Chapter

3

Numerical Descriptive Measures

3.1 Measures of Central Tendency for Ungrouped Data

Case Study 3–1 High-Priced Tickets in Big Markets

Case Study 3–2 Median Annual Starting Salary for MBAs

3.2 Measures of Dispersion for Ungrouped Data

3.3 Mean, Variance, and Standard Deviation for Grouped Data

3.4 Use of Standard Deviation

Case Study 3–3 Here Comes the SD

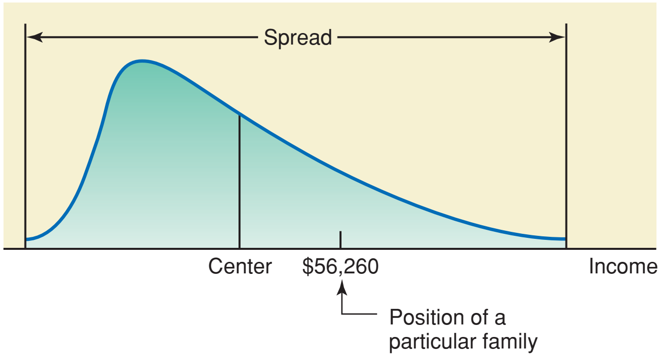
3.5 Measures of Position

3.6 Box-and-Whisker Plot

Which Major League Baseball team do you think has the highest average ticket price? Do you think it is one of the two New York teams, Mets or Yankees, assuming these teams play in one of the most expensive cities in the world? No, you are not even close. Then, is it one of the teams playing in Los Angeles—Dodgers or Angels? Still wrong. Actually it is the Boston Red Sox that had the highest average ticket price ($40.77) among all Major League Baseball teams in 2004. The next highest was $28.45 for the Chicago Cubs. See Case Study 3–1.

In Chapter 2 we discussed how to summarize data using different methods and to display data using graphs. Graphs are one important component of statistics; however it is also important to numerically describe the main characteristics of a data set. The numerical summary measures, such as the ones that identify the center and spread of a distribution, identify many important features of a distribution. For example, the techniques learned in Chapter 2 can help us graph data on family incomes. However, if we want to know the income of a "typical" family (given by the center of the distribution), the spread of the distribution of incomes, or the relative position of a family with a particular income, the numerical summary measures can provide more detailed information (see Figure 3.1). The measures that we discuss in this chapter include measures of (1) central tendency, (2) dispersion (or spread), and (3) position.

Figure 3.1



3.1 Measures of Central Tendency for Ungrouped Data

We often represent a data set by numerical summary measures, usually called the *typical values*. A **measure of central tendency** gives the center of a histogram or a frequency distribution curve. This section discusses three different measures of central tendency: the mean, the median, and the mode; however, a few other measures of central tendency, such as the trimmed mean, the weighted mean, and the geometric mean, are explained in exercises following this section. We will learn how to calculate each of these measures for ungrouped data. Recall from Chapter 2, the data that give information on each member of the population or sample individually are called *ungrouped data*, whereas *grouped data* are presented in the form of a frequency distribution table.

3.1.1 Mean

The **mean**, also called the *arithmetic mean*, is the most frequently used measure of central tendency. This book will use the words *mean* and *average* synonymously. For ungrouped data, the mean is obtained by dividing the sum of all values by the number of values in the data set.

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

The mean calculated for sample data is denoted by $\bar{x}$ (read as "x bar"), and the mean calculated for population data is denoted by μ (Greek letter *mu*). We know from the discussion in Chapter 2 that the number of values in a data set is denoted by n for a sample and by N for a population. In Chapter 1, we learned that a variable is denoted by x and the sum of all values of x is denoted by Σx. Using these notations, we can write the following formulas for the mean.

Calculating Mean for Ungrouped Data The *mean for ungrouped data* is obtained by dividing the sum of all values by the number of values in the data set. Thus,

$$\text{Mean for population data:} \quad \mu = \frac{\Sigma x}{N}$$

$$\text{Mean for sample data:} \quad \bar{x} = \frac{\Sigma x}{n}$$

where Σx is the sum of all values, N is the population size, n is the sample size, μ is the population mean, and $\bar{x}$ is the sample mean.

■ EXAMPLE 3–1

Calculating the sample mean for ungrouped data.

Table 3.1 lists the total number of identity fraud victims in 2004 for six states.

Table 3.1 Identity Fraud Victims in 2004 for Six States

State	Total Identity Fraud Victims in 2004
California	43,839
Florida	16,062
Illinois	11,138
New York	17,680
Ohio	6956
Texas	26,454

Source: Federal Trade Commission's Identity Theft Data Clearinghouse.

Find the mean number of identity fraud victims in 2004 for these six states.

Solution The variable in this example is the number of identity fraud victims in 2004 for six states. Let us denote it by x. Then, the six values of x are

$$x_1 = 43{,}839, \quad x_2 = 16{,}062, \quad x_3 = 11{,}138, \quad x_4 = 17{,}680, \quad x_5 = 6956, \quad \text{and} \quad x_6 = 26{,}454$$

where $x_1 = 43{,}839$ represents the number of identity fraud victims in 2004 for California, $x_2 = 16{,}062$ represents the number of identity fraud victims in 2004 for Florida, and so on. The sum of the numbers of identity fraud victims for these six states is

$$\begin{aligned}\Sigma x &= x_1 + x_2 + x_3 + x_4 + x_5 + x_6 \\ &= 43{,}839 + 16{,}062 + 11{,}138 + 17{,}680 + 6956 + 26{,}454 = 122{,}129\end{aligned}$$

Note that the given data includes only six states. Hence, it represents a sample. Because the data set contains six values, $n = 6$. Substituting the values of Σx and n in the sample formula, we obtain the mean number of identity fraud victims in 2004 for these six states:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{122{,}129}{6} = \mathbf{20{,}354.83}$$

Thus, the mean number of identity fraud victims in 2004 for these six states is 20,354.83. ■

■ EXAMPLE 3–2

Calculating the population mean for ungrouped data.

The following are the ages of all eight employees of a small company:

53 32 61 27 39 44 49 57

Find the mean age of these employees.

Solution Because the given data set includes *all* eight employees of the company, it represents the population. Hence, $N = 8$.

$$\Sigma x = 53 + 32 + 61 + 27 + 39 + 44 + 49 + 57 = 362$$

The population mean is

$$\mu = \frac{\Sigma x}{N} = \frac{362}{8} = \mathbf{45.25\ years}$$

Thus, the mean age of all eight employees of this company is 45.25 years, or 45 years and 3 months. ■

Reconsider Example 3–2. If we take a sample of three employees from this company and calculate the mean age of those three employees, this mean will be denoted by $\bar{x}$. Suppose the three values included in the sample are 32, 39, and 57. Then, the mean age for this sample is

$$\bar{x} = \frac{32 + 39 + 57}{3} = 42.67 \text{ years}$$

If we take a second sample of three employees of this company, the value of $\bar{x}$ will (most likely) be different. Suppose the second sample includes the values 53, 27, and 44. Then, the mean age for this sample is

$$\bar{x} = \frac{53 + 27 + 44}{3} = 41.33 \text{ years}$$

Consequently, we can state that the value of the population mean μ is constant. However, the value of the sample mean $\bar{x}$ varies from sample to sample. The value of $\bar{x}$ for a particular sample depends on what values of the population are included in that sample.

