on the random samples, the engineer may then draw a conclusion about the proportion of defective items in the entire population of drives.

These are just a few examples illustrating the use of statistical inference in business situations. In the rest of this chapter, we will introduce the descriptive statistics needed to carry out basic statistical analyses. The following chapters will develop the elements of inference from samples to populations.

PROBLEMS

1–1. A survey by an electric company contains questions on the following:

1. Age of household head.
2. Sex of household head.
3. Number of people in household.
4. Use of electric heating (yes or no).
5. Number of large appliances used daily.
6. Thermostat setting in winter.
7. Average number of hours heating is on.
8. Average number of heating days.
9. Household income.
10. Average monthly electric bill.
11. Ranking of this electric company as compared with two previous electricity suppliers.

Describe the variables implicit in these 11 items as quantitative or qualitative, and describe the scales of measurement.

1–2. Discuss the various data collection methods described in this section.

1–3. Discuss and compare the various scales of measurement.

1–4. Describe each of the following variables as qualitative or quantitative.

The Richest People on Earth 2007

Name	Wealth ($ billion)	Age	Industry	Country of Citizenship
William Gates III	56	51	Technology	U.S.A.
Warren Buffett	52	76	Investment	U.S.A.
Carlos Slim Helú	49	67	Telecom	Mexico
Ingvar Kamprad	33	80	Retail	Sweden
Bernard Arnault	26	58	Luxury goods	France

Source: *Forbes*, March 26, 2007 (the "billionaires" issue), pp. 104–156.

1–5. Five ice cream flavors are rank-ordered by preference. What is the scale of measurement?

1–6. What is the difference between a qualitative and a quantitative variable?

1–7. A town has 15 neighborhoods. If you interviewed everyone living in one particular neighborhood, would you be interviewing a population or a sample from the town?

Would this be a random sample? If you had a list of everyone living in the town, called a **frame,** and you randomly selected 100 people from all the neighborhoods, would this be a random sample?

1–8. What is the difference between a sample and a population?

1–9. What is a random sample?

1–10. For each tourist entering the United States, the U.S. Immigration and Naturalization Service computer is fed the tourist's nationality and length of intended stay. Characterize each variable as quantitative or qualitative.

1–11. What is the scale of measurement for the color of a karate belt?

1–12. An individual federal tax return form asks, among other things, for the following information: income (in dollars and cents), number of dependents, whether filing singly or jointly with a spouse, whether or not deductions are itemized, amount paid in local taxes. Describe the scale of measurement of each variable, and state whether the variable is qualitative or quantitative.

1–2 Percentiles and Quartiles

Given a set of numerical observations, we may order them according to magnitude. Once we have done this, it is possible to define the boundaries of the set. Any student who has taken a nationally administered test, such as the Scholastic Aptitude Test (SAT), is familiar with *percentiles*. Your score on such a test is compared with the scores of all people who took the test at the same time, and your position within this group is defined in terms of a percentile. If you are in the 90th percentile, 90% of the people who took the test received a score lower than yours. We define a percentile as follows.

The Pth **percentile** of a group of numbers is that value below which lie P% (P percent) of the numbers in the group. The position of the Pth percentile is given by $(n + 1)P/100$, where n is the number of data points.

Let's look at an example.

EXAMPLE 1–2

The magazine *Forbes* publishes annually a list of the world's wealthiest individuals. For 2007, the net worth of the 20 richest individuals, in billions of dollars, in no particular order, is as follows:[4]

33, 26, 24, 21, 19, 20, 18, 18, 52, 56, 27, 22, 18, 49, 22, 20, 23, 32, 20, 18

Find the 50th and 80th percentiles of this set of the world's top 20 net worths.

Solution First, let's order the data from smallest to largest:

18, 18, 18, 18, 19, 20, 20, 20, 21, 22, 22, 23, 24, 26, 27, 32, 33, 49, 52, 56

To find the 50th percentile, we need to determine the data point in position $(n + 1)P/100 = (20 + 1)(50/100) = (21)(0.5) = 10.5$. Thus, we need the data point in position 10.5. Counting the observations from smallest to largest, we find that the 10th observation is 22, and the 11th is 22. Therefore, the observation that would lie in position 10.5 (halfway between the 10th and 11th observations) is 22. Thus, the 50th percentile is 22.

Similarly, we find the 80th percentile of the data set as the observation lying in position $(n + 1)P/100 = (21)(80/100) = 16.8$. The 16th observation is 32, and the 17th is 33; therefore, the 80th percentile is a point lying 0.8 of the way from 32 to 33, that is, 32.8.

[4] *Forbes,* March 26, 2007 (the "billionaires" issue), pp. 104–186.

Certain percentiles have greater importance than others because they break down the **distribution** of the data (the way the data points are distributed along the number line) into four groups. These are the quartiles. **Quartiles** are the percentage points that break down the data set into quarters—first quarter, second quarter, third quarter, and fourth quarter.

The **first quartile** is the 25th percentile. It is that point below which lie one-fourth of the data.

Similarly, the second quartile is the 50th percentile, as we computed in Example 1–2. This is a most important point and has a special name—the *median*.

The **median** is the point below which lie half the data. It is the 50th percentile.

We define the third quartile correspondingly:

The **third quartile** is the 75th percentile point. It is that point below which lie 75 percent of the data.

The 25th percentile is often called the **lower quartile;** the 50th percentile point, the median, is called the **middle quartile;** and the 75th percentile is called the **upper quartile.**

EXAMPLE 1–3

Find the lower, middle, and upper quartiles of the billionaires data set in Example 1–2.

Solution

Based on the procedure we used in computing the 80th percentile, we find that the lower quartile is the observation in position $(21)(0.25) = 5.25$, which is 19.25. The middle quartile was already computed (it is the 50th percentile, the median, which is 22). The upper quartile is the observation in position $(21)(75/100) = 15.75$, which is 30.75.

We define the **interquartile range** as the difference between the first and third quartiles.

The interquartile range is a measure of the spread of the data. In Example 1–2, the interquartile range is equal to Third quartile − First quartile = 30.75 − 19.25 = 11.5.

PROBLEMS

1–13. The following data are numbers of passengers on flights of Delta Air Lines between San Francisco and Seattle over 33 days in April and early May.

128, 121, 134, 136, 136, 118, 123, 109, 120, 116, 125, 128, 121, 129, 130, 131, 127, 119, 114, 134, 110, 136, 134, 125, 128, 123, 128, 133, 132, 136, 134, 129, 132

Find the lower, middle, and upper quartiles of this data set. Also find the 10th, 15th, and 65th percentiles. What is the interquartile range?

1–14. The following data are annualized returns on a group of 15 stocks.

12.5, 13, 14.8, 11, 16.7, 9, 8.3, −1.2, 3.9, 15.5, 16.2, 18, 11.6, 10, 9.5

Find the median, the first and third quartiles, and the 55th and 85th percentiles for these data.

1–15. The following data are the total 1-year return, in percent, for 10 midcap mutual funds:[5]

0.7, 0.8, 0.1, −0.7, −0.7, 1.6, 0.2, −0.5, −0.4, −1.3

Find the median and the 20th, 30th, 60th, and 90th percentiles.

1–16. Following are the numbers of daily bids received by the government of a developing country from firms interested in winning a contract for the construction of a new port facility.

2, 3, 2, 4, 3, 5, 1, 1, 6, 4, 7, 2, 5, 1, 6

Find the quartiles and the interquartile range. Also find the 60th percentile.

1–17. Find the median, the interquartile range, and the 45th percentile of the following data.

23, 26, 29, 30, 32, 34, 37, 45, 57, 80, 102, 147, 210, 355, 782, 1,209

1–3 Measures of Central Tendency

CHAPTER 1

Percentiles, and in particular quartiles, are measures of the relative positions of points within a data set or a population (when our data set constitutes the entire population). The median is a special point, since it lies in the center of the data in the sense that half the data lie below it and half above it. The median is thus a measure of the *location* or *centrality* of the observations.

In addition to the median, two other measures of central tendency are commonly used. One is the *mode* (or modes—there may be several of them), and the other is the *arithmetic mean,* or just the *mean*. We define the mode as follows.

The **mode** of the data set is the value that occurs most frequently.

Let us look at the frequencies of occurrence of the data values in Example 1–2, shown in Table 1–2. We see that the value 18 occurs most frequently. Four data points have this value—more points than for any other value in the data set. Therefore, the mode is equal to 18.

The most commonly used measure of central tendency of a set of observations is the mean of the observations.

TABLE 1–2 Frequencies of Occurrence of Data Values in Example 1–2

Value	Frequency
18	4
19	1
20	3
21	1
22	2
23	1
24	1
26	1
27	1
32	1
33	1
49	1
52	1
56	1

The **mean** of a set of observations is their **average.** It is equal to the sum of all observations divided by the number of observations in the set.

Let us denote the observations by $x_1, x_2, \ldots x_n$. That is, the first observation is denoted by x_1, the second by x_2, and so on to the nth observation, x_n. (In Example 1–2, $x_1 = 33$, $x_2 = 26, \ldots$, and $x_n = x_{20} = 18$.) The sample mean is denoted by $\bar{x}$

Mean of a sample:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad (1\text{–}1)$$

where Σ is summation notation. The summation extends over all data points.

[5]"The Money 70," *Money,* March 2007, p. 63.

When our observation set constitutes an entire population, instead of denoting the mean by $\bar{x}$ we use the symbol μ (the Greek letter mu). For a population, we use N as the number of elements instead of n. The population mean is defined as follows.

Mean of a population:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1–2}$$

The mean of the observations in Example 1–2 is found as

$$\begin{aligned}\bar{x} &= (x_1 + x_2 + \cdots + x_{20})/20 = (33 + 26 + 24 + 21 + 19 \\ &\quad + 20 + 18 + 18 + 52 + 56 + 27 + 22 + 18 + 49 + 22 \\ &\quad + 20 + 23 + 32 + 20 + 18)/20 \\ &= 538/20 = 26.9\end{aligned}$$

The mean of the observations of Example 1–2, their average, is 26.9.

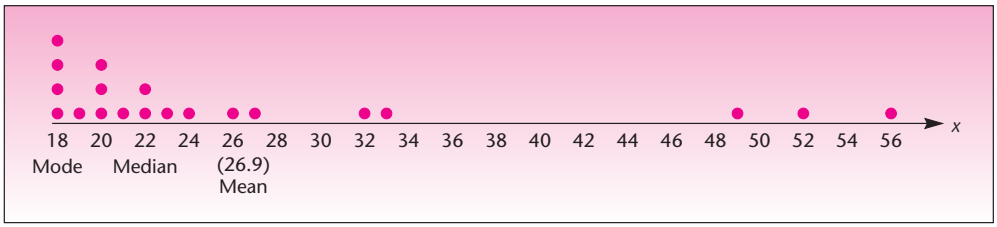
Figure 1–1 shows the data of Example 1–2 drawn on the number line along with the mean, median, and mode of the observations. If you think of the data points as little balls of equal weight located at the appropriate places on the number line, the mean is that point where all the weights balance. It is the *fulcrum* of the point-weights, as shown in Figure 1–1.

What characterizes the three measures of centrality, and what are the relative merits of each? The mean summarizes all the information in the data. It is the average of all the observations. The mean is a single point that can be viewed as the point where all the mass–the weight–of the observations is concentrated. It is the center of mass of the data. If all the observations in our data set were the same size, then (assuming the total is the same) each would be equal to the mean.

The median, on the other hand, is an observation (or a point between two observations) in the center of the data set. One-half of the data lie above this observation, and one-half of the data lie below it. When we compute the median, we do not consider the exact location of each data point on the number line; we only consider whether it falls in the half lying above the median or in the half lying below the median.

What does this mean? If you look at the picture of the data set of Example 1–2, Figure 1–1, you will note that the observation $x_{10} = 56$ lies to the far right. If we shift this particular observation (or any other observation to the right of 22) to the right, say, move it from 56 to 100, what will happen to the median? The answer is: absolutely *nothing* (prove this to yourself by calculating the new median). The exact location of any data point is not considered in the computation of the median, only

FIGURE 1–1 **Mean, Median, and Mode for Example 1–2**



its relative standing with respect to the central observation. *The median is resistant to extreme observations.*

The mean, on the other hand, is sensitive to extreme observations. Let us see what happens to the mean if we change x_{10} from 56 to 100. The new mean is

$$\begin{aligned}\bar{x} &= (33 + 26 + 24 + 21 + 19 + 20 + 18 + 18 + 52 + 100 + 27 \\ &\quad + 22 + 18 + 49 + 22 + 20 + 23 + 32 + 20 + 18)/20 \\ &= 29.1\end{aligned}$$

We see that the mean has shifted 2.2 units to the right to accommodate the change in the single data point x_{10}.

The mean, however, does have strong advantages as a measure of central tendency. *The mean is based on information contained in all the observations in the data set,* rather than being an observation lying "in the middle" of the set. The mean also has some desirable mathematical properties that make it useful in many contexts of statistical inference. In cases where we want to guard against the influence of a few outlying observations (called *outliers*), however, we may prefer to use the median.

EXAMPLE 1–4

To continue with the condominium prices from Example 1–1, a larger sample of asking prices for two-bedroom units in Boston (numbers in thousand dollars, rounded to the nearest thousand) is

789, 813, 980, 880, 650, 700, 2,990, 850, 690

What are the mean and the median? Interpret their meaning in this case.

Solution

Arranging the data from smallest to largest, we get

650, 690, 700, 789, 813, 850, 880, 980, 2,990

There are nine observations, so the median is the value in the middle, that is, in the fifth position. That value is 813 thousand dollars.