Sometime a data set may contain a few very small or a few very large values. As mentioned in Chapter 2 on page 58, such values are called *outliers* or *extreme values.*

A major shortcoming of the mean as a measure of central tendency is that it is very sensitive to outliers. Example 3–3 illustrates this point.

■ EXAMPLE 3–3

Illustrating the effect of an outlier on the mean.

Table 3.2 lists the total philanthropic givings (in million dollars) by six donors during their lifetimes until 2004.

Table 3.2 Total Philanthropic Givings in Lifetime

Donors	Total Philanthropic Giving in Lifetime (millions of dollars)
Bill and Melinda Gates	27,976
Warren Buffett	2730
George Soros	5171
Michael and Susan Dell	1230
Walton Family	1000
Ted Turner	1200

Source: Business Week, November 29, 2004.

Notice that the lifetime givings of Bill and Melinda Gates are very large compared to the lifetime givings of other donors. Hence, it is an outlier. Show how the inclusion of this outlier affects the value of the mean.

Solution If we do not include the lifetime givings of Bill and Melinda Gates (the outlier), the mean of the lifetime givings of the remaining five donors is

$$\text{Mean} = \frac{2730 + 5171 + 1230 + 1000 + 1200}{5} = \frac{11{,}331}{5} = \mathbf{\$2266.20 \text{ million}}$$

Now, to see the impact of the outlier on the value of the mean, we include the lifetime givings of Bill and Melinda Gates and find the mean lifetime givings of the six donors. This mean is

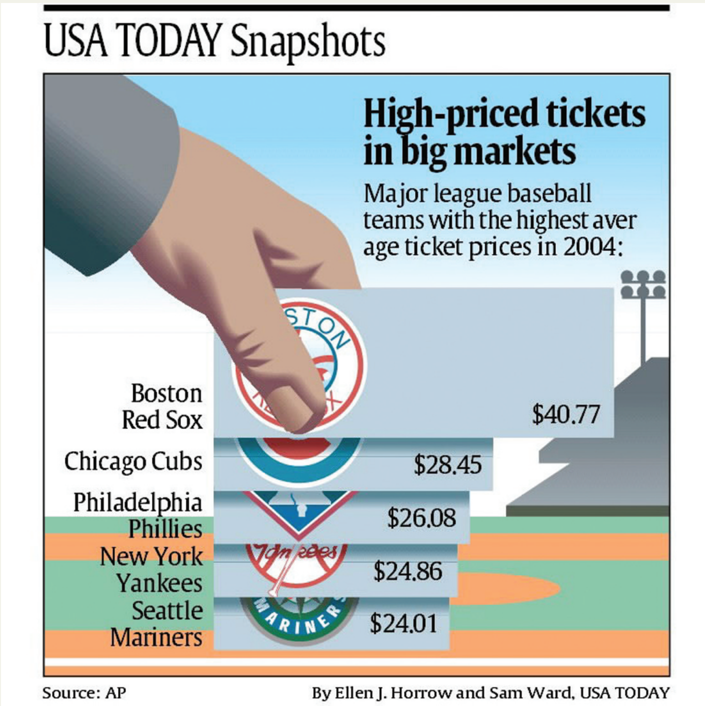
$$\text{Mean} = \frac{27{,}976 + 2730 + 5171 + 1230 + 1000 + 1200}{6} = \frac{39{,}307}{6} = \mathbf{\$6551.17 \text{ million}}$$

Thus, including the lifetime givings of Bill and Melinda Gates causes almost a threefold increase in the value of the mean, as it changes from \$2266.20 million to \$6551.17 million. ■

CASE STUDY

3–1

HIGH-PRICED TICKETS IN BIG MARKETS



Source: USA TODAY, April 26, 2004. Copyright © 2004, *USA TODAY*. Chart reproduced with permission.

The above chart, reproduced from *USA TODAY*, shows the average ticket prices of the five Major League Baseball teams that had the highest average ticket prices in 2004. According to the information given in the chart, the highest average price for an MLB team was for the Boston Red Sox, which was $40.77. The Chicago Cubs had the second highest average ticket price of $28.45.

The preceding example should encourage us to be cautious. We should remember that the mean is not always the best measure of central tendency because it is heavily influenced by outliers. Sometimes other measures of central tendency give a more accurate impression of a data set. For example, when a data set has outliers, instead of using the mean, we can use either the trimmed mean (defined in Exercise 3.33) or the median (to be discussed next) as a measure of central tendency.

3.1.2 Median

Another important measure of central tendency is the **median**. It is defined as follows.

Definition

Median The *median* is the value of the middle term in a data set that has been ranked in increasing order.

As is obvious from the definition of the median, it divides a ranked data set into two equal parts. The calculation of the median consists of the following two steps:

1. Rank the data set in increasing order.
2. Find the middle term. The value of this term is the median.[1]

[1]The value of the middle term in a data set ranked in *decreasing* order will also give the value of the median.

Note that if the number of observations in a data set is *odd*, then the median is given by the value of the middle term in the ranked data. However, if the number of observations is *even*, then the median is given by the average of the values of the two middle terms.

■ EXAMPLE 3–4

Calculating the median for ungrouped data: odd number of data values.

The following data give the weight lost (in pounds) by a sample of five members of a health club at the end of two months of membership.

10 5 19 8 3

Find the median.

Solution First, we rank the given data in increasing order as follows:

3 5 8 10 19

Since there are five terms in the data set and the middle term is the third term, the median is given by the value of the third term in the ranked data.

3 5 8 10 19
↑
Median (8)

The median weight loss for this sample of five members of this health club is **8 pounds**. ■

■ EXAMPLE 3–5

Calculating the median for ungrouped data: even number of data values.

Table 3.3 lists the number of car thefts during 2003 in 12 cities.

Table 3.3 Number of Car Thefts in 2003 in 12 Cities

City	Number of Car Thefts
Phoenix-Mesa, Arizona	40,769
Washington, D.C.	33,956
Miami, Florida	21,088
Atlanta, Georgia	29,920
Chicago, Illinois	42,082
Kansas City, Kansas	11,669
Baltimore, Maryland	13,435
Detroit, Michigan	40,197
St. Louis, Missouri	18,215
Las Vegas, Nevada	18,103
Newark, New Jersey	14,413
Dallas, Texas	26,343

Source: National Insurance Crime Bureau.

Find the median for these data.

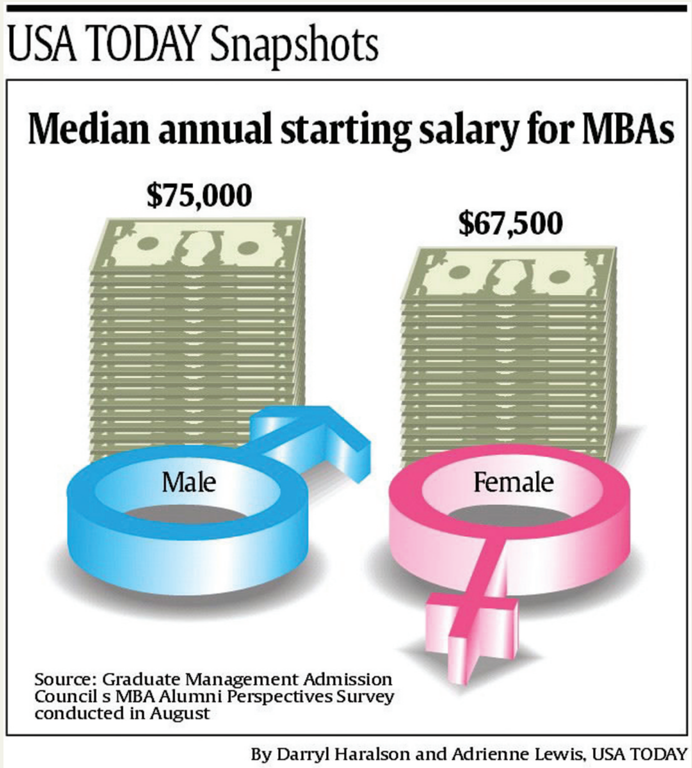
Solution First we rank the given data on car thefts in increasing order as follows:

11,669 13,435 14,413 18,103 18,215 21,088 26,343 29,920 33,956 40,197 40,769 42,082

There are 12 values in the data set. Because there is an even number of values in the data set, the median will be given by the mean of the two middle values. The two middle values

CASE STUDY 3–2

MEDIAN ANNUAL STARTING SALARY FOR MBAs



Source: USA TODAY, February 4, 2004. Copyright © 2004, *USA TODAY.* Chart reproduced with permission.

The above chart, reproduced from *USA TODAY*, shows the median annual starting salary of MBAs. These salaries are based on a survey conducted in August 2003. According to this survey, the median starting salary for males with an MBA degree was $75,000 and that of females was $67,500.

are the sixth and seventh values in the above arranged data, which are 21,088 and 26,343. The median, which is given by the average of these two values, is calculated below.

11,669 13,435 14,413 18,103 18,215 **21,088 ↑ 26,343** 29,920 33,956 40,197 40,769 42,082

Median

$$\text{Median} = \frac{21{,}088 + 26{,}343}{2} = \frac{47{,}431}{2} = \mathbf{23{,}715.50\ car\ thefts}$$

Thus, the median number of car thefts in 2003 for these 12 cities was 23,715.50. ■

The median gives the center of a histogram, with half of the data values to the left of the median and half to the right of the median. The advantage of using the median as a measure of central tendency is that it is not influenced by outliers. Consequently, the median is preferred over the mean as a measure of central tendency for data sets that contain outliers.

3.1.3 Mode

Mode is a French word that means *fashion*—an item that is most popular or common. In statistics, the mode represents the most common value in a data set.

Definition

Mode The *mode* is the value that occurs with the highest frequency in a data set.

■ EXAMPLE 3–6

Calculating the mode for ungrouped data.

The following data give the speeds (in miles per hour) of eight cars that were stopped on I-95 for speeding violations.

77	82	74	81	79	84	74	78

Find the mode.

Solution In this data set, 74 occurs twice and each of the remaining values occurs only once. Because 74 occurs with the highest frequency, it is the mode. Therefore,

Mode = **74 miles per hour** ■

A major shortcoming of the mode is that a data set may have none or may have more than one mode, whereas it will have only one mean and only one median. For instance, a data set with each value occurring only once has no mode. A data set with only one value occurring with the highest frequency has only one mode. The data set in this case is called **unimodal**. A data set with two values that occur with the same (highest) frequency has two modes. The distribution, in this case, is said to be **bimodal**. If more than two values in a data set occur with the same (highest) frequency, then the data set contains more than two modes and it is said to be **multimodal**.

■ EXAMPLE 3–7

Data set with no mode.

Last year's incomes of five randomly selected families were $46,150, $95,750, $64,985, $87,490, and $53,740. Find the mode.

Solution Because each value in this data set occurs only once, this data set contains **no mode**. ■

■ EXAMPLE 3–8

Data set with two modes.

The prices of the same brand of television set at eight stores are found to be $895, $886, $903, $895, $870, $905, $870, and $899. Find the mode.

Solution In this data set, each of the two values $895 and $870 occurs twice and each of the remaining values occurs only once. Therefore, this data set has two modes: **$895** and **$870**. ■

■ EXAMPLE 3–9

Data set with three modes.

The ages of 10 randomly selected students from a class are 21, 19, 27, 22, 29, 19, 25, 21, 22, and 30. Find the mode.

Solution This data set has three modes: **19**, **21**, and **22**. Each of these three values occurs with a (highest) frequency of 2. ■

One advantage of the mode is that it can be calculated for both kinds of data, quantitative and qualitative, whereas the mean and median can be calculated for only quantitative data.

■ EXAMPLE 3–10

Finding the mode for qualitative data.

The status of five students who are members of the student senate at a college are senior, sophomore, senior, junior, senior. Find the mode.

Solution Because **senior** occurs more frequently than the other categories, it is the mode for this data set. We cannot calculate the mean and median for this data set. ■

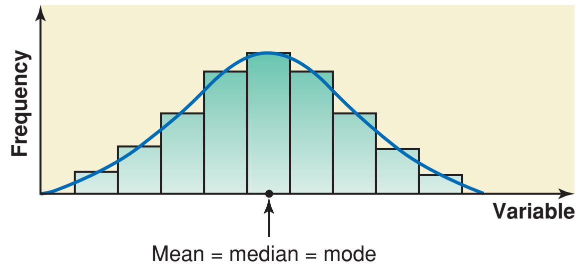
To sum up, we cannot say for sure which of the three measures of central tendency is a better measure overall. Each of them may be better under different situations. Probably the mean is the most used measure of central tendency, followed by the median. The mean has the advantage that its calculation includes each value of the data set. The median is a better measure when a data set includes outliers. The mode is simple to locate, but it is not of much use in practical applications.

3.1.4 Relationships among the Mean, Median, and Mode

As discussed in Chapter 2, two of the many shapes that a histogram or a frequency distribution curve can assume are symmetric and skewed. This section describes the relationships among the mean, median, and mode for three such histograms and frequency distribution curves. Knowing the values of the mean, median, and mode can give us some idea about the shape of a frequency distribution curve.

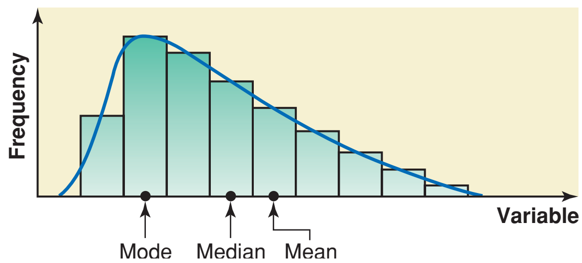
1. For a symmetric histogram and frequency distribution curve with one peak (see Figure 3.2), the values of the mean, median, and mode are identical, and they lie at the center of the distribution.

Figure 3.2 Mean, median, and mode for a symmetric histogram and frequency distribution curve.



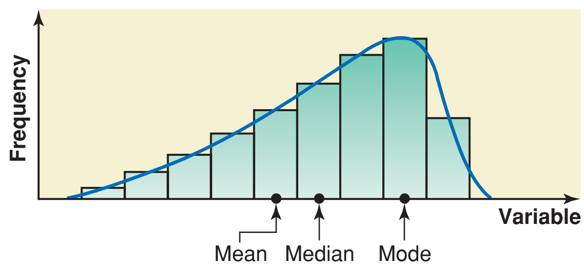
2. For a histogram and a frequency distribution curve skewed to the right (see Figure 3.3), the value of the mean is the largest, that of the mode is the smallest, and the value of the median lies between these two. (Notice that the mode always occurs at the peak point.) The value of the mean is the largest in this case because it is sensitive to outliers that occur in the right tail. These outliers pull the mean to the right.