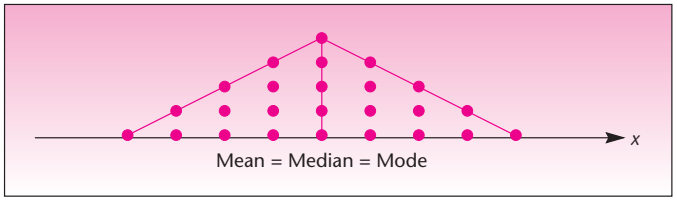
To compute the mean, we add all data values and divide by 9, giving 1,038 thousand dollars—that is, $1,038,000. Now notice some interesting facts. The value 2,990 is clearly an *outlier.* It lies far to the right, away from the rest of the data bunched together in the 650–980 range.

In this case, the median is a very descriptive measure of this data set: it tells us where our data (with the exception of the outlier) are located. The mean, on the other hand, pays so much attention to the large observation 2,990 that it locates itself at 1,038, a value larger than our largest observation, except for the outlier. If our outlier had been more like the rest of the data, say, 820 instead of 2,990, the mean would have been 796.9. Notice that the median does not change and is still 813. This is so because 820 is on the same side of the median as 2,990.

Sometimes an outlier is due to an error in recording the data. In such a case it should be removed. Other times it is "out in left field" (actually, right field in this case) for good reason.

As it turned out, the condominium with asking price of $2,990,000 was quite different from the rest of the two-bedroom units of roughly equal square footage and location. This unit was located in a prestigious part of town (away from the other units, geographically as well). It had a large whirlpool bath adjoining the master bedroom; its floors were marble from the Greek island of Paros; all light fixtures and faucets were gold-plated; the chandelier was Murano crystal. "This is not your average condominium," the realtor said, inadvertently reflecting a purely statistical fact in addition to the intended meaning of the expression.

FIGURE 1–2 A Symmetrically Distributed Data Set



The mode tells us our data set's most frequently occurring value. There may be several modes. In Example 1–2, our data set actually possesses three modes: 18, 20, and 22. Of the three measures of central tendency, we are most interested in the mean.

If a data set or population is *symmetric* (i.e., if one side of the distribution of the observations is a mirror image of the other) and if the distribution of the observations has only one mode, then the mode, the median, and the mean are all equal. Such a situation is demonstrated in Figure 1–2. Generally, when the data distribution is not symmetric, then the mean, median, and mode will not all be equal. The relative positions of the three measures of centrality in such situations will be discussed in section 1–6.

In the next section, we discuss measures of variability of a data set or population.

PROBLEMS

1–18. Discuss the differences among the three measures of centrality.

1–19. Find the mean, median, and mode(s) of the observations in problem 1–13.

1–20. Do the same as problem 1–19, using the data of problem 1–14.

1–21. Do the same as problem 1–19, using the data of problem 1–15.

1–22. Do the same as problem 1–19, using the data of problem 1–16.

1–23. Do the same as problem 1–19, using the observation set in problem 1–17.

1–24. Do the same as problem 1–19 for the data in Example 1–1.

1–25. Find the mean, mode, and median for the data set 7, 8, 8, 12, 12, 12, 14, 15, 20, 47, 52, 54.

1–26. For the following stock price one-year percentage changes, plot the data and identify any outliers. Find the mean and median.[6]

Intel	−6.9%
AT&T	46.5
General Electric	12.1
ExxonMobil	20.7
Microsoft	16.9
Pfizer	17.2
Citigroup	16.5

[6]"Stocks," *Money*, March 2007, p. 128.

1–27. The following data are the median returns on investment, in percent, for 10 industries.[7]

Consumer staples	24.3%
Energy	23.3
Health care	22.1
Financials	21.0
Industrials	19.2
Consumer discretionary	19.0
Materials	18.1
Information technology	15.1
Telecommunication services	11.0
Utilities	10.4

Find the median of these medians and their mean.

1–4 Measures of Variability

CHAPTER 1

Consider the following two data sets.

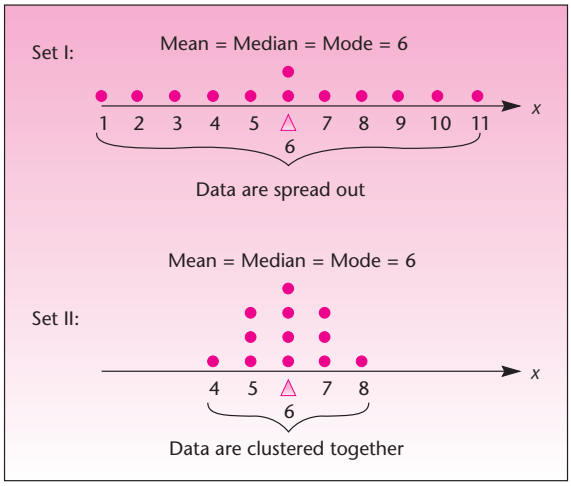
Set I: 1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11
Set II: 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8

Compute the mean, median, and mode of each of the two data sets. As you see from your results, the two data sets have the same mean, the same median, and the same mode, all equal to 6. The two data sets also happen to have the same number of observations, $n = 12$. But the two data sets are different. What is the main difference between them?

Figure 1–3 shows data sets I and II. The two data sets have the same central tendency (as measured by any of the three measures of centrality), but they have a different *variability*. In particular, we see that data set I is more variable than data set II. The values in set I are more spread out: they lie farther away from their mean than do those of set II.

There are several measures of **variability,** or **dispersion.** We have already discussed one such measure—the interquartile range. (Recall that the interquartile range

FIGURE 1–3 Comparison of Data Sets I and II



[7] "Sector Snapshot," *BusinessWeek,* March 26, 2007, p. 62.

is defined as the difference between the upper quartile and the lower quartile.) The interquartile range for data set I is 5.5, and the interquartile range of data set II is 2 (show this). The interquartile range is one measure of the dispersion or variability of a set of observations. Another such measure is the *range*.

The **range** of a set of observations is the difference between the largest observation and the smallest observation.

The range of the observations in Example 1–2 is Largest number − Smallest number = 56 − 18 = 38. The range of the data in set I is 11 − 1 = 10, and the range of the data in set II is 8 − 4 = 4. We see that, conforming with what we expect from looking at the two data sets, the range of set I is greater than the range of set II. Set I is more variable.

The range and the interquartile range are measures of the dispersion of a set of observations, the interquartile range being more resistant to extreme observations. There are also two other, more commonly used measures of dispersion. These are the *variance* and the square root of the variance—the *standard deviation*.

The variance and the standard deviation are more useful than the range and the interquartile range because, like the mean, they use the information contained in all the observations in the data set or population. (The range contains information only on the distance between the largest and smallest observations, and the interquartile range contains information only about the difference between upper and lower quartiles.) We define the variance as follows.

The **variance** of a set of observations is the average squared deviation of the data points from their mean.

When our data constitute a sample, the variance is denoted by s^2, and the averaging is done by dividing the sum of the squared deviations from the mean by $n - 1$. (The reason for this will become clear in Chapter 5.) When our observations constitute an entire population, the variance is denoted by σ^2, and the averaging is done by dividing by N. (And σ is the Greek letter sigma; we call the variance *sigma squared*. The capital sigma is known to you as the symbol we use for summation, Σ.)

Sample variance:

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

Recall that $\bar{x}$ is the sample mean, the average of all the observations in the sample. Thus, the numerator in equation 1–3 is equal to the sum of the squared differences of the data points x_i (where $i = 1, 2, \ldots, n$) from their mean $\bar{x}$. When we divide the numerator by the denominator $n - 1$, we get a kind of average of the items summed in the numerator. This average is based on the assumption that there are only $n - 1$ data points. (Note, however, that the summation in the numerator extends over all n data points, not just $n - 1$ of them.) This will be explained in section 5–5.

When we have an entire population at hand, we denote the total number of observations in the population by N. We define the population variance as follows.

Population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad (1\text{–}4)$$

where μ is the population mean.

Unless noted otherwise, we will assume that all our data sets are samples and do not constitute entire populations; thus, we will use equation 1–3 for the variance, and not equation 1–4. We now define the standard deviation.

The **standard deviation** of a set of observations is the (positive) square root of the variance of the set.

The standard deviation of a sample is the square root of the sample variance, and the standard deviation of a population is the square root of the variance of the population.[8]

Sample standard deviation:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad (1\text{–}5)$$

Population standard deviation:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} \qquad (1\text{–}6)$$

Why would we use the standard deviation when we already have its square, the variance? The standard deviation is a more meaningful measure. The variance is the average squared deviation from the mean. It is squared because if we just compute the deviations from the mean and then averaged them, we get zero (prove this with any of the data sets). Therefore, when seeking a measure of the variation in a set of observations, we square the deviations from the mean; this removes the negative signs, and thus the measure is not equal to zero. The measure we obtain—the variance—is still a *squared* quantity; it is an average of squared numbers. By taking its square root, we "unsquare" the units and get a quantity denoted in the original units of the problem (e.g., dollars instead of dollars squared, which would have little meaning in most applications). The variance tends to be large because it is in squared units. Statisticians like to work with the variance because its mathematical properties simplify computations. People applying statistics prefer to work with the standard deviation because it is more easily interpreted.

Let us find the variance and the standard deviation of the data in Example 1–2. We carry out hand computations of the variance by use of a table for convenience. After doing the computation using equation 1–3, we will show a shortcut that will help in the calculation. Table 1–3 shows how the mean $\bar{x}$ is subtracted from each of the values and the results are squared and added. At the bottom of the last column we find the sum of all squared deviations from the mean. Finally, the sum is divided by $n - 1$, giving s^2, the sample variance. Taking the square root gives us s, the sample standard deviation.

[8]A note about calculators: If your calculator is designed to compute means and standard deviations, find the key for the standard deviation. Typically, there will be two such keys. Consult your owner's handbook to be sure you are using the key that will produce the correct computation for a sample (division by $n - 1$) versus a population (division by N).

TABLE 1–3 Calculations Leading to the Sample Variance in Example 1–2

x	$x - \bar{x}$	$(x - \bar{x})^2$
18	$18 - 26.9 = -8.9$	79.21
18	$18 - 26.9 = -8.9$	79.21
18	$18 - 26.9 = -8.9$	79.21
18	$18 - 26.9 = -8.9$	79.21
19	$19 - 26.9 = -7.9$	62.41
20	$20 - 26.9 = -6.9$	47.61
20	$20 - 26.9 = -6.9$	47.61
20	$20 - 26.9 = -6.9$	47.61
21	$21 - 26.9 = -5.9$	34.81
22	$22 - 26.9 = -4.9$	24.01
22	$22 - 26.9 = -4.9$	24.01
23	$23 - 26.9 = -3.9$	15.21
24	$24 - 26.9 = -2.9$	8.41
26	$26 - 26.9 = -0.9$	0.81
27	$27 - 26.9 = 0.1$	0.01
32	$32 - 26.9 = 5.1$	26.01
33	$33 - 26.9 = 6.1$	37.21
49	$49 - 26.9 = 22.1$	488.41
52	$52 - 26.9 = 25.1$	630.01
56	$56 - 26.9 = 29.1$	846.81
	0	2,657.8

By equation 1–3, the variance of the sample is equal to the sum of the third column in the table, 2,657.8, divided by $n - 1$: $s^2 = 2{,}657.8/19 = 139.88421$. The standard deviation is the square root of the variance: $s = \sqrt{139.88421} = 11.827266$, or, using two-decimal accuracy,[9] $s = 11.83$.

If you have a calculator with statistical capabilities, you may avoid having to use a table such as Table 1–3. If you need to compute by hand, there is a shortcut formula for computing the variance and the standard deviation.

Shortcut formula for the sample variance:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n}{n - 1} \qquad (1\text{–}7)$$

Again, the standard deviation is just the square root of the quantity in equation 1–7. We will now demonstrate the use of this computationally simpler formula with the data of Example 1–2. We will then use this simpler formula and compute the variance and the standard deviation of the two data sets we are comparing: set I and set II.

As before, a table will be useful in carrying out the computations. The table for finding the variance using equation 1–7 will have a column for the data points x and

[9] In quantitative fields such as statistics, decimal accuracy is always a problem. How many digits after the decimal point should we carry? This question has no easy answer; everything depends on the required level of accuracy. As a rule, we will use only two decimals, since this suffices in most applications in this book. In some procedures, such as regression analysis, more digits need to be used in computations (these computations, however, are usually done by computer).

TABLE 1–4 Shortcut Computations for the Variance in Example 1–2

x	x^2
18	324
18	324
18	324
18	324
19	361
20	400
20	400
20	400
21	441
22	484
22	484
23	529
24	576
26	676
27	729
32	1,024
33	1,089
49	2,401
52	2,704
56	3,136
538	17,130

a column for the squared data points x^2. Table 1–4 shows the computations for the variance of the data in Example 1–2.

Using equation 1–7, we find

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n}{n-1} = \frac{17{,}130 - (538)^2/20}{19} = \frac{17{,}130 - 289{,}444/20}{19} = 139.88421$$

The standard deviation is obtained as before: $s = \sqrt{139.88421} = 11.83$. Using the same procedure demonstrated with Table 1–4, we find the following quantities leading to the variance and the standard deviation of set I and of set II. Both are assumed to be samples, not populations.

Set I: $\Sigma x = 72$, $\Sigma x^2 = 542$, $s^2 = 10$, and $s = \sqrt{10} = 3.16$

Set II: $\Sigma x = 72$, $\Sigma x^2 = 446$, $s^2 = 1.27$, and $s = \sqrt{1.27} = 1.13$

As expected, we see that the variance and the standard deviation of set II are smaller than those of set I. While each has a mean of 6, set I is more variable. That is, the values in set I vary more about their mean than do those of set II, which are clustered more closely together.

The sample standard deviation and the sample mean are very important statistics used in inference about populations.

EXAMPLE 1–5

In financial analysis, the standard deviation is often used as a measure of *volatility* and of the *risk* associated with financial variables. The data below are exchange rate values of the British pound, given as the value of one U.S. dollar's worth in pounds. The first column of 10 numbers is for a period in the beginning of 1995, and the second column of 10 numbers is for a similar period in the beginning of 2007.[10] During which period, of these two precise sets of 10 days each, was the value of the pound more volatile?