Figure 3.3 Mean, median, and mode for a histogram and frequency distribution curve skewed to the right.



3. If a histogram and a frequency distribution curve are skewed to the left (see Figure 3.4), the value of the mean is the smallest and that of the mode is the largest, with the value of the median lying between these two. In this case, the outliers in the left tail pull the mean to the left.

Figure 3.4 Mean, median, and mode for a histogram and frequency distribution curve skewed to the left.



EXERCISES

CONCEPTS AND PROCEDURES

3.1 Explain how the value of the median is determined for a data set that contains an odd number of observations and for a data set that contains an even number of observations.

3.2 Briefly explain the meaning of an outlier. Is the mean or the median a better measure of central tendency for a data set that contains outliers? Illustrate with the help of an example.

3.3 Using an example, show how outliers can affect the value of the mean.

3.4 Which of the three measures of central tendency (the mean, the median, and the mode) can be calculated for quantitative data only, and which can be calculated for both quantitative and qualitative data? Illustrate with examples.

3.5 Which of the three measures of central tendency (the mean, the median, and the mode) can assume more than one value for a data set? Give an example of a data set for which this summary measure assumes more than one value.

3.6 Is it possible for a (quantitative) data set to have no mean, no median, or no mode? Give an example of a data set for which this summary measure does not exist.

3.7 Explain the relationships among the mean, median, and mode for symmetric and skewed histograms. Illustrate these relationships with graphs.

3.8 Prices of cars have a distribution that is skewed to the right with outliers in the right tail. Which of the measures of central tendency is the best to summarize this data set? Explain.

3.9 The following data set belongs to a population:

5	−7	2	0	−9	16	10	7

Calculate the mean, median, and mode.

3.10 The following data set belongs to a sample:

14	18	−10	8	8	−16

Calculate the mean, median, and mode.

APPLICATIONS

Exercises 3.11 and 3.12 are based on the following data.

The following table gives the sticker prices and dealer's prices for base models of 10 two-door small cars as of January 2004.

Make/Model	Sticker Price	Dealer's Cost
Acura RSX	$20,025	$18,261
Chevrolet Cavalier LS	15,820	14,792
Ford Focus ZX3 Comfort	14,495	13,566
Honda Civic EX	16,860	15,410
Hyundai Accent GL	10,899	10,201
Mini Cooper	16,449	14,887
Oldsmobile Alero GL2	21,900	20,039
Pontiac Sunfire	14,930	13,810
Toyota Celica GT	17,390	15,735
Volkswagen Golf GL	15,580	14,593

Sources: MONEY, March 2004.

3.11 Calculate the mean and median for the data on sticker prices for these cars.

3.12 Find the mean and median for the data on dealers' costs for these cars.

3.13 The following data give the number of workers (in thousands) employed by small companies in all 50 states (*USA Today*, June 20, 2005). The data are entered in alphabetic order for states.

786	128	930	476	6800	981	759	171	2900	1500
253	259	2600	1300	642	588	734	853	292	1100
1500	2000	1200	452	1200	210	383	402	302	1800

319	3800	1600	161	2300	645	736	2500	238	739
189	1000	3800	430	162	1400	1200	303	1300	123

a. Calculate the mean and median for these data. Are these values of the mean and median the sample statistics or population parameters?

b. Do these data have a mode? Explain.

3.14 The following data give the 2004 profits (in millions of dollars) of the nine computer and office equipment companies included in the Fortune 1000 (*FORTUNE*, April 18, 2005). The data, entered in that order, are for International Business Machines, Hewlett-Packard, Dell, Xerox, Sun Microsystems, Apple Computer, NCR, Pitney Bowes, and Gateway.

8430	3497	3043	859	−388	276	290	481	−568

Find the mean and median for these data. Do these data have a mode?

3.15 The following data give the annual salaries (in dollars) of governors of 13 western states for 2004 (*Source:* Council of State Governments, *The Book of the States, 2004; The New York Times Almanac,* 2005). The salaries, listed in that order are for AK, HI, CA, OR, WA, ID, MT, WY, CO, UT, NV, AZ, and NM.

85,776	94,780	175,000	93,600	139,087
98,500	93,089	130,000	90,000	100,600
117,000	95,000	110,000		

Find the mean and median for these data.

3.16 The following data give the numbers of car thefts that occurred in a city during the past 12 days.

6	3	7	11	4	3	8	7	2	6	9	15

Find the mean, median, and mode.

3.17 The following data give the revenues (in millions of dollars) for the last available fiscal year for a sample of six charitable organizations that are related to serious diseases (*Forbes*, December 13, 2004). The values listed in that order are for Alzheimer's Association, American Cancer Society, American Diabetes Association, American Heart Association, American Lung Association, and Cystic Fibrosis Foundation.

136	816	192	513	158	152

Compute the mean and median. Do these data have a mode? Why or why not?

3.18 The following table gives the numbers of *takeaways* (recoveries of opponents' fumbles and interceptions of opponents' passes) during the 2004 season for all 16 teams in the National Conference of the National Football League.

Team	Takeaways
Carolina	38
Seattle	35
New Orleans	33
Philadelphia	28
Detroit	24
N.Y. Giants	28
Atlanta	32
Arizona	30
Minnesota	22
Washington	26
Chicago	29
Tampa Bay	27
Green Bay	15
Dallas	22
San Francisco	21
St. Louis	15

Source: USA TODAY, January 5, 2005.

Compute the mean and median for the data on *takeaways.* Do these data have a mode? Why or why not?

3.19 Due to antiquated equipment and frequent windstorms, the town of Oak City often suffers power outages. The following data give the numbers of power outages for the past 12 months.

4	5	7	3	2	0	2	3	2	1	2	4

Compute the mean, median, and mode for these data.

3.20 A brochure from the department of public safety in a northern state recommends that motorists should carry 12 items (flashlights, blankets, and so forth) in their vehicles for emergency use while driving in winter. The following data give the number of items out of these 12 that were carried in their vehicles by 15 randomly selected motorists.

5	3	7	8	0	1	0	5	12	10	7	6	7	11	9

Find the mean, median, and mode for these data. Are the values of these summary measures population parameters or sample statistics? Explain.

3.21 Nixon Corporation manufactures computer monitors. The following data are the numbers of computer monitors produced at the company for a sample of 10 days.

24	32	27	23	35	33	29	40	23	28

Calculate the mean, median, and mode for these data.

3.22 The Tri-City School District has instituted a zero-tolerance policy for students carrying any objects that could be used as weapons. The following data give the number of students suspended during each of the past 12 weeks for violating this school policy.

15	9	12	11	7	6	9	10	14	3	6	5

Calculate the mean, median, and mode for these data.

3.23 The following data give the numbers of casinos in 11 states as of December 21, 2003 (*USA TODAY*, July 16, 2004). The data entered in that order are for CO, IL, IN, IA, LA, MI, MS, MO, NV, NJ, and SD.

44	9	10	13	18	3	29	11	256	12	38

a. Calculate the mean and median for these data.
b. Do these data contain an outlier? If so, drop the outlier and recalculate the mean and median. Which of these two summary measures changes by a larger amount when you drop the outlier?
c. Which is the better summary measure for these data, the mean or the median? Explain.

3.24 The following data, based on the AAA Foundation for Traffic Safety estimates, give the number of fatal crashes caused by road debris from 1999 to 2001 in 10 states with the most such accidents (*USA TODAY*, June 16, 2004). The data entered in that order are for TX, FL, MO, VA, OK, MD, AZ, LA, WI, and IN.

33	17	13	6	6	5	5	4	3	3

Compute the mean and median for these data. Do these data have modes? Why or why not?

***3.25** One property of the mean is that if we know the means and sample sizes of two (or more) data sets, we can calculate the **combined mean** of both (or all) data sets. The combined mean for two data sets is calculated by using the formula

$$\text{Combined mean} = \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

where n_1 and n_2 are the sample sizes of the two data sets and $\bar{x}_1$ and $\bar{x}_2$ are the means of the two data sets, respectively. Suppose a sample of 10 statistics books gave a mean price of \$95 and a sample of 8 mathematics books gave a mean price of \$104. Find the combined mean. (*Hint:* For this example: $n_1 = 10$, $n_2 = 8$, $\bar{x}_1 = \$95$, $\bar{x}_2 = \$104$.)

***3.26** Twenty business majors and 18 economics majors go bowling. Each student bowls one game. The scorekeeper announces that the mean score for the 18 economics majors is 144 and the mean score for the entire group of 38 students is 150. Find the mean score for the 20 business majors.

***3.27** For any data, the sum of all values is equal to the product of the sample size and mean; that is, $\Sigma x = n\bar{x}$. Suppose the average amount of money spent on shopping by 10 persons during a given week is \$105.50. Find the total amount of money spent on shopping by these 10 persons.

***3.28** The mean 2005 income for five families was \$79,520. What was the total 2005 income of these five families?

***3.29** The mean age of six persons is 46 years. The ages of five of these six persons are 57, 39, 44, 51, and 37 years. Find the age of the sixth person.

***3.30** Seven airline passengers in economy class on the same flight paid an average of \$361 per ticket. Because the tickets were purchased at different times and from different sources, the prices varied. The first five passengers paid \$420, \$210, \$333, \$695, and \$485. The sixth and seventh tickets were purchased by a couple who paid identical fares. What price did each of them pay?

***3.31** Consider the following two data sets.

Data Set I:	12	25	37	8	41
Data Set II:	19	32	44	15	48

Notice that each value of the second data set is obtained by adding 7 to the corresponding value of the first data set. Calculate the mean for each of these two data sets. Comment on the relationship between the two means.

***3.32** Consider the following two data sets.

Data Set I:	4	8	15	9	11
Data Set II:	8	16	30	18	22

Notice that each value of the second data set is obtained by multiplying the corresponding value of the first data set by 2. Calculate the mean for each of these two data sets. Comment on the relationship between the two means.

***3.33** The **trimmed mean** is calculated by dropping a certain percentage of values from each end of a ranked data set. The trimmed mean is especially useful as a measure of central tendency when a data set contains a few outliers at each end. Suppose the following data give the ages of 10 employees of a company:

47	53	38	26	39	49	19	67	31	23

To calculate the 10% trimmed mean, first rank these data values in increasing order; then drop 10% of the smallest values and 10% of the largest values. The mean of the remaining 80% of the values will give the 10% trimmed mean. Note that this data set contains 10 values, and 10% of 10 is 1. Thus, if we drop the smallest value and the largest value from this data set, the mean of the remaining 8 values will be called the 10% trimmed mean. Calculate the 10% trimmed mean for this data set.

***3.34** The following data give the prices (in thousands of dollars) of 20 houses sold recently in a city.

184	297	365	309	245	387	369	438	195	390
323	578	410	679	307	271	457	795	259	590

Find the 20% trimmed mean for this data set.

3.35** In some applications, certain values in a data set may be considered more important than others. For example, to determine students' grades in a course, an instructor may assign a weight to the final exam twice as much as to each of the other exams. In such cases, it is more appropriate to use the **weighted mean**. In general, for a sequence of n data values $x_1, x_2, \ldots x_n$ that are assigned weights $w_1, w_2, \ldots w_n$, respectively, the ***weighted mean is found by the formula

$$\text{Weighted mean} = \frac{\Sigma xw}{\Sigma w}$$

where Σxw is obtained by multiplying each data value by its weight and then adding the products. Suppose an instructor gives two exams and a final, assigning the final exam a weight twice that of each of the other exams. Find the weighted mean for a student who scores 73 and 67 on the first two exams, and 85 on the final. (*Hint:* Here, $x_1 = 73$, $x_2 = 67$, $x_3 = 85$, and $w_1 = w_2 = 1$, and $w_3 = 2$.)

***3.36** When studying phenomena such as inflation or population changes, which involve periodic increases or decreases, the **geometric mean** is used to find the average change over the entire period under study. To calculate the geometric mean of a sequence of n values $x_1, x_2, \ldots x_n$, we multiply them together and then find the nth root of this product. Thus

$$\text{Geometric mean} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n}$$

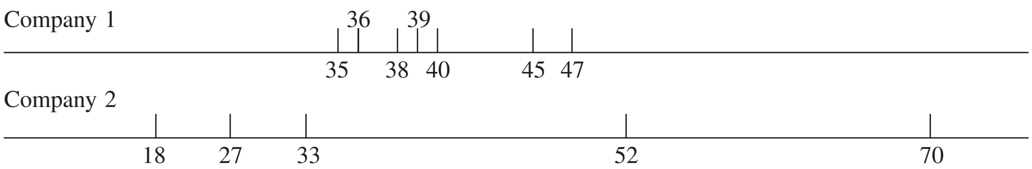
Suppose that the inflation rates for the last five years are 4%, 3%, 5%, 6%, and 8%, respectively. Thus at the end of the first year, the price index will be 1.04 times the price index at the beginning of the year, and so on. Find the mean rate of inflation over the five-year period by finding the geometric mean of the data set 1.04, 1.03, 1.05, 1.06, and 1.08. (*Hint:* Here, $n = 5$, $x_1 = 1.04$, $x_2 = 1.03$, etc. Use the $x^{1/n}$ key on your calculator to find the fifth root. Note that the mean inflation rate will be obtained by subtracting 1 from the geometric mean.)

3.2 Measures of Dispersion for Ungrouped Data

The measures of central tendency, such as the mean, median, and mode, do not reveal the whole picture of the distribution of a data set. Two data sets with the same mean may have completely different spreads. The variation among the values of observations for one data set may be much larger or smaller than for the other data set. (Note that the words *dispersion*, *spread*, and *variation* have the same meaning.) Consider the following two data sets on the ages of all workers in each of two small companies.

Company 1:	47	38	35	40	36	45	39
Company 2:	70	33	18	52	27		

The mean age of workers in both these companies is the same, 40 years. If we do not know the ages of individual workers in these two companies and are told only that the mean age of the workers in both companies is the same, we may deduce that the workers in these two companies have a similar age distribution. But as we can observe, the variation in the workers' ages for each of these two companies is very different. As illustrated in the diagram, the ages of the workers in the second company have a much larger variation than the ages of the workers in the first company.



Thus, the mean, median, or mode by itself is usually not a sufficient measure to reveal the shape of the distribution of a data set. We also need a measure that can provide some information about the variation among data values. The measures that help us learn about the spread of a data set are called the **measures of dispersion**. The measures of central tendency and dispersion taken together give a better picture of a data set than the measures of central tendency alone. This section discusses three measures of dispersion: range, variance, and standard deviation.

3.2.1 Range

The **range** is the simplest measure of dispersion to calculate. It is obtained by taking the difference between the largest and the smallest values in a data set.