1995	2007
0.6332	0.5087
0.6254	0.5077
0.6286	0.5100
0.6359	0.5143
0.6336	0.5149
0.6427	0.5177
0.6209	0.5164
0.6214	0.5180
0.6204	0.5096
0.6325	0.5182

Solution We are looking at two *populations* of 10 specific days at the start of each year (rather than a random sample of days), so we will use the formula for the population standard deviation. For the 1995 period we get $\sigma = 0.007033$. For the 2007 period we get $\sigma = 0.003938$. We conclude that during the 1995 ten-day period the British pound was

[10]From data reported in "Business Day," *The New York Times*, in March 2007, and from Web information.

more volatile than in the same period in 2007. Notice that if these had been random samples of days, we would have used the sample standard deviation. In such cases we might have been interested in statistical inference to some population.

EXAMPLE 1–6

The data for second quarter earnings per share (EPS) for major banks in the Northeast are tabulated below. Compute the mean, the variance, and the standard deviation of the data.

Name	EPS
Bank of New York	$2.53
Bank of America	4.38
Banker's Trust/New York	7.53
Chase Manhattan	7.53
Citicorp	7.96
Brookline	4.35
MBNA	1.50
Mellon	2.75
Morgan JP	7.25
PNC Bank	3.11
Republic	7.44
State Street	2.04
Summit	3.25

Solution

$$\sum x = \$61.62; \quad \bar{x} = \$4.74; \quad \sum x^2 = 363.40;$$
$$s^2 = 5.94; \quad s = \$2.44.$$

Figure 1–4 shows how Excel commands can be used for obtaining a group of the most useful and common descriptive statistics using the data of Example 1–2. In section 1–10, we will see how a complete set of descriptive statistics can be obtained from a spreadsheet template.

FIGURE 1–4 Using Excel for Example 1–2

	A	B	C	D	E	F	G	H
1								
2		Wealth ($billion)						
3		33						
4		26						
5		24						
6		21						
7		19						
8		20						
9		18						
10		18						
11		52						
12		56						
13		27						
14		22						
15		18						
16		49						
17		22						
18		20						
19		23						
20		32						
21		20						
22		18						
23								

Descriptive Statistics	Excel Command	Result
Mean	=AVERAGE(A3:A22)	26.9
Median	=MEDIAN(A3:A22)	22
Mode	=MODE(A3:A22)	18
Standard Deviation	=STDEV(A3:A22)	11.8272656
Standard Error	=F11/SQRT(20)	2.64465698
Kurtosis	=KURT(A3:A22)	1.60368514
Skewness	=SKEW(A3:A22)	1.65371559
Range	=MAX(A3:A22)-MIN(A3:A22)	38
Minimum	=MIN(A3:A22)	18
Maximum	=MAX(A3:A22)	56
Sum	=SUM(A3:A22)	538
Count	=COUNT(A3:A22)	20

PROBLEMS

1–28. Explain why we need measures of variability and what information these measures convey.

1–29. What is the most important measure of variability and why?

1–30. What is the computational difference between the variance of a sample and the variance of a population?

1–31. Find the range, the variance, and the standard deviation of the data set in problem 1–13 (assumed to be a sample).

1–32. Do the same as problem 1–31, using the data in problem 1–14.

1–33. Do the same as problem 1–31, using the data in problem 1–15.

1–34. Do the same as problem 1–31, using the data in problem 1–16.

1–35. Do the same as problem 1–31, using the data in problem 1–17.

1–5 Grouped Data and the Histogram

Data are often grouped. This happened naturally in Example 1–2, where we had a group of four points with a value of 18, a group of three points with a value of 20, and a group of two points with a value of 22. In other cases, especially when we have a large data set, the collector of the data may break the data into groups even if the points in each group are not equal in value. The data collector may set some (often arbitrary) group boundaries for ease of recording the data. When the salaries of 5,000 executives are considered, for example, the data may be reported in the form: 1,548 executives in the salary range \$60,000 to \$65,000; 2,365 executives in the salary range \$65,001 to \$70,000; and so on. In this case, the data collector or analyst has processed all the salaries and put them into groups with defined boundaries. In such cases, there is a loss of information. We are unable to find the mean, variance, and other measures because we do not know the actual values. (Certain formulas, however, allow us to find the approximate mean, variance, and standard deviation. The formulas assume that all data points in a group are placed in the midpoint of the interval.) In this example, we assume that all 1,548 executives in the \$60,000–\$65,000 *class* make exactly (\$60,000 + \$65,000)/2 = \$62,500; we estimate similarly for executives in the other groups.

We define a group of data values within specified group boundaries as a **class.**

When data are grouped into classes, we may also plot a frequency distribution of the data. Such a frequency plot is called a *histogram.*

A **histogram** is a chart made of bars of different heights. The height of each bar represents the **frequency** of values in the class represented by the bar. Adjacent bars share sides.

We demonstrate the use of histograms in the following example. Note that a histogram is used only for measured, or ordinal, data.

EXAMPLE 1–7

Management of an appliance store recorded the amounts spent at the store by the 184 customers who came in during the last day of the big sale. The data, amounts spent, were grouped into categories as follows: \$0 to less than \$100, \$100 to less than \$200, and so on up to \$600, a bound higher than the amount spent by any single buyer. The classes and the frequency of each class are shown in Table 1–5. The frequencies, denoted by $f(x)$, are shown in a histogram in Figure 1–5.

TABLE 1–5 Classes and Frequencies for Example 1–7

x Spending Class ($)	*f*(*x*) Frequency (Number of Customers)
0 to less than 100	30
100 to less than 200	38
200 to less than 300	50
300 to less than 400	31
400 to less than 500	22
500 to less than 600	13
	184

FIGURE 1–5 A Histogram of the Data in Example 1–7

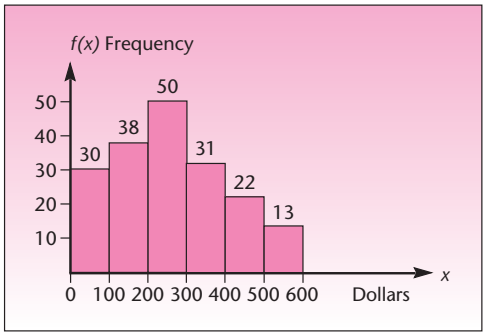


TABLE 1–6 Relative Frequencies for Example 1–7

x Class ($)	*f*(*x*) Relative Frequency
0 to less than 100	0.163
100 to less than 200	0.207
200 to less than 300	0.272
300 to less than 400	0.168
400 to less than 500	0.120
500 to less than 600	0.070
	1.000

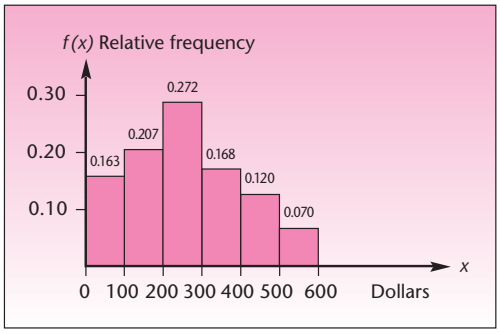
As you can see from Figure 1–5, a histogram is just a convenient way of plotting the frequencies of grouped data. Here the frequencies are *absolute frequencies* or **counts** of data points. It is also possible to plot *relative frequencies*.

The **relative frequency** of a class is the count of data points in the class divided by the total number of data points.

Solution The relative frequency in the first class, $0 to less than $100, is equal to count/total = 30/184 = 0.163. We can similarly compute the relative frequencies for the other classes. The advantage of relative frequencies is that they are standardized: They add to 1.00. The relative frequency in each class represents the proportion of the total sample in the class. Table 1–6 gives the relative frequencies of the classes.

Figure 1–6 is a histogram of the relative frequencies of the data in this example. Note that the shape of the histogram of the relative frequencies is the same as that of

FIGURE 1–6 A Histogram of the Relative Frequencies in Example 1–7



the absolute frequencies, the counts. The shape of the histogram does not change; only the labeling of the $f(x)$ axis is different.

Relative frequencies—proportions that add to 1.00—may be viewed as probabilities, as we will see in the next chapter. Hence, such frequencies are very useful in statistics, and so are their histograms.

CHAPTER 1
CHAPTER 3

1–6 Skewness and Kurtosis

In addition to measures of location, such as the mean or median, and measures of variation, such as the variance or standard deviation, two more attributes of a frequency distribution of a data set may be of interest to us. These are *skewness* and *kurtosis*.

Skewness is a measure of the degree of asymmetry of a frequency distribution.

When the distribution stretches to the right more than it does to the left, we say that the distribution is *right skewed*. Similarly, a *left-skewed* distribution is one that stretches asymmetrically to the left. Four graphs are shown in Figure 1–7: a symmetric distribution, a right-skewed distribution, a left-skewed distribution, and a symmetrical distribution with two modes.

Recall that a symmetric distribution with a single mode has mode = mean = median. Generally, for a right-skewed distribution, the mean is to the right of the median, which in turn lies to the right of the mode (assuming a single mode). The opposite is true for left-skewed distributions.

Skewness is calculated[11] and reported as a number that may be positive, negative, or zero. *Zero skewness* implies a symmetric distribution. A *positive skewness* implies a right-skewed distribution, and a *negative skewness* implies a left-skewed distribution.

Two distributions that have the same mean, variance, and skewness could still be significantly different in their shape. We may then look at their kurtosis.

Kurtosis is a measure of the peakedness of a distribution.

The larger the kurtosis, the more peaked will be the distribution. The kurtosis is calculated[12] and reported either as an absolute or a relative value. *Absolute kurtosis* is

[11] The formula used for calculating the skewness of a population is $\sum_{i=1}^{N}\left[\frac{x_i-\mu}{\sigma}\right]^3 \Big/ N.$

[12] The formula used for calculating the absolute kurtosis of a population is $\sum_{i=1}^{N}\left[\frac{x_i-\mu}{\sigma}\right]^4 \Big/ N.$

FIGURE 1–7 Skewness of Distributions

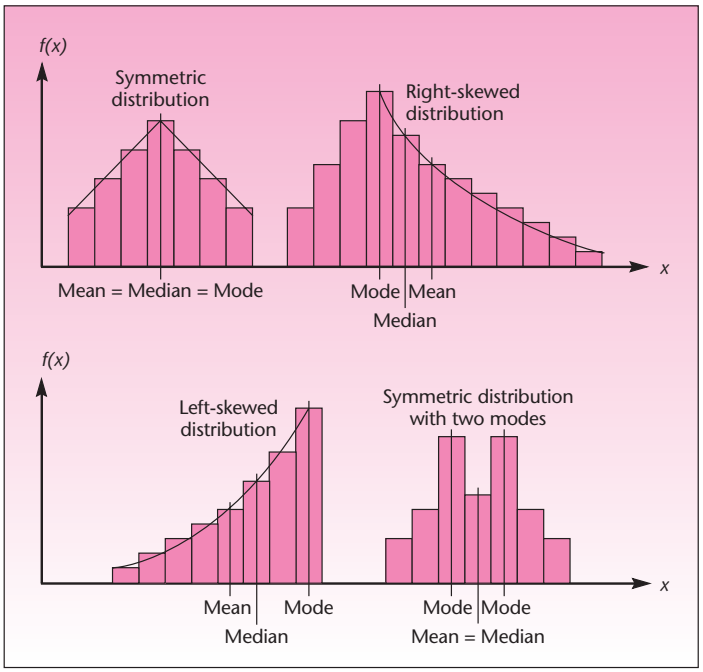
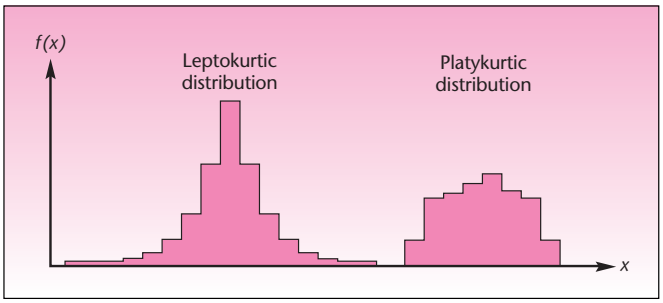


FIGURE 1–8 Kurtosis of Distributions



always a positive number. The absolute kurtosis of a *normal distribution,* a famous distribution about which we will learn in Chapter 4, is 3. This value of 3 is taken as the datum to calculate the *relative kurtosis*. The two are related by the equation

$$\text{Relative kurtosis} = \text{Absolute kurtosis} - 3$$

The relative kurtosis can be negative. We will always work with relative kurtosis. As a result, in this book, "kurtosis" means "relative kurtosis."

A negative kurtosis implies a flatter distribution than the normal distribution, and it is called *platykurtic*. A positive kurtosis implies a more peaked distribution than the normal distribution, and it is called *leptokurtic*. Figure 1–8 shows these examples.

1–7 Relations between the Mean and the Standard Deviation

The mean is a measure of the centrality of a set of observations, and the standard deviation is a measure of their spread. There are two general rules that establish a relation between these measures and the set of observations. The first is called Chebyshev's theorem, and the second is the empirical rule.

Chebyshev's Theorem

A mathematical theorem called **Chebyshev's theorem** establishes the following rules:

1. At least three-quarters of the observations in a set will lie within 2 standard deviations of the mean.
2. At least eight-ninths of the observations in a set will lie within 3 standard deviations of the mean.

In general, the rule states that at least $1 - 1/k^2$ of the observations will lie within k standard deviations of the mean. (We note that k does not have to be an integer.) In Example 1–2 we found that the mean was 26.9 and the standard deviation was 11.83. According to rule 1 above, at least three-quarters of the observations should fall in the interval Mean $\pm$ $2s = 26.9 \pm 2(11.83)$, which is defined by the points 3.24 and 50.56. From the data set itself, we see that all but the three largest data points lie within this range of values. Since there are 20 observations in the set, seventeen-twentieths are within the specified range, so the rule that at least three-quarters will be within the range is satisfied.

The Empirical Rule

If the distribution of the data is mound-shaped—that is, if the histogram of the data is more or less symmetric with a single mode or high point—then tighter rules will apply. This is the **empirical rule:**

1. Approximately 68% of the observations will be within 1 standard deviation of the mean.
2. Approximately 95% of the observations will be within 2 standard deviations of the mean.
3. A vast majority of the observations (all, or almost all) will be within 3 standard deviations of the mean.

Note that Chebyshev's theorem states *at least* what percentage will lie within k standard deviations in any distribution, whereas the empirical rule states *approximately* what percentage will lie within k standard deviations in a *mound-shaped* distribution.

For the data set in Example 1–2, the distribution of the data set is not symmetric, and the empirical rule holds only approximately.

PROBLEMS

1–36. Check the applicability of Chebyshev's theorem and the empirical rule for the data set in problem 1–13.

1–37. Check the applicability of Chebyshev's theorem and the empirical rule for the data set in problem 1–14.

1–38. Check the applicability of Chebyshev's theorem and the empirical rule for the data set in problem 1–15.

1–39. Check the applicability of Chebyshev's theorem and the empirical rule for the data set in problem 1–16.

1–40. Check the applicability of Chebyshev's theorem and the empirical rule for the data set in problem 1–17.

1–8 Methods of Displaying Data

In section 1–5, we saw how a histogram is used to display frequencies of occurrence of values in a data set. In this section, we will see a few other ways of displaying data, some of which are descriptive only. We will introduce frequency polygons, cumulative frequency plots (called *ogives*), pie charts, and bar charts. We will also see examples of how descriptive graphs can sometimes be misleading. We will start with pie charts.

Pie Charts

A **pie chart** is a simple descriptive display of data that sum to a given total. A pie chart is probably the most illustrative way of displaying quantities as percentages of a given total. The total area of the pie represents 100% of the quantity of interest (the sum of the variable values in all categories), and the size of each slice is the percentage of the total represented by the category the slice denotes. Pie charts are used to present frequencies for categorical data. The scale of measurement may be nominal or ordinal. Figure 1–9 is a pie chart of the percentages of all kinds of investments in a typical family's portfolio.

Bar Charts

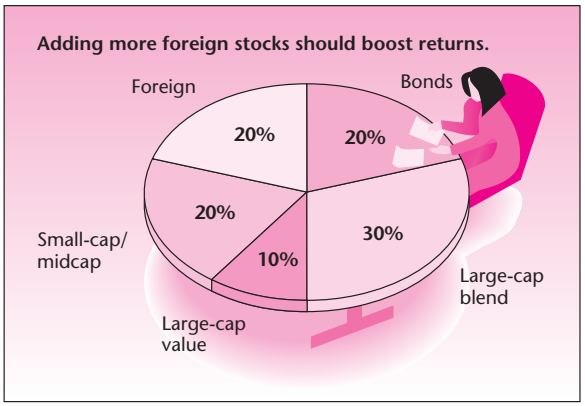
Bar charts (which use horizontal or vertical rectangles) are often used to display categorical data where there is no emphasis on the percentage of a total represented by each category. The scale of measurement is nominal or ordinal.

Charts using horizontal bars and those using vertical bars are essentially the same. In some cases, one may be more convenient than the other for the purpose at hand. For example, if we want to write the name of each category inside the rectangle that represents that category, then a horizontal bar chart may be more convenient. If we want to stress the height of the different columns as measures of the quantity of interest, we use a vertical bar chart. Figure 1–10 is an example of how a bar chart can be used effectively to display and interpret information.

Frequency Polygons and Ogives

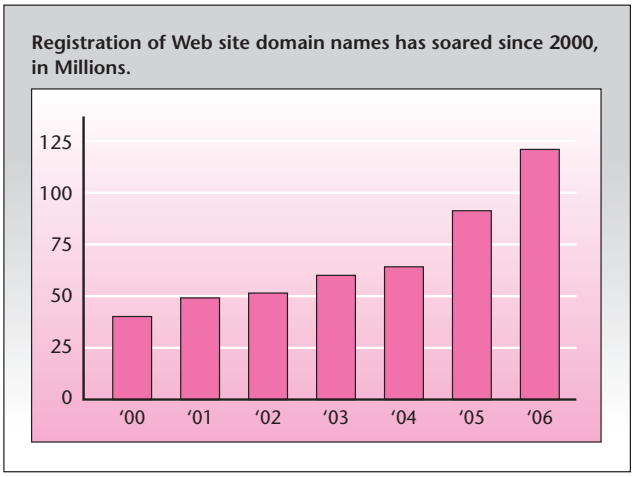
A **frequency polygon** is similar to a histogram except that there are no rectangles, only a point in the midpoint of each interval at a height proportional to the frequency

FIGURE 1–9 **Investments Portfolio Composition**



Source: Carolyn Bigda, "The Fast Track to Kicking Back," *Money*, March 2007, p. 60.

FIGURE 1–10 The Web Takes Off

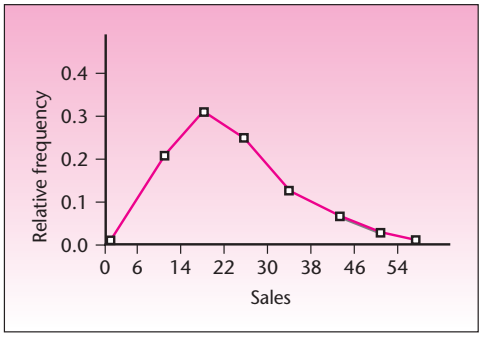


Source: S. Hammand and M. Tucker, "How Secure Is Your Domain," *BusinessWeek*, March 26, 2007, p. 118.

TABLE 1–7 Pizza Sales

Sales ($000)	Relative Frequency
6–14	0.20
15–22	0.30
23–30	0.25
31–38	0.15
39–46	0.07
47–54	0.03

FIGURE 1–11 Relative-Frequency Polygon for Pizza Sales



or relative frequency (in a relative-frequency polygon) of the category of the interval. The rightmost and leftmost points are zero. Table 1–7 gives the relative frequency of sales volume, in thousands of dollars per week, for pizza at a local establishment.

A relative-frequency polygon for these data is shown in Figure 1–11. Note that the frequency is located in the middle of the interval as a point with height equal to the relative frequency of the interval. Note also that the point zero is added at the left

FIGURE 1–12 Excel-Produced Graph of the Data in Example 1–2

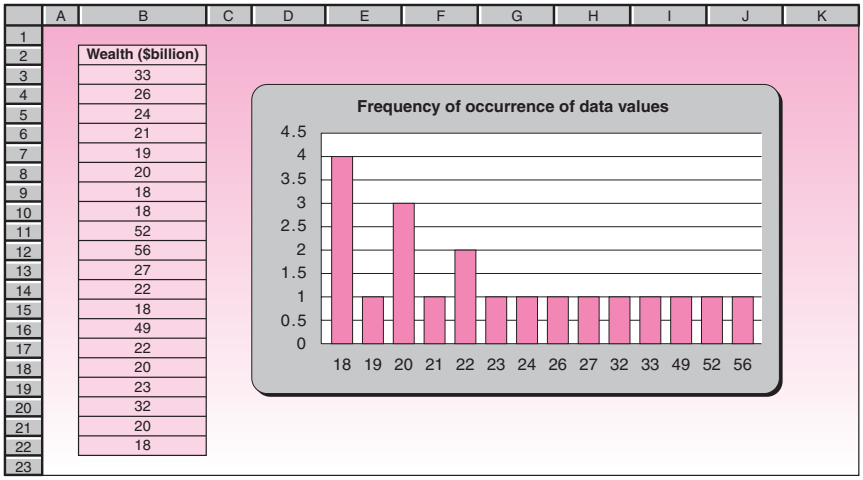
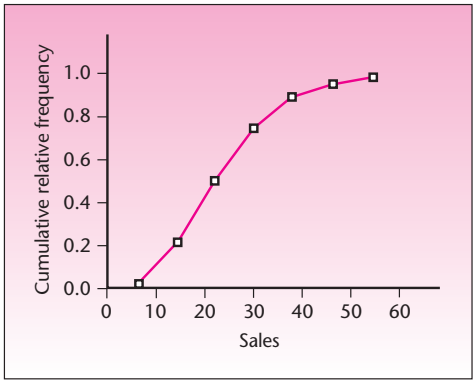


FIGURE 1–13 Ogive of Pizza Sales



boundary and the right boundary of the data set: The polygon starts at zero and ends at zero relative frequency.

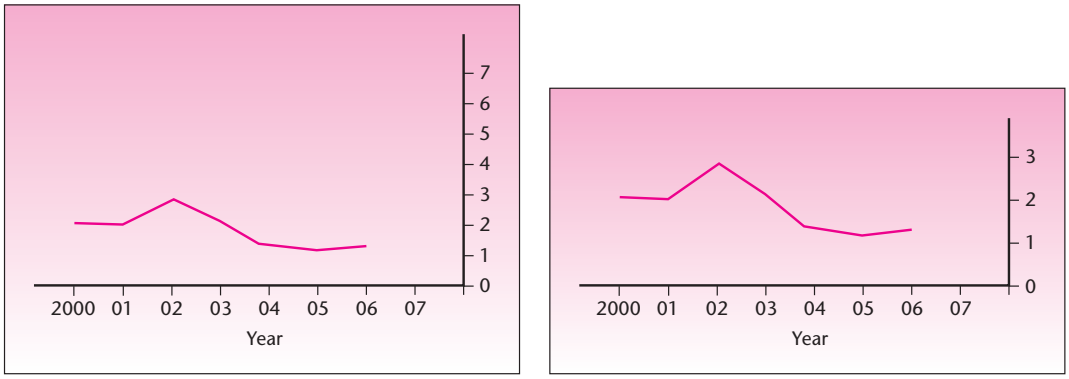
Figure 1–12 shows the worth of the 20 richest individuals from Example 1–2 displayed as a column chart. This is done using Excel's Chart Wizard.

An **ogive** is a cumulative-frequency (or cumulative relative-frequency) graph. An ogive starts at 0 and goes to 1.00 (for a relative-frequency ogive) or to the maximum cumulative frequency. The point with height corresponding to the cumulative frequency is located at the right endpoint of each interval. An ogive for the data in Table 1–7 is shown in Figure 1–13. While the ogive shown is for the cumulative *relative* frequency, an ogive can also be used for the cumulative absolute frequency.

A Caution about Graphs

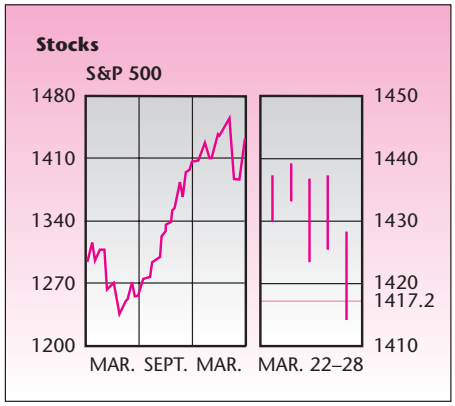
A picture is indeed worth a thousand words, but pictures can sometimes be deceiving. Often, this is where "lying with statistics" comes in: presenting data graphically on a stretched or compressed scale of numbers with the aim of making the data show whatever you want them to show. This is one important argument against a

FIGURE 1–14 German Wage Increases (%)



Source: "Economic Focus," *The Economist*, March 3, 2007, p. 82. Reprinted by permission.

FIGURE 1–15 The S&P 500, One Year, to March 2007



Source: Adapted from "Economic Focus," *The Economist,* March 3, 2007, p. 82.

merely descriptive approach to data analysis and an argument for statistical *inference*. Statistical tests tend to be more objective than our eyes and are less prone to deception as long as our assumptions (random sampling and other assumptions) hold. As we will see, statistical inference gives us tools that allow us to objectively evaluate what we see in the data.

Pictures are sometimes deceptive even though there is no intention to deceive. When someone shows you a graph of a set of numbers, there may really be no particular scale of numbers that is "right" for the data.

The graph on the left in Figure 1–14 is reprinted from *The Economist*. Notice that there is *no scale* that is the "right" one for this graph. Compare this graph with the one on the right side, which has a different scale.

Time Plots

Often we want to graph changes in a variable over time. An example is given in Figure 1–15.

PROBLEMS

1–41. The following data are estimated worldwide appliance sales (in millions of dollars). Use the data to construct a pie chart for the worldwide appliance sales of the listed manufacturers.

Electrolux	$5,100
General Electric	4,350
Matsushita Electric	4,180
Whirlpool	3,950
Bosch-Siemens	2,200
Philips	2,000
Maytag	1,580

1–42. Draw a bar graph for the data on the first five stocks in problem 1–14. Is any one of the three kinds of plot more appropriate than the others for these data? If so, why?

1–43. Draw a bar graph for the endowments (stated in billions of dollars) of each of the universities specified in the following list.

Harvard	$3.4
Texas	2.5
Princeton	1.9
Yale	1.7
Stanford	1.4
Columbia	1.3
Texas A&M	1.1

1–44. The following are the top 10 private equity deals of all time, in billions of dollars.[13]

38.9, 32.7, 31.1, 27.4, 25.7, 21.6, 17.6, 17.4, 15.0, 13.9

Find the mean, median, and standard deviation. Draw a bar graph.

1–45. The following data are credit default swap values:[14] 6, 10, 12, 13, 18, 21 (in trillions of dollars). Draw a pie chart of these amounts. Find the mean and median.