Finding Range for Ungrouped Data

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

■ EXAMPLE 3–11

Table 3.4 gives the total areas in square miles of the four western South-Central states of the United States.

Calculating the range for ungrouped data.

Table 3.4

State	Total Area (square miles)
Arkansas	53,182
Louisiana	49,651
Oklahoma	69,903
Texas	267,277

Find the range for this data set.

Solution The maximum total area for a state in this data set is 267,277 square miles, and the smallest area is 49,651 square miles. Therefore,

$$\begin{aligned}\text{Range} &= \text{Largest value} - \text{Smallest value}\\ &= 267{,}277 - 49{,}651 = \mathbf{217{,}626\ square\ miles}\end{aligned}$$

Thus, the total areas of these four states are spread over a range of 217,626 square miles. ■

The range, like the mean, has the disadvantage of being influenced by outliers. In Example 3–11, if the state of Texas with a total area of 267,277 square miles is dropped, the range decreases from 217,626 square miles to 20,252 square miles. Consequently, the range is not a good measure of dispersion to use for a data set that contains outliers.

Another disadvantage of using the range as a measure of dispersion is that its calculation is based on two values only: the largest and the smallest. All other values in a data set are ignored when calculating the range. Thus, the range is not a very satisfactory measure of dispersion.

3.2.2 Variance and Standard Deviation

The **standard deviation** is the most used measure of dispersion. The value of the standard deviation tells how closely the values of a data set are clustered around the mean. In general, a lower value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively smaller range around the mean. In contrast, a larger value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively larger range around the mean.

The *standard deviation is obtained by taking the positive square root of the* **variance**. The variance calculated for population data is denoted by $\boldsymbol{\sigma^2}$ (read as *sigma squared*),[2] and the variance calculated for sample data is denoted by $\boldsymbol{s^2}$. Consequently, the standard deviation calculated for population data is denoted by $\boldsymbol{\sigma}$, and the standard deviation calculated for sample data is denoted by $\boldsymbol{s}$. Following are what we will call the *basic formulas* that are used to calculate the variance:[3]

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

where σ^2 is the population variance and s^2 is the sample variance.

The quantity $x - \mu$ or $x - \bar{x}$ in the above formulas is called the *deviation* of the x value from the mean. The sum of the deviations of the x values from the mean is always zero; that is, $\Sigma(x-\mu) = 0$ and $\Sigma(x-\bar{x}) = 0$.

For example, suppose the midterm scores of a sample of four students are 82, 95, 67, and 92. Then, the mean score for these four students is

$$\bar{x} = \frac{82 + 95 + 67 + 92}{4} = 84$$

The deviations of the four scores from the mean are calculated in Table 3.5. As we can observe from the table, the sum of the deviations of the x values from the mean is zero; that is, $\Sigma(x-\bar{x}) = 0$. For this reason we square the deviations to calculate the variance and standard deviation.

Table 3.5

x	$x - \bar{x}$
82	$82 - 84 = -2$
95	$95 - 84 = +11$
67	$67 - 84 = -17$
92	$92 - 84 = +8$
	$\Sigma(x-\bar{x}) = 0$

[2]Note that Σ is uppercase sigma and σ is lowercase sigma of the Greek alphabet.

[3]From the formula for σ^2, it can be stated that the population variance is the mean of the squared deviations of x values from the mean. However, this is not true for the variance calculated for a sample data set.

From the computational point of view, it is easier and more efficient to use *short-cut formulas* to calculate the variance and standard deviation. By using the short-cut formulas, we reduce the computation time and round-off errors. Use of the basic formulas for ungrouped data is illustrated in Section A3.1.1 of Appendix 3.1 of this chapter. The short-cut formulas for calculating the variance and standard deviation are given next.

Short-Cut Formulas for the Variance and Standard Deviation for Ungrouped Data

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$$

where σ^2 is the population variance and s^2 is the sample variance.

The standard deviation is obtained by taking the positive square root of the variance.

Population standard deviation: $\sigma = \sqrt{\sigma^2}$

Sample standard deviation: $s = \sqrt{s^2}$

Note that the denominator in the formula for the population variance is N, but that in the formula for the sample variance it is $n - 1$.[4]

■ EXAMPLE 3–12

Calculating the sample variance and standard deviation for ungrouped data.

The following table, based on *Forbes* magazine's list of the wealthiest people in the world, gives the total wealth (in billions of dollars) of five persons (*USA TODAY*, March 11, 2005).

Billinaire	Total Wealth (billions of dollars)
Bill Gates	46.5
Helen Walton	18.0
Michael Dell	16.0
Keith Rupert Murdoch	7.8
George Soros	7.2

Find the variance and standard deviation for these data.

Solution Let x denote the total wealth (in billions of dollars) of a person. The values of Σx and Σx^2 are calculated in Table 3.6.

Table 3.6

x	x^2
46.5	2162.25
18.0	324.00
16.0	256.00
7.8	60.84
7.2	51.84
$\Sigma x = 95.5$	$\Sigma x^2 = 2854.93$

The calculation of the variance involves the following steps.

[4]The reason that the denominator in the sample formula is $n - 1$ and not n follows: The sample variance underestimates the population variance when the denominator in the sample formula for variance is n. However, the sample variance does not underestimate the population variance if the denominator in the sample formula for variance is $n - 1$. In Chapter 8 we will learn that $n - 1$ is called the degrees of freedom.

Step 1. *Calculate* $\sum x$.

The sum of the values in the first column of Table 3.6 gives the value of $\sum x$, which is 95.5.

Step 2. *Find* $\sum x^2$.

The value of $\sum x^2$ is obtained by squaring each value of x and then adding the squared values. The results of this step are shown in the second column of Table 3.6. Notice that $\sum x^2 = 2854.93$.

Step 3. *Determine the variance.*

Substitute all the values in the variance formula and simplify. Because the given data are on the wealth of a sample of five persons, we use the formula for the sample variance.

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{2854.93 - \frac{(95.5)^2}{5}}{5-1} = \frac{2854.93 - 1824.05}{4} = \mathbf{257.72}$$

Step 4. *Obtain the standard deviation.*

The standard deviation is obtained by taking the (positive) square root of the variance.

$$s = \sqrt{257.72} = \mathbf{16.05366} = \mathbf{\$16.05\ billion}$$

Thus, the standard deviation of the wealth of these five individuals is $16.05 billion. ■

Two Observations ►

1. **The values of the variance and the standard deviation are never negative**. That is, the numerator in the formula for the variance should never produce a negative value. Usually the values of the variance and standard deviation are positive, but if a data set has no variation, then the variance and standard deviation are both zero. For example, if four persons in a group are the same age—say, 35 years—then the four values in the data set are

   35 35 35 35

   If we calculate the variance and standard deviation for these data, their values are zero. This is because there is no variation in the values of this data set.

2. **The measurement units of variance are always the square of the measurement units of the original data**. This is so because the original values are squared to calculate the variance. In Example 3–12, the measurement units of the original data are billions of dollars. However, the measurement units of the variance are squared billions of dollars, which, of course, does not make any sense. Thus, the variance of the wealth of these five persons in Example 3–12 is 257.72 squared billion dollars. But the measurement units of the standard deviation are the same as the measurement units of the original data because the standard deviation is obtained by taking the square root of the variance.

■ EXAMPLE 3–13

Calculating the population variance and standard deviation for ungrouped data.