1–46. The following are the amounts from the sales slips of a department store (in dollars): 3.45, 4.52, 5.41, 6.00, 5.97, 7.18, 1.12, 5.39, 7.03, 10.25, 11.45, 13.21, 12.00, 14.05, 2.99, 3.28, 17.10, 19.28, 21.09, 12.11, 5.88, 4.65, 3.99, 10.10, 23.00, 15.16, 20.16. Draw a frequency polygon for these data (start by defining intervals of the data and counting the data points in each interval). Also draw an ogive and a column graph.

1–9 Exploratory Data Analysis

Exploratory data analysis (EDA) is the name given to a large body of statistical and graphical techniques. These techniques provide ways of looking at data to determine relationships and trends, identify outliers and influential observations, and quickly describe or summarize data sets. Pioneering methods in this field, as well as the name *exploratory data analysis,* derive from the work of John W. Tukey [John W. Tukey, *Exploratory Data Analysis* (Reading, Massachusetts: Addison-Wesley, 1977)].

[13]R. Kirkland, "Private Money," *Fortune,* March 5, 2007, p. 58.

[14]John Ferry, "Gimme Shelter," *Worth,* April 2007, p. 89.

CHAPTER 1

Stem-and-Leaf Displays

A **stem-and-leaf display** is a quick way of looking at a data set. It contains some of the features of a histogram but avoids the loss of information in a histogram that results from aggregating the data into intervals. The stem-and-leaf display is based on the tallying principle: | || ||| |||| 𝍸; but it also uses the decimal base of our number system. In a stem-and-leaf display, the *stem* is the number without its rightmost digit (the *leaf*). The stem is written to the left of a vertical line separating the stem from the leaf. For example, suppose we have the numbers 105, 106, 107, 107, 109. We display them as

```
10 | 56779
```

With a more complete data set with different stem values, the last digit of each number is displayed at the appropriate place to the right of its stem digit(s). Stem-and-leaf displays help us identify, at a glance, numbers in our data set that have high frequency. Let's look at an example.

EXAMPLE 1–8

Virtual reality is the name given to a system of simulating real situations on a computer in a way that gives people the feeling that what they see on the computer screen is a real situation. Flight simulators were the forerunners of virtual reality programs. A particular virtual reality program has been designed to give production engineers experience in real processes. Engineers are supposed to complete certain tasks as responses to what they see on the screen. The following data are the time, in seconds, it took a group of 42 engineers to perform a given task:

11, 12, 12, 13, 15, 15, 15, 16, 17, 20, 21, 21, 21, 22, 22, 22, 23, 24, 26, 27, 27, 27, 28, 29, 29, 30, 31, 32, 34, 35, 37, 41, 41, 42, 45, 47, 50, 52, 53, 56, 60, 62

Use a stem-and-leaf display to analyze these data.

Solution The data are already arranged in increasing order. We see that the data are in the 10s, 20s, 30s, 40s, 50s, and 60s. We will use the first digit as the stem and the second digit of each number as the leaf. The stem-and-leaf display of our data is shown in Figure 1–16.

As you can see, the stem-and-leaf display is a very quick way of arranging the data in a kind of a histogram (turned sideways) that allows us to see what the data look like. Here, we note that the data do not seem to be symmetrically distributed; rather, they are skewed to the right.

FIGURE 1–16 Stem-and-Leaf Display of the Task Performance Times of Example 1–8

```
1 | 122355567
2 | 0111222346777899
3 | 012457
4 | 11257
5 | 0236
6 | 02
```

We may feel that this display does not convey very much information because there are too many values with first digit 2. To solve this problem, we may split the groups into two subgroups. We will denote the stem part as 1* for the possible numbers 10, 11, 12, 13, 14 and as 1. for the possible numbers 15, 16, 17, 18, 19. Similarly, the stem 2* will be used for the possible numbers 20, 21, 22, 23, and 24; stem 2. will be used for the numbers 25, 26, 27, 28, and 29; and so on for the other numbers. Our stem-and-leaf diagram for the data of Example 1–8 using this convention is shown in Figure 1–17. As you can see from the figure, we now have a more spread-out histogram of the data. The data still seem skewed to the right.

If desired, a further refinement of the display is possible by using the symbol * for a stem followed by the leaf values 0 and 1; the symbol t for leaf values 2 and 3; the symbol f for leaf values 4 and 5; s for 6 and 7; and . for 8 and 9. Also, the class containing the median observation is often denoted with its stem value in parentheses.

We demonstrate this version of the display for the data of Example 1–8 in Figure 1–18. Note that the median is 27 (why?).

Note that for the data set of this example, the refinement offered in Figure 1–18 may be too much: We may have lost the general picture of the data. In cases where there are many observations with the same value (for example, 22, 22, 22, 22, 22, 22, 22, . . .), the use of a more stretched-out display may be needed in order to get a good picture of the way our data are clustered.

FIGURE 1–17 **Refined Stem-and-Leaf Display for Data of Example 1–8**

1*	1223
1.	55567
2*	011122234
2.	6777899
3*	0124
3.	57
4*	112
4.	57
5*	023
5.	6
6*	02

CHAPTER 1

Box Plots

A *box plot* (also called a *box-and-whisker plot*) is another way of looking at a data set in an effort to determine its central tendency, spread, skewness, and the existence of outliers.

A **box plot** is a set of five summary measures of the distribution of the data:

1. The median of the data
2. The lower quartile
3. The upper quartile
4. The smallest observation
5. The largest observation

These statements require two qualifications. First, we will assume that the *hinges* of the box plot are essentially the quartiles of the data set. (We will define hinges shortly.) The median is a line inside the box.

FIGURE 1–18 **Further Refined Stem-and-Leaf Display of Data of Example 1–8**

	1*	1
	t	223
	f	555
	s	67
	.	
	2*	0111
	t	2223
	f	4
(Median in this class)	(s)	6777
	.	899
	3*	01
	t	2
	f	45
	s	7
	.	
	4*	11
	t	2
	f	5
	s	7
	.	
	5*	0
	t	23
	f	
	s	6
	.	
	6*	0
	t	2

Second, the **whiskers** of the box plot are made by extending a line from the upper quartile to the largest observation and from the lower quartile to the smallest observation, only if the largest and smallest observations are within a distance of 1.5 times the interquartile range from the appropriate hinge (quartile). If one or more observations are farther away than that distance, they are marked as suspected outliers. If these observations are at a distance of over 3 times the interquartile range from the appropriate hinge, they are marked as outliers. The whisker then extends to the largest or smallest observation that is at a distance less than or equal to 1.5 times the interquartile range from the hinge.

Let us make these definitions clearer by using a picture. Figure 1–19 shows the parts of a box plot and how they are defined. The median is marked as a vertical line across the box. The **hinges** of the box are the upper and lower quartiles (the rightmost and leftmost sides of the box). The interquartile range (IQR) is the distance from the upper quartile to the lower quartile (the length of the box from hinge to hinge): IQR = $Q_U - Q_L$. We define the **inner fence** as a point at a distance of 1.5(IQR) above the upper quartile; similarly, the lower inner fence is $Q_L - 1.5(\text{IQR})$. The **outer fences** are defined similarly but are at a distance of 3(IQR) above or below the appropriate hinge. Figure 1–20 shows the fences (these are not shown on the actual box plot; they are only guidelines for defining the whiskers, suspected outliers, and outliers) and demonstrates how we mark outliers.

FIGURE 1–19 The Box Plot

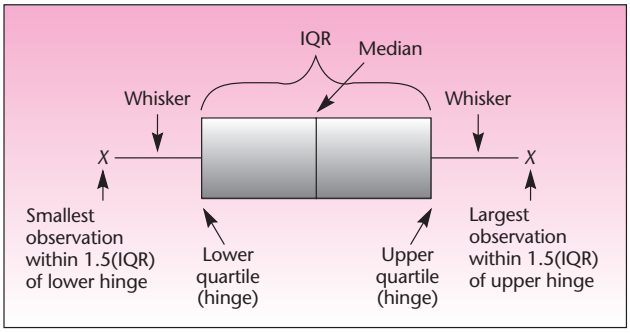
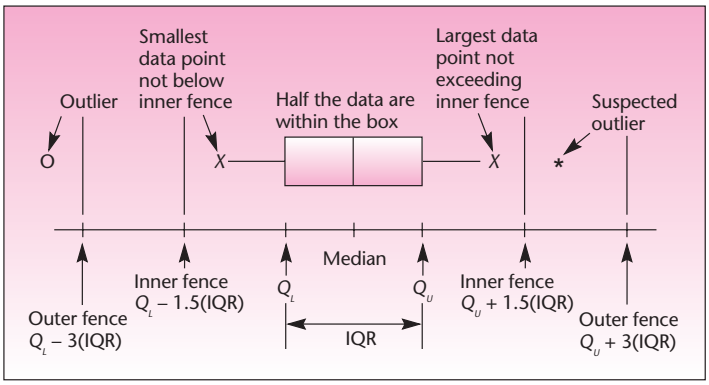


FIGURE 1–20 The Elements of a Box Plot



Box plots are very useful for the following purposes.

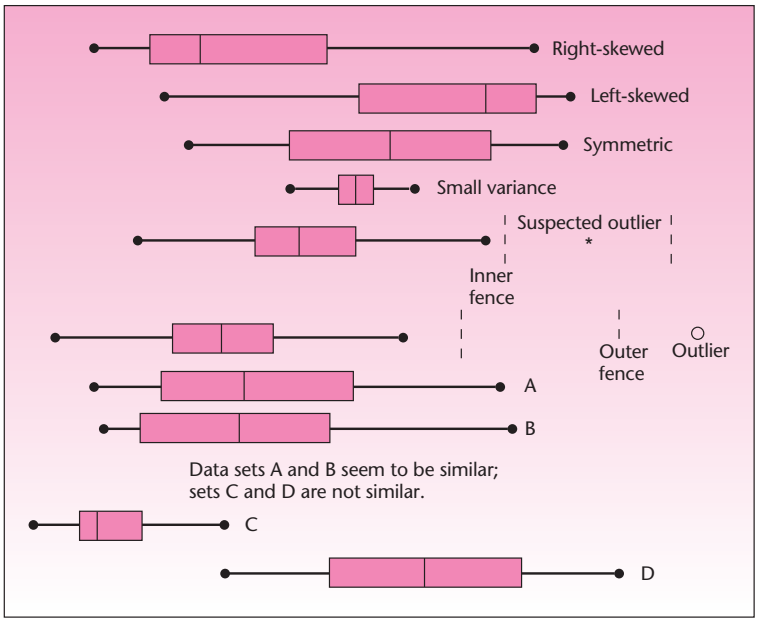
1. To identify the location of a data set based on the median.
2. To identify the spread of the data based on the length of the box, hinge to hinge (the interquartile range), and the length of the whiskers (the range of the data without extreme observations: outliers or suspected outliers).
3. To identify possible skewness of the distribution of the data set. If the portion of the box to the right of the median is longer than the portion to the left of the median, and/or the right whisker is longer than the left whisker, the data are right-skewed. Similarly, a longer left side of the box and/or left whisker implies a left-skewed data set. If the box and whiskers are symmetric, the data are symmetrically distributed with no skewness.
4. To identify suspected outliers (observations beyond the inner fences but within the outer fences) and outliers (points beyond the outer fences).
5. To compare two or more data sets. By drawing a box plot for each data set and displaying the box plots on the same scale, we can compare several data sets.

A special form of a box plot may even be used for conducting a test of the equality of two population medians. The various uses of a box plot are demonstrated in Figure 1–21.

Let us now construct a box plot for the data of Example 1–8. For this data set, the median is 27, and we find that the lower quartile is 20.75 and the upper quartile is 41. The interquartile range is IQR = 41 − 20.75 = 20.25. One and one-half times this distance is 30.38; hence, the inner fences are −9.63 and 71.38. Since no observation lies beyond either point, there are no suspected outliers and no outliers, so the whiskers extend to the extreme values in the data: 11 on the left side and 62 on the right side.

As you can see from the figure, there are no outliers or suspected outliers in this data set. The data set is skewed to the right. This confirms our observation of the skewness from consideration of the stem-and-leaf diagrams of the same data set, in Figures 1–16 to 1–18.

FIGURE 1–21 **Box Plots and Their Uses**



PROBLEMS

1–47. The following data are monthly steel production figures, in millions of tons.

7.0, 6.9, 8.2, 7.8, 7.7, 7.3, 6.8, 6.7, 8.2, 8.4, 7.0, 6.7, 7.5, 7.2, 7.9, 7.6, 6.7, 6.6, 6.3, 5.6, 7.8, 5.5, 6.2, 5.8, 5.8, 6.1, 6.0, 7.3, 7.3, 7.5, 7.2, 7.2, 7.4, 7.6

Draw a stem-and-leaf display of these data.

1–48. Draw a box plot for the data in problem 1–47. Are there any outliers? Is the distribution of the data symmetric or skewed? If it is skewed, to what side?

1–49. What are the uses of a stem-and-leaf display? What are the uses of a box plot?

1–50. Worker participation in management is a new concept that involves employees in corporate decision making. The following data are the percentages of employees involved in worker participation programs in a sample of firms. Draw a stem-and-leaf display of the data.

5, 32, 33, 35, 42, 43, 42, 45, 46, 44, 47, 48, 48, 48, 49, 49, 50, 37, 38, 34, 51, 52, 52, 47, 53, 55, 56, 57, 58, 63, 78

1–51. Draw a box plot of the data in problem 1–50, and draw conclusions about the data set based on the box plot.

1–52. Consider the two box plots in Figure 1–24 (on page 38), and draw conclusions about the data sets.

1–53. Refer to the following data on distances between seats in business class for various airlines. Find μ, σ, σ^2, draw a box plot, and find the mode and any outliers.