Following are the 2005 earnings (in thousands of dollars) before taxes for all six employees of a small company.

48.50 38.40 65.50 22.60 79.80 54.60

Calculate the variance and standard deviation for these data.

Solution Let x denote the 2005 earnings before taxes of an employee of this company. The values of Σx and Σx^2 are calculated in Table 3.7.

Table 3.7

x	x^2
48.50	2352.25
38.40	1474.56
65.50	4290.25
22.60	510.76
79.80	6368.04
54.60	2981.16
$\Sigma x = 309.40$	$\Sigma x^2 = 17{,}977.02$

Because the data are on earnings of *all* employees of this company, we use the population formula to compute the variance. Thus, the variance is

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N} = \frac{17{,}977.02 - \frac{(309.40)^2}{6}}{6} = \mathbf{337.0489}$$

The standard deviation is obtained by taking the (positive) square root of the variance:

$$\sigma = \sqrt{337.0489} = \mathbf{\$18.359\ thousand = \$18{,}359}$$

Thus, the standard deviation of the 2005 earnings of all six employees of this company is $18,359. ■

◀ *Warning*

Note that Σx^2 is not the same as $(\Sigma x)^2$. The value of Σx^2 is obtained by squaring the x values and then adding them. The value of $(\Sigma x)^2$ is obtained by squaring the value of Σx.

The uses of the standard deviation are discussed in Section 3.4. Later chapters explain how the mean and the standard deviation taken together can help in making inferences about the population.

3.2.3 Population Parameters and Sample Statistics

A numerical measure such as the mean, median, mode, range, variance, or standard deviation calculated for a population data set is called a *population parameter*, or simply a **parameter**. A summary measure calculated for a sample data set is called a *sample statistic*, or simply a **statistic**. Thus, μ and σ are population parameters, and $\bar{x}$ and s are sample statistics. As an illustration, $\bar{x} = 20{,}354.83$ in Example 3–1 is a sample statistic, and $\mu = 45.25$ years in Example 3–2 is a population parameter. Similarly, $s = \$16.05$ billion in Example 3–12 is a sample statistic, whereas $\sigma = \$18{,}359$ in Example 3–13 is a population parameter.

EXERCISES

CONCEPTS AND PROCEDURES

3.37 The range, as a measure of spread, has the disadvantage of being influenced by outliers. Illustrate this with an example.

3.38 Can the standard deviation have a negative value? Explain.

3.39 When is the value of the standard deviation for a data set zero? Give one example. Calculate the standard deviation for the example and show that its value is zero.

3.40 Briefly explain the difference between a population parameter and a sample statistic. Give one example of each.

3.41 The following data set belongs to a population:

5 −7 2 0 −9 16 10 7

Calculate the range, variance, and standard deviation.

3.42 The following data set belongs to a sample:

14 18 −10 8 8 −16

Calculate the range, variance, and standard deviation.

APPLICATIONS

3.43 The following data give the number of shoplifters apprehended during each of the past eight weeks at a large department store.

7 10 8 3 15 12 6 11

a. Find the mean for these data. Calculate the deviations of the data values from the mean. Is the sum of these deviations zero?
b. Calculate the range, variance, and standard deviation.

3.44 The following data give the prices of seven textbooks randomly selected from a university bookstore.

$89 $67 $104 $113 $36 $121 $147

a. Find the mean for these data. Calculate the deviations of the data values from the mean. Is the sum of these deviations zero?
b. Calculate the range, variance, and standard deviation.

3.45 The following data give the numbers of car thefts that occurred in a city in the past 12 days.

6 3 7 11 4 3 8 7 2 6 9 15

Calculate the range, variance, and standard deviation.

3.46 During the 2004 presidential election campaign, spending on television commercials was high, particularly in key states where the vote was expected to be close. The following data give the expenditures on television commercials (in millions of dollars) by all candidates in 10 states where such spending was the highest. The data, entered in that order, are for Florida, California, Ohio, Pennsylvania, Missouri, New Jersey, Delaware, Michigan, Wisconsin, and North Carolina (*USA TODAY*, November 26, 2004).

236.7 190.7 166.8 133.9 98.0 88.3 65.3 61.6 54.4 51.6

Find the range, variance, and standard deviation for these data.

3.47 The following data give the numbers of pieces of junk mail received by 10 families during the past month.

41 33 28 21 29 19 14 31 39 36

Find the range, variance, and standard deviation.

3.48 The following data give the number of highway collisions with large wild animals, such as deer or moose, in one of the northeastern states during each week of a nine-week period.

7 10 3 8 2 5 7 4 9

Find the range, variance, and standard deviation.

3.49 Attacks by stinging insects, such as bees or wasps, may become medical emergencies if either the victim is allergic to venom or multiple stings are involved. The following data give the number of patients treated each week for such stings in a large regional hospital during 13 weeks last summer.

1 5 2 3 0 4 1 7 0 1 2 0 1

Compute the range, variance, and standard deviation for these data.

3.50 The following data give the number of hot dogs consumed by 10 participants in a hot-dog-eating contest.

21 17 32 8 20 15 17 23 9 18

Calculate the range, variance, and standard deviation for these data.

3.51 Following are the temperatures (in degrees Fahrenheit) observed during eight wintry days in a midwestern city:

23	14	6	−7	−2	11	16	19

Compute the range, variance, and standard deviation.

3.52 The following data give the numbers of hours spent partying by 10 randomly selected college students during the past week.

7	14	5	0	9	7	10	4	0	8

Compute the range, variance, and standard deviation.

3.53 The following data, based on *Forbes* Magazine's rankings of the wealthiest people in the world, give the net worth (in billions of dollars) of the 10 wealthiest people in the world (*USA TODAY*, March 11, 2005). The data, entered in that order, are for Bill Gates, Warren Buffett, Lakshmi Mittal, Carlos Slim Helu, Prince Alwaleed Bin Talal Alsaud, Ingvar Kamprad, Paul Allen, Karl Albrecht, Lawrence Ellison, and S. Robson Walton.

46.5	44.0	25.0	23.8	23.7	23.0	21.0	18.5	18.4	18.3

Find the range, variance, and standard deviation for these data.

3.54 The following data give the average speeds (rounded to the nearest mile per hour) at the Indianapolis 500 auto race for the years 1995 to 2004 (*The New York Times 2005 Almanac*).

154	148	146	145	153	168	142	166	156	139

Find the range, variance, and standard deviation for these data.

3.55 The following data give the hourly wage rates of eight employees of a company.

12	12	12	12	12	12	12	12

Calculate the standard deviation. Is its value zero? If yes, why?

3.56 The following data are the ages (in years) of six students.

19	19	19	19	19	19

Calculate the standard deviation. Is its value zero? If yes, why?

***3.57** One disadvantage of the standard deviation as a measure of dispersion is that it is a measure of absolute variability and not of relative variability. Sometimes we may need to compare the variability of two different data sets that have different units of measurement. The **coefficient of variation** is one such measure. The coefficient of variation, denoted by CV, expresses standard deviation as a percentage of the mean and is computed as follows:

$$\text{For population data:} \quad CV = \frac{\sigma}{\mu} \times 100\%$$

$$\text{For sample data:} \quad CV = \frac{s}{\bar{x}} \times 100\%$$

The yearly salaries of all employees who work for a company have a mean of \$62,350 and a standard deviation of \$6820. The years of experience for the same employees have a mean of 15 years and a standard deviation of 2 years. Is the relative variation in the salaries greater or less than that in years of experience for these employees?