Characteristics of Business-Class Carriers

	Distance between Rows (in cm)
Europe	
Air France	122
Alitalia	140
British Airways	127
Iberia	107
KLM/Northwest	120
Lufthansa	101
Sabena	122
SAS	132
SwissAir	120
Asia	
All Nippon Airw	127
Cathay Pacific	127
JAL	127
Korean Air	127
Malaysia Air	116
Singapore Airl	120
Thai Airways	128
Vietnam Airl	140
North America	
Air Canada	140
American Airl	127
Continental	140
Delta Airlines	130
TWA	157
United	124

1–54. The following data are the daily price quotations for a certain stock over a period of 45 days. Construct a stem-and-leaf display for these data. What can you conclude about the distribution of daily stock prices over the period under study?

10, 11, 10, 11, 11, 12, 12, 13, 14, 16, 15, 11, 18, 19, 20, 15, 14, 14, 22, 25, 27, 23, 22, 26, 27, 29, 28, 31, 32, 30, 32, 34, 33, 38, 41, 40, 42, 53, 52, 47, 37, 23, 11, 32, 23

1–55. Discuss ways of dealing with outliers—their detection and what to do about them once they are detected. Can you always discard an outlier? Why or why not?

1–56. Define the inner fences and the outer fences of a box plot; also define the whiskers and the hinges. What portion of the data is represented by the box? By the whiskers?

1–57. The following data are the number of ounces of silver per ton of ore for two mines.

Mine A: 34, 32, 35, 37, 41, 42, 43, 45, 46, 45, 48, 49, 51, 52, 53, 60, 73, 76, 85

Mine B: 23, 24, 28, 29, 32, 34, 35, 37, 38, 40, 43, 44, 47, 48, 49, 50, 51, 52, 59

Construct a stem-and-leaf display for each data set and a box plot for each data set. Compare the two displays and the two box plots. Draw conclusions about the data.

1–58. Can you compare two *populations* by looking at box plots or stem-and-leaf displays of random samples from the two populations? Explain.

1–59. The following data are daily percentage changes in stock prices for 20 stocks called "The Favorites."[15]

−0.1, 0.5, 0.6, 0.7, 1.4, 0.7, 1.3, 0.3, 1.6, 0.6, −3.5, 0.6, 1.1, 1.3, −0.1, 2.5, −0.3, 0.3, 0.2, 0.4

Draw a box plot of these data.

1–60. Consult the following data on a sports car 0 to 60 times, in seconds.[16]

4.9, 4.6, 4.2, 5.1, 5.2, 5.1, 4.8, 4.7, 4.9, 5.3

Find the mean and the median. Compare the two. Also construct a box plot. Interpret your findings.

1–10 Using the Computer

Using Excel for Descriptive Statistics and Plots

If you need to develop any statistical or engineering analyses, you can use the **Excel Analysis Toolpack.** One of the applicable features available in the Analysis Toolpack is Descriptive Statistics. To access this tool, click Data Analysis in the Analysis Group on the Data tab. Then choose Descriptive Statistics. You can define the range of input and output in this window. Don't forget to select the Summary Statistics check box. Then press OK. A table containing the descriptive statistics of your data set will be created in the place that you have specified for output range.

If the Data Analysis command is not available in the Data tab, you need to load the Analysis Toolpack add-in program. For this purpose follow the next steps:

- Click the Microsoft Office button, and then click Excel Options.
- Click Add-ins, and then in the Manage box, select Excel Add-ins.
- Click Go.
- In the Add-ins Available box, select the Analysis Toolpack check box, and then click OK.

[15]Data reported in "Business Day," *The New York Times,* Thursday, March 15, 2007, p. C11.

[16]"Sports Stars," *BusinessWeek,* March 5, 2007, p. 140.

FIGURE 1–22 Template for Calculating Basic Statistics [Basic Statistics.xls]

Basic Statistics from Raw Data — Sales Data

Measures of Central tendency					
Mean	26.9	Median	22	Mode	18

Measures of Dispersion	If the data is of a Sample	If the data is of a Population		
Variance	139.884211	132.89	Range	38
St. Dev.	11.8272656	11.5277925	IQR	8.5

Skewness and Kurtosis	If the data is of a Sample	If the data is of a Population
Skewness	1.65371559	1.52700876
(Relative) Kurtosis	1.60368514	0.94417958

Percentile and Percentile Rank Calculations				
x	x-th Percentile	y	Percentile rank of y
50	22	22.0	47
80	32.2	32.2	80
90	49.3	49.3	90

Quartiles			
1st Quartile	19.75		
Median	22	IQR	8.5
3rd Quartile	28.25		

Other Statistics	
Sum	538
Size	22
Maximum	56
Minimum	18

	Data Entry
1	33
2	26
3	24
4	21
5	19
6	20
7	18
8	18
9	52
10	56
11	27
12	22
13	18
14	49
15	22
16	20
17	23
18	32
19	20
20	18

In addition to the useful features of the Excel Analysis Toolpak and the direct use of Excel commands as shown in Figure 1–4, we also will discuss the use of Excel templates that we have developed for computations and charts covered in the chapter. General instructions about using templates appear on the Student CD.

Figure 1–22 shows the template that can be used for calculating basic statistics of a data set. As soon as the data are entered in the shaded area in column K, all the statistics are automatically calculated and displayed. All the statistics have been explained in this chapter, but some aspects of this template will be discussed next.

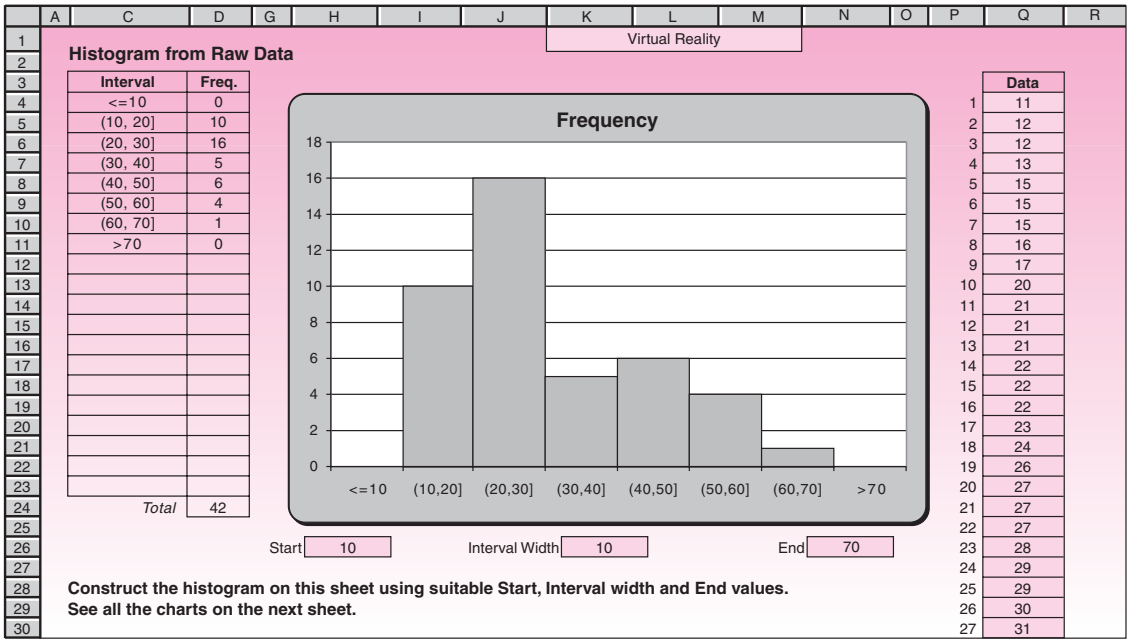
PERCENTILE AND PERCENTILE RANK COMPUTATION

The percentile and percentile rank computations are done slightly differently in Excel. Do not be alarmed if your manual calculation differs (slightly) from the result you see in the template. These discrepancies in percentile and percentile rank computations occur because of approximation and rounding off. In Figure 1–22, notice that the 50th percentile is 22, but the percentile rank of 22 is 47. Such discrepancies will get smaller as the size of the data set increases. For large data sets, the discrepancy will be negligible or absent.

HISTOGRAMS

A histogram can be drawn either from raw data or from grouped data, so the workbook contains one sheet for each case. Figure 1–23 shows the template that used raw data. After entering the data in the shaded area in column Q, select appropriate values for the start, interval width, and end values for the histogram in

FIGURE 1–23 Template for Histograms and Related Charts [Histogram.xls; Sheet: from Raw Data]



cells H26, K26, and N26 respectively. When selecting the start and end values, make sure that the first bar and the last bar of the chart have zero frequencies. This will ensure that no value in the data has been omitted. The interval width should be selected to make the histogram a good representation of the distribution of the data.

After constructing the histogram on this sheet, go to the next sheet, named "Charts," to see all the related charts: Relative Frequency, Frequency Polygon, Relative Frequency Polygon, and Ogive.

At times, you may have grouped data rather than raw data to start with. In this case, go to the grouped data sheet and enter the data in the shaded area on the right. This sheet contains a total of five charts. If any of these is not needed, unprotect the sheet and delete it before printing. Another useful template provided in the CD is Frequency Polygon.xls, which is used to compare two distributions.

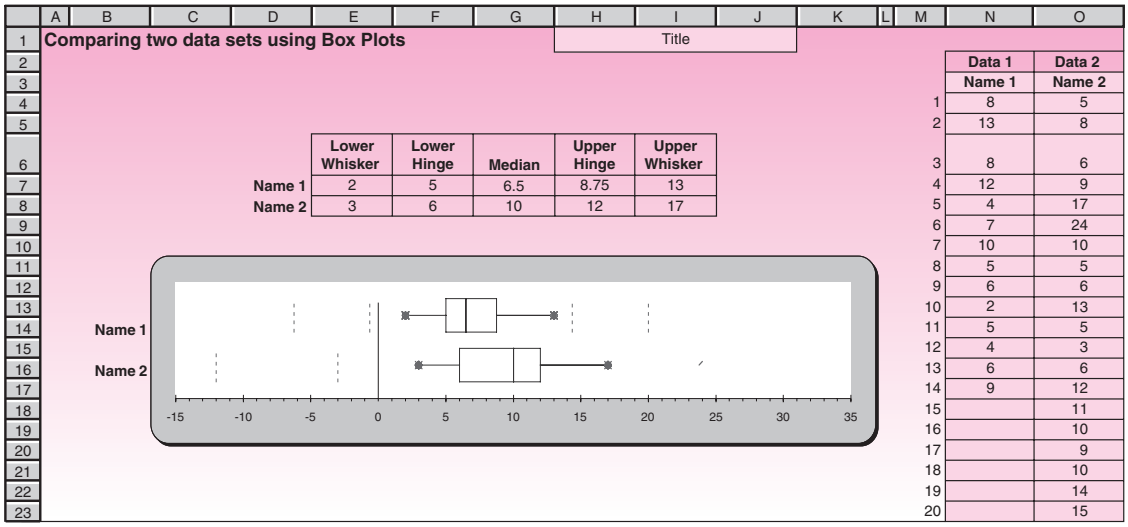
An advantage of frequency polygons is that unlike histograms, we can superpose two or more polygons to compare the distributions.

PIE CHARTS

Pie chart.xls is one of the templates in the CD for creating pie charts. Note that the data entered in this template for creating a pie chart need not be percentages, and even if they are percentages, they need not add up to 100%, since the spreadsheet recalculates the proportions.

If you wish to modify the format of the chart, for example, by changing the colors of the slices or the location of legends, unprotect the sheet and use the Chart Wizard. To use the Chart Wizard, click on the icon that looks like this:  . Protect the sheet after you are done.

FIGURE 1–24 Box Plot Template to Compare Two Data Sets [Box Plot 2.xls]



BAR CHARTS

Bar chart.xls is the template that can be used to draw bar charts. Many refinements are possible on the bar charts, such as making it a 3-D chart. You can unprotect the sheet and use the Chart Wizard to make the refinements.

BOX PLOTS

Box plot.xls is the template that can be used to create box plots. Box plot2.xls is the template that draws two box plots of two different data sets. Thus it can be used to compare two data sets. Figure 1–24 shows the comparison between two data sets using this template. Cells N3 and O3 are used to enter the name for each data set. The comparison shows that the second data set is more varied and contains relatively larger numbers than the first set.

TIME PLOTS

Time plot.xls is the template that can be used to create time plots.

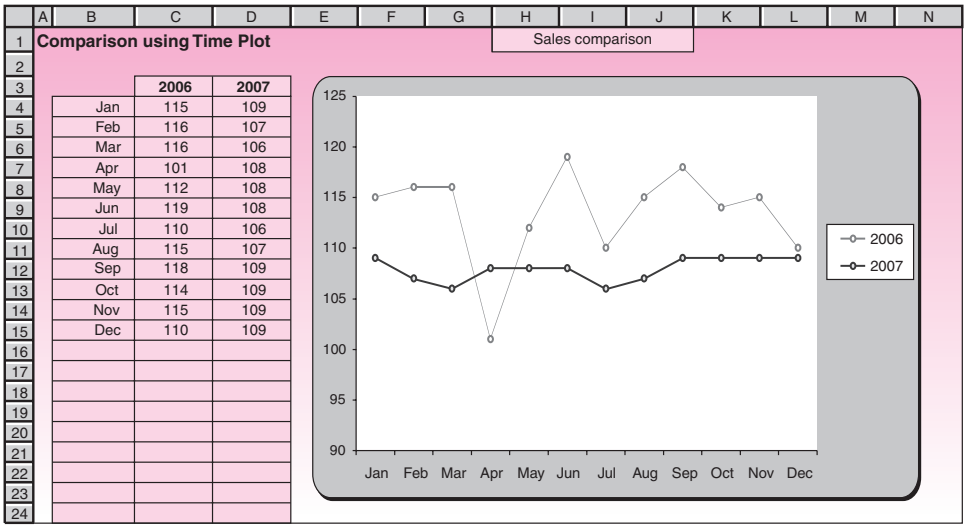
To compare two data sets, use the template timeplot2.xls. Comparing sales in years 2006 and 2007, Figure 1–25 shows that Year 2007 sales were consistently below those of Year 2006, except in April. Moreover, the Year 2007 sales show less variance than those of Year 2006. Reasons for both facts may be worth investigating.

SCATTER PLOTS

Scatter plots are used to identify and report any underlying relationships among pairs of data sets. For example, if we have the data on annual sales of a product and on the annual advertising budgets for that product during the same period, then we can plot them on the same graph to see if a pattern emerges that brings out a relationship between the data sets. We might expect that whenever the advertising budget was high, the sales would also be high. This can be verified on a scatter plot.

The plot consists of a scatter of points, each point representing an observation. For instance, if the advertising budget in one year was x and the sales in the same year was y, then a point is marked on the plot at coordinates (x, y). Scatter plot.xls is the template that can be used to create a scatter plot.

FIGURE 1–25 Time Plot Comparison [Time Plot 2.xls]



Sometimes we have several data sets, and we may want to know if a relation exists between any two of them. Plotting every pair of them can be tedious, so it would be faster and easier if a bunch of scatter plots are produced together. The template Scatter plot.xls has another sheet named "5 Variables" which accommodates data on five variables and produces a scatter plot for every pair of variables. A glance at the scatter plots can quickly reveal an apparent correlation between any pair.

Using MINITAB for Descriptive Statistics and Plots

MINITAB can use data from different sources: previously saved MINITAB worksheet files, text files, and Microsoft Excel files. To place data in MINITAB, we can:

- Type directly into MINITAB.
- Copy and paste from other applications.
- Open from a variety of file types, including Excel or text files.

In this section we demonstrate the use of MINITAB in producing descriptive statistics and corresponding plots with the data of Example 1–2. If you are using a keyboard to type the data into the worksheet, begin in the row above the horizontal line containing the numbered row. This row is used to provide a label for each variable. In the first column (labeled C1) enter the label of your variable (wealth) and press Enter. By moving the cursor to the cell in the next row, you can start entering data in the first column.

To open data from a file, choose File ▶ Open Worksheet. This will provide you with the open worksheet dialog box. Many different files, including Minitab worksheet files (.MTW), Microsoft Excel (.XLS), data (.DAT), and text (.TXT), can be opened from this dialog box. Make sure that the proper file type appears in the List of Files of Type Box. You can also use the Session window and type the command to set the data into the columns.

For obtaining descriptive statistics, you can type the appropriate command in the Session window or use the menu. Figure 1–26 shows the command, data, and output for Example 1–2.

FIGURE 1–26 Using MINITAB to Describe Data

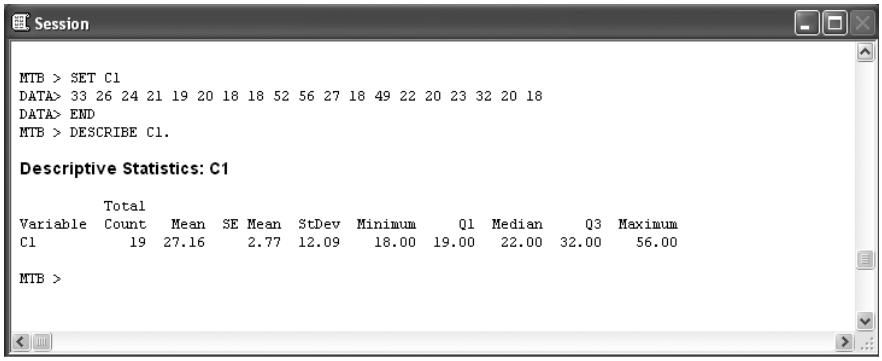
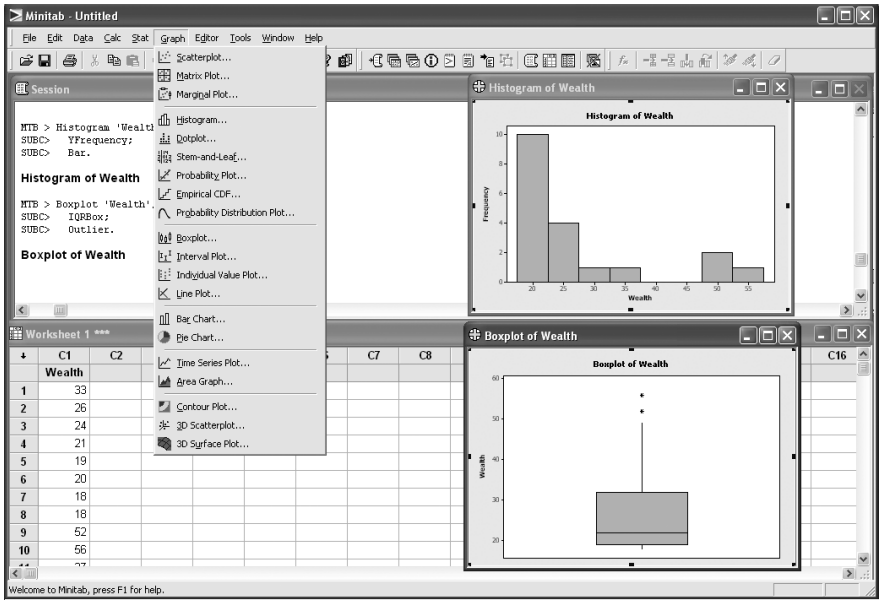


FIGURE 1–27 MINITAB Output



To obtain descriptive statistics using the menu, choose Stat ▶ Basic Statistics ▶ Display Descriptive Statistics. In the Descriptive Statistics dialog box choose C1 in the Variable List box and then press zero. The result will be shown in the Session window. Some users find menu commands quicker to use than session commands.

As was mentioned earlier in the chapter, we can use graphs to explore data and assess relationships among the variables. You can access MINITAB's graph from the Graph and Stat menus. Using the Graph menu enables you to obtain a large variety of graphs. Figure 1–27 shows the histogram and box plot obtained using the Graph menu.

Finally, note that MINITAB does not display the command prompt by default. To enter commands directly into the Session window, you must enable this prompt by choosing Editor ▶ Enable Commands. A check appears next to the menu item.

When you execute a command from a menu and session commands are enabled, the corresponding session command appears in the Session window along with the text output. This technique provides a convenient way to learn session commands.

1–11 Summary and Review of Terms

In this chapter we introduced many terms and concepts. We defined a **population** as the set of all measurements in which we are interested. We defined a **sample** as a smaller group of measurements chosen from the larger population (the concept of random sampling will be discussed in detail in Chapter 4). We defined the process of using the sample for drawing conclusions about the population as **statistical inference.**

We discussed **descriptive statistics** as quantities computed from our data. We also defined the following statistics: **percentile,** a point below which lie a specified percentage of the data, and **quartile,** a percentile point in multiples of 25. The first quartile, the 25th percentile point, is also called the **lower quartile.** The 50th percentile point is the second quartile, also called the middle quartile, or the **median.** The 75th percentile is the **third quartile,** or the upper quartile. We defined the **interquartile range** as the difference between the upper and lower quartiles. We said that the median is a measure of central tendency, and we defined two other measures of central tendency: the **mode,** which is a *most frequent* value, and the **mean.** We called the mean the most important measure of central tendency, or location, of the data set. We said that the mean is the average of all the data points and is the point where the entire distribution of data points balances.

We defined measures of variability: the **range,** the **variance,** and the **standard deviation.** We defined the range as the difference between the largest and smallest data points. The variance was defined as the average squared deviation of the data points from their mean. For a sample (rather than a population), we saw that this averaging is done by dividing the sum of the squared deviations from the mean by $n - 1$ instead of by n. We defined the standard deviation as the square root of the variance.

We discussed grouped data and **frequencies** of occurrence of data points in **classes** defined by intervals of numbers. We defined **relative frequencies** as the absolute frequencies, or counts, divided by the total number of data points. We saw how to construct a **histogram** of a data set: a graph of the frequencies of the data. We mentioned **skewness,** a measure of the asymmetry of the histogram of the data set. We also mentioned **kurtosis,** a measure of the flatness of the distribution. We introduced **Chebyshev's theorem** and the **empirical rule** as ways of determining the proportions of data lying within several standard deviations of the mean.

We defined four scales of measurement of data: **nominal**—name only; **ordinal**—data that can be ordered as greater than or less than; **interval**—with meaningful distances as intervals of numbers; and **ratio**—a scale where ratios of distances are also meaningful.

The next topic we discussed was graphical techniques. These extended the idea of a histogram. We saw how a **frequency polygon** may be used instead of a histogram. We also saw how to construct an **ogive:** a cumulative frequency graph of a data set. We also talked about **bar charts** and **pie charts,** which are types of charts for displaying data, both categorical and numerical.

Then we discussed **exploratory data analysis,** a statistical area devoted to analyzing data using graphical techniques and other techniques that do not make restrictive assumptions about the structure of the data. Here we encountered two useful techniques for plotting data in a way that sheds light on their structure: **stem-and-leaf displays** and **box plots.** We saw that a stem-and-leaf display, which can be drawn quickly, is a type of histogram that makes use of the decimal structure of our number system. We saw how a box plot is made out of five quantities: the median, the two **hinges,** and the two **whiskers.** And we saw how the whiskers, as well as outliers and suspected outliers, are determined by the **inner fences** and **outer fences;** the first lies at a distance of 1.5 times the interquartile range from the hinges, and the second is found at 3 times the interquartile range from the hinges.

Finally, was saw the use of **templates** to compute population parameters and sample statistics, create histograms and frequency polygons, create bar charts and pie charts, draw box plots, and produce scatter plots.

ADDITIONAL PROBLEMS

1–61. Open the workbook named Problem 1–61.xls. Study the statistics that have been calculated in the worksheet. Of special interest to this exercise are the two cells marked Mult and Add. If you enter 2 under Mult, all the data points will be multiplied by 2, as seen in the modified data column. Entering 1 under Mult leaves the data unchanged, since multiplying a number by 1 does not affect it. Similarly, entering 5 under Add will add 5 to all the data points. Entering 0 under Add will leave the data unchanged.

1. Set Mult = 1 and Add = 5, which corresponds to adding 5 to all data points. Observe how the statistics have changed in the modified statistics column. Keeping Mult = 1 and changing Add to different values, observe how the statistics change. Then make a formal statement such as "If we add x to all the data points, then the average would increase by x," for each of the statistics, starting with average.
2. Add an explanation for each statement made in part 1 above. For the average, this will be "If we add x to all the data points, then the sum of all the numbers will increase by $x*n$ where n is the number of data points. The sum is divided by n to get the average. So the average will increase by x."
3. Repeat part 1 for multiplying all the data points by some number. This would require setting Mult equal to desired values and Add = 0.
4. Repeat part 1 for multiplying and adding at once. This would require setting both Mult and Add to desired values.

1–62. *Fortune* published a list of the 10 largest "green companies"–those that follow environmental policies. Their annual revenues, in $ billions, are given below.[17]

Company	Revenue $ Billion
Honda	$84.2
Continental Airlines	13.1
Suncor	13.6
Tesco	71.0
Alcan	23.6
PG&E	12.5
S.C. Johnson	7.0
Goldman Sachs	69.4
Swiss RE	24.0
Hewlett-Packard	91.7

Find the mean, variance, and standard deviation of the annual revenues.

1–63. The following data are the number of tons shipped weekly across the Pacific by a shipping company.

398, 412, 560, 476, 544, 690, 587, 600, 613, 457, 504, 477, 530, 641, 359, 566, 452, 633, 474, 499, 580, 606, 344, 455, 505, 396, 347, 441, 390, 632, 400, 582

Assume these data represent an entire population. Find the population mean and the population standard deviation.

1–64. Group the data in problem 1–63 into classes, and draw a histogram of the frequency distribution.

[17] "Green Is Good: Ten Green Giants," *Fortune,* April 2, 2007, pp. 44–50.

1–65. Find the 90th percentile, the quartiles, and the range of the data in problem 1–63.

1–66. The following data are numbers of color television sets manufactured per day at a given plant: 15, 16, 18, 19, 14, 12, 22, 23, 25, 20, 32, 17, 34, 25, 40, 41. Draw a frequency polygon and an ogive for these data.

1–67. Construct a stem-and-leaf display for the data in problem 1–66.

1–68. Construct a box plot for the data in problem 1–66. What can you say about the data?

1–69. The following data are the number of cars passing a point on a highway per minute: 10, 12, 11, 19, 22, 21, 23, 22, 24, 25, 23, 21, 28, 26, 27, 27, 29, 26, 22, 28, 30, 32, 25, 37, 34, 35, 62. Construct a stem-and-leaf display of these data. What does the display tell you about the data?

1–70. For the data problem 1–69, construct a box plot. What does the box plot tell you about these data?

1–71. An article by Julia Moskin in the *New York Times* reports on the use of cheap wine in cooking.[18] Assume that the following results are taste-test ratings, from 1 to 10, for food cooked in cheap wine.

7, 7, 5, 6, 9, 10, 10, 10, 10, 7, 3, 8, 10, 10, 9

Find the mean, median, and modes of these data. Based on these data alone, do you think cheap wine works?

1–72. The following are a sample of Motorola's stock prices in March 2007.[19]

20, 20.5, 19.8, 19.9, 20.1, 20.2, 20.7, 20.6, 20.8, 20.2, 20.6, 20.2

Find the mean and the variance, plot the data, determine outliers, and construct a box plot.

1–73. Consult the corporate data shown below. Plot data; find μ, σ, σ^2; and identify outliers.

Morgan Stanley	91.36%
Merrill Lynch	40.26
Travelers	39.42
Warner-Lambert	35.00
Microsoft	32.95
J.P. Morgan & Co.	29.62
Lehman Brothers	28.25
US Airways	26.71
Sun Microsystems	25.99
Marriott	25.81
Bankers Trust	25.53
General Mills	25.41
MCI	24.39
AlliedSignal	24.23
ITT Industries	24.14

1–74. The following are quoted interest rates (%) on Italian bonds.

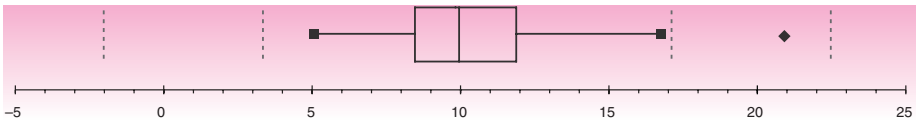
2.95, 4.25, 3.55, 1.90, 2.05, 1.78, 2.90, 1.85, 3.45, 1.75, 3.50, 1.69, 2.85, 4.10, 3.80, 3.85, 2.85, 8.70, 1.80, 2.87, 3.95, 3.50, 2.90, 3.45, 3.40, 3.55, 4.25, 1.85, 2.95

Plot the data; find μ, σ, and σ^2; and identify outliers (one is private, the rest are banks and government).

[18]Julia Moskin, "It Boils Down to This: Cheap Wine Works Fine," *The New York Times,* March 21, 2007, p. D1.

[19]Adapted from a chart in R. Farzad, "Activist Investors Not Welcome," *BusinessWeek,* April 9, 2007, p. 36.

1–75. Refer to the box plot below to answer the questions.



1. What is the interquartile range for this data set?
2. What can you say about the skewness of this data set?
3. For this data set, the value of 9.5 is more likely to be (choose one)
   a. The first quartile rather than the median.
   b. The median rather than the first quartile.
   c. The mean rather than the mode.
   d. The mode rather than the mean.
4. If a data point that was originally 13 is changed to 14, how would the box plot be affected?

1–76. The following table shows changes in bad loans and in provisions for bad loans, from 2005 to 2006, for 19 lending institutions.[20] Verify the reported averages, and find the medians. Which measure is more meaningful, in your opinion? Also find the standard deviation and identify outliers for change in bad loans and change in provision for bad loans.

Menacing Loans

Bank/Assets $ Billions	Change in Bad Loans* 12/06 vs. 12/05	Change in Provisions for Bad Loans
Bank of America ($1,459.0)	16.8%	12.1%
Wachovia (707.1)	91.7	23.3
Wells Fargo (481.9)	24.5	−2.8
Suntrust Banks (182.2)	123.5	4.4
Bank of New York (103.4)	42.3	−12.0
Fifth Third Bancorp (100.7)	19.7	3.6
Northern Trust (60.7)	15.2	12.0
Comerica (58.0)	55.1	−4.5
M&T Bank (57.0)	44.9	1.9
Marshall & Isley (56.2)	96.5	15.6
Commerce Bancorp ($45.3)	45.5	13.8
TD Banknorth (40.2)	116.9	25.4
First Horizon National (37.9)	79.2	14.0
Huntington Bancshares (35.3)	22.9	1.4
Compass Bancshares (34.2)	17.3	8.9
Synovus Financial (31.9)	17.6	8.6
Associated Banc-Corp (21.0)	43.4	0.0
Mercantile Bankshares (17.72)	37.2	−8.7
W Holding (17.2)	159.1	37.3
Average** (149.30)	11.00	4.1

*Nonperforming loans.
**At 56 banks with more than $10 billion in assets.
Data: SNL financial.

[20]Mara der Hovanesian, "Lender Woes Go beyond Subprime," *BusinessWeek,* March 12, 2007, p. 38. Reprinted by permission.

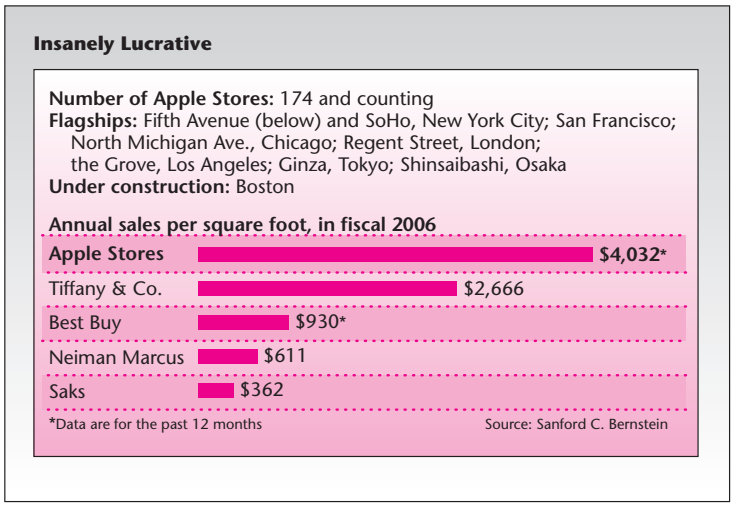
1–77. Repeat problem 1–76 for the bank assets data, shown in parentheses in the table at the bottom of the previous page.

1–78. A country's percentage approval of its citizens in European Union membership is given below.[21]

Ireland	78%	Luxembourg	75%	Netherlands	70%
Belgium	68	Spain	62	Denmark	60
Germany	58	Greece	57	Italy	52
France	52	Portugal	48	Sweden	48
Finland	40	Austria	39	Britain	37

Find the mean, median, and standard deviation for the percentage approval. Compare the mean and median to the entire EU approval percentage, 53%.

1–79. The following display is adapted from an article in *Fortune*.[22]



Interpret the chart, and find the mean and standard deviation of the data, viewed as a population.

1–80. The future Euroyen is the price of the Japanese yen as traded in the European futures market. The following are 30-day Euroyen prices on an index from 0 to 100%: 99.24, 99.37, 98.33, 98.91, 98.51, 99.38, 99.71, 99.21, 98.63, 99.10. Find μ, σ, σ^2, and the median.

1–81. The daily expenditure on food by a traveler, in dollars in summer 2006, was as follows: 17.5, 17.6, 18.3, 17.9, 17.4, 16.9, 17.1, 17.1, 18.0, 17.2, 18.3, 17.8, 17.1, 18.3, 17.5, 17.4. Find the mean, standard deviation, and variance.

1–82. For the following data on financial institutions' net income, find the mean and the standard deviation.[23]

Goldman Sachs	$ 9.5 billion
Lehman Brothers	4.0 billion
Moody's	$753 million
T. Rowe Price	$530 million
PNC Financial	$ 2.6 billion

[21]"Four D's for Europe: Dealing with the Dreaded Democratic Deficit," *The Economist,* March 17, 2007, p. 16.

[22]Jerry Useem, "Simply Irresistible: Why Apple Is the Best Retailer in America," *Fortune,* March 19, 2007, p. 108.

[23]"The Rankings," *BusinessWeek,* March 26, 2007, pp. 74–90.

1–83. The following are the percentage profitability data (%) for the top 12 American corporations.[24]

39, 33, 63, 41, 46, 32, 27, 13, 55, 35, 32, 30

Find the mean, median, and standard deviation of the percentages.

1–84. Find the daily stock price of Wal-Mart for the last three months. (A good source for the data is http://moneycentral.msn.com. You can ask for the three-month chart and export the data to a spreadsheet.)

1. Calculate the mean and the standard deviation of the stock prices.
2. Get the corresponding data for Kmart and calculate the mean and the standard deviation.
3. The coefficient of variation (CV) is defined as the ratio of the standard deviation over the mean. Calculate the CV of Wal-Mart and Kmart stock prices.
4. If the CV of the daily stock prices is taken as an indicator of risk of the stock, how do Wal-Mart and Kmart stocks compare in terms of risk? (There are better measures of risk, but we will use CV in this exercise.)
5. Get the corresponding data of the Dow Jones Industrial Average (DJIA) and compute its CV. How do Wal-Mart and Kmart stocks compare with the DJIA in terms of risk?
6. Suppose you bought 100 shares of Wal-Mart stock three months ago and held it. What are the mean and the standard deviation of the daily market price of your holding for the three months?

1–85. To calculate variance and standard deviation, we take the deviations from the mean. At times, we need to consider the deviations from a target value rather than the mean. Consider the case of a machine that bottles cola into 2-liter (2,000-cm^3) bottles. The target is thus 2,000 cm^3. The machine, however, may be bottling 2,004 cm^3 on average into every bottle. Call this 2,004 cm^3 the *process mean*. The damage from process errors is determined by the deviations from the target rather than from the process mean. The variance, though, is calculated with deviations from the process mean, and therefore is not a measure of the damage. Suppose we want to calculate a new variance using deviations from the target value. Let "SSD(Target)" denote the sum of the squared deviations from the target. [For example, SSD(2,000) denotes the sum of squared deviations when the deviations are taken from 2,000.] Dividing the SSD by the number of data points gives the Average SSD(Target).

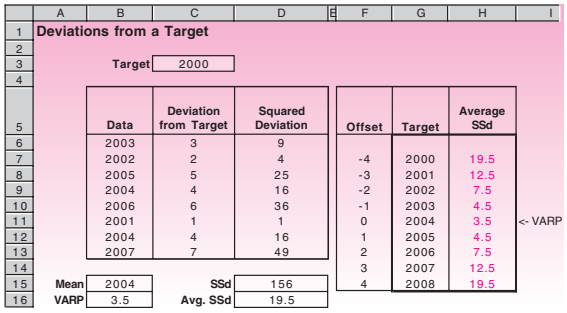
The following spreadsheet is set up to calculate the deviations from the target, SSD(Target), and the Average SSD(Target). Column B contains the data, showing a process mean of 2,004. (Strictly speaking, this would be sample data. But to simplify matters, let us assume that this is population data.) Note that the population variance (VARP) is 3.5 and the Average SSD(2,000) is 19.5.

In the range G5:H13, a table has been created to see the effect of changing the target on Average SSD(Target). The offset refers to the difference between the target and the process mean.

1. Study the table and find an equation that relates the Average SSD to VARP and the Offset. [Hint: Note that while calculating SSD, the deviations are squared, so think in squares.]
2. Using the equation you found in part 1, prove that the Average SSD(Target) is minimized when the target equals the process mean.

[24]From "Inside the Rankings," *BusinessWeek,* March 26, 2007, p. 92.

Working with Deviations from a Target [Problem 1–85.xls]



	A	B	C	D	E	F	G	H	I
1	Deviations from a Target								
2									
3		Target	2000						
4									
5		Data	Deviation from Target	Squared Deviation		Offset	Target	Average SSd	
6		2003	3	9					
7		2002	2	4		-4	2000	19.5	
8		2005	5	25		-3	2001	12.5	
9		2004	4	16		-2	2002	7.5	
10		2006	6	36		-1	2003	4.5	
11		2001	1	1		0	2004	3.5	<- VARP
12		2004	4	16		1	2005	4.5	
13		2007	7	49		2	2006	7.5	
14						3	2007	12.5	
15		Mean	2004	SSd	156	4	2008	19.5	
16		VARP	3.5	Avg. SSd	19.5				

1–86. The Consumer Price Index (CPI) is an important indicator of the general level of prices of essential commodities. It is widely used in making cost of living adjustments to salaries, for example.

1. Log on to the Consumer Price Index (CPI) home page of the Bureau of Labor Statistics Web site (stats.bls.gov/cpihome.htm). Get a table of the last 48 months' CPI for U.S. urban consumers with 1982–1984 as the base. Make a time plot of the data. Discuss any seasonal pattern you see in the data.
2. Go to the Average Price Data area and get a table of the last 48 months' average price of unleaded regular gasoline. Make a comparison time plot of the CPI data in part 1 and the gasoline price data. Comment on the gasoline prices.

1–87. Log on to the Center for Disease Control Web site and go to the HIV statistics page (www.cdc.gov/hiv/stats.htm).

1. Download the data on the cumulative number of AIDS cases reported in the United States and its age-range breakdown. Draw a pie chart of the data.
2. Download the race/ethnicity breakdown of the data. Draw a pie chart of the data.

1–88. Search the Web for major league baseball (MLB) players' salaries. ESPN and *USA Today* are good sources.

1. Get the Chicago Cubs players' salaries for the current year. Draw a box plot of the data. (Enter the data in thousands of dollars to make the numbers smaller.) Are there any outliers?
2. Get the Chicago White Sox players' salaries for the current year. Make a comparison box plot of the two data. Describe your comparison based on the plot.

1-89. The following data are bank yields (in percent) for 6-month CDs.[25]

3.56, 5.44, 5.37, 5.28, 5.19, 5.35, 5.48, 5.27, 5.39

Find the mean and standard deviation.

[25]"Wave and You've Paid," *Money*, March 2007, p. 40.

CASE 1 NASDAQ Volatility

The NASDAQ Combined Composite Index is a measure of the aggregate value of technological stocks. During the year 2007, the index moved up and down considerably, indicating the rapid changes in e-business that took place in that year and the high uncertainty in the profitability of technology-oriented companies. Historical data of the index are available at many Web sites, including **Finance.Yahoo.com.**

1. Download the monthly data of the index for the calendar year 2007 and make a time plot of the data. Comment on the volatility of the index, looking at the plot. Report the standard deviation of the data.
2. Download the monthly data of the index for the calendar year 2006 and compare the data for 2006 and 2007 on a single plot. Which year has been more volatile? Calculate the standard deviations of the two sets of data. Do they confirm your answer about the relative volatility of the two years?
3. Download the monthly data of the S&P 500 index for the year 2007. Compare this index with the NASDAQ index for the same year on a single plot. Which index has been more volatile? Calculate and report the standard deviations of the two sets of data.
4. Download the monthly data of the Dow Jones Industrial Average for the year 2007. Compare this index with the NASDAQ index for the same year on a single plot. Which index has been more volatile? Calculate and report the standard deviations of the two sets of data.
5. Repeat part 1 with the monthly data for the latest 12 full months.