***3.58** The SAT scores of 100 students have a mean of 975 and a standard deviation of 105. The GPAs of the same 100 students have a mean of 3.16 and a standard deviation of .22. Is the relative variation in SAT scores greater or less than that in GPAs?

***3.59** Consider the following two data sets.

Data Set I:	12	25	37	8	41
Data Set II:	19	32	44	15	48

Note that each value of the second data set is obtained by adding 7 to the corresponding value of the first data set. Calculate the standard deviation for each of these two data sets using the formula for sample data. Comment on the relationship between the two standard deviations.

***3.60** Consider the following two data sets.

Data Set I:	4	8	15	9	11
Data Set II:	8	16	30	18	22

Note that each value of the second data set is obtained by multiplying the corresponding value of the first data set by 2. Calculate the standard deviation for each of these two data sets using the formula for population data. Comment on the relationship between the two standard deviations.

3.3 Mean, Variance, and Standard Deviation for Grouped Data

In Sections 3.1.1 and 3.2.2, we learned how to calculate the mean, variance, and standard deviation for ungrouped data. In this section, we will learn how to calculate the mean, variance, and standard deviation for grouped data.

3.3.1 Mean for Grouped Data

We learned in Section 3.1.1 that the mean is obtained by dividing the sum of all values by the number of values in a data set. However, if the data are given in the form of a frequency table, we no longer know the values of individual observations. Consequently, in such cases, we cannot obtain the sum of individual values. We find an approximation for the sum of these values using the procedure explained in the next paragraph and example. The formulas used to calculate the mean for grouped data follow.

Calculating Mean for Grouped Data

Mean for population data: $$\mu = \frac{\Sigma mf}{N}$$

Mean for sample data: $$\bar{x} = \frac{\Sigma mf}{n}$$

where m is the midpoint and f is the frequency of a class.

To calculate the mean for grouped data, first find the midpoint of each class and then multiply the midpoints by the frequencies of the corresponding classes. The sum of these products, denoted by Σmf, gives an approximation for the sum of all values. To find the value of the mean, divide this sum by the total number of observations in the data.

■ EXAMPLE 3–14

Calculating the population mean for grouped data.

Table 3.8 gives the frequency distribution of the daily commuting times (in minutes) from home to work for *all* 25 employees of a company.

Table 3.8

Daily Commuting Time (minutes)	Number of Employees
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Calculate the mean of the daily commuting times.

Solution Note that because the data set includes *all* 25 employees of the company, it represents the population. Table 3.9 shows the calculation of Σmf. Note that in Table 3.9, m denotes the midpoints of the classes.

Table 3.9

Daily Commuting Time (minutes)	f	m	mf
0 to less than 10	4	5	20
10 to less than 20	9	15	135
20 to less than 30	6	25	150
30 to less than 40	4	35	140
40 to less than 50	2	45	90
	$N = 25$		$\Sigma mf = 535$

To calculate the mean, we first find the midpoint of each class. The class midpoints are recorded in the third column of Table 3.9. The products of the midpoints and the corresponding frequencies are listed in the fourth column. The sum of the fourth column values, denoted by Σmf, gives the approximate total daily commuting time (in minutes) for all 25 employees. The mean is obtained by dividing this sum by the total frequency. Therefore,

$$\mu = \frac{\Sigma mf}{N} = \frac{535}{25} = \mathbf{21.40 \text{ minutes}}$$

Thus, the employees of this company spend an average of 21.40 minutes a day commuting from home to work. ■

What do the numbers 20, 135, 150, 140, and 90 in the column labeled mf in Table 3.9 represent? We know from this table that 4 employees spend 0 to less than 10 minutes commuting per day. If we assume that the time spent commuting by these 4 employees is evenly spread in the interval 0 to less than 10, then the midpoint of this class (which is 5) gives the mean time spent commuting by these 4 employees. Hence, $4 \times 5 = 20$ is the approximate total time (in minutes) spent commuting per day by these 4 employees. Similarly, 9 employees spend 10 to less than 20 minutes commuting per day, and the total time spent commuting by these 9 employees is approximately 135 minutes a day. The other numbers in this column can be interpreted the same way. Note that these numbers give the approximate commuting times for these employees based on the assumption of an even spread within classes. The total commuting time for all 25 employees is approximately 535 minutes. Consequently, 21.40 minutes is an approximate and not the exact value of the mean. We can find the exact value of the mean only if we know the exact commuting time for each of the 25 employees of the company.

■ EXAMPLE 3–15

Calculating the sample mean for grouped data.

Table 3.10 gives the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company.

Table 3.10

Number of Orders	Number of Days
10–12	4
13–15	12
16–18	20
19–21	14

Calculate the mean.

Solution Because the data set includes only 50 days, it represents a sample. The value of Σmf is calculated in Table 3.11.

Table 3.11

Number of Orders	f	m	mf
10–12	4	11	44
13–15	12	14	168
16–18	20	17	340
19–21	14	20	280
	$n = 50$		$\Sigma mf = 832$

The value of the sample mean is

$$\bar{x} = \frac{\Sigma mf}{n} = \frac{832}{50} = \mathbf{16.64\ orders}$$

Thus, this mail-order company received an average of 16.64 orders per day during these 50 days. ■

3.3.2 Variance and Standard Deviation for Grouped Data

Following are what we will call the *basic formulas* used to calculate the population and sample variances for grouped data:

$$\sigma^2 = \frac{\Sigma f(m - \mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\Sigma f(m - \bar{x})^2}{n - 1}$$

where σ^2 is the population variance, s^2 is the sample variance, and m is the midpoint of a class.

In either case, the standard deviation is obtained by taking the positive square root of the variance.

Again, the *short-cut formulas* are more efficient for calculating the variance and standard deviation. Section A3.1.2 of Appendix 3.1 at the end of this chapter shows how to use the basic formulas to calculate the variance and standard deviation for grouped data.

Short-Cut Formulas for the Variance and Standard Deviation for Grouped Data

$$\sigma^2 = \frac{\Sigma m^2 f - \dfrac{(\Sigma mf)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\Sigma m^2 f - \dfrac{(\Sigma mf)^2}{n}}{n - 1}$$

where σ^2 is the population variance, s^2 is the sample variance, and m is the midpoint of a class.

The standard deviation is obtained by taking the positive square root of the variance.

Population standard deviation: $\sigma = \sqrt{\sigma^2}$

Sample standard deviation: $s = \sqrt{s^2}$

Examples 3–16 and 3–17 illustrate the use of these formulas to calculate the variance and standard deviation.

Calculating the population variance and standard deviation for grouped data.

■ EXAMPLE 3–16

The following data, reproduced from Table 3.8 of Example 3–14, give the frequency distribution of the daily commuting times (in minutes) from home to work for all 25 employees of a company.

Daily Commuting Time (minutes)	Number of Employees
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Calculate the variance and standard deviation.

Solution All four steps needed to calculate the variance and standard deviation for grouped data are shown after Table 3.12.

Table 3.12

Daily Commuting Time (minutes)	f	m	mf	m^2f
0 to less than 10	4	5	20	100
10 to less than 20	9	15	135	2025
20 to less than 30	6	25	150	3750
30 to less than 40	4	35	140	4900
40 to less than 50	2	45	90	4050
	$N = 25$		$\Sigma mf = 535$	$\Sigma m^2f = 14{,}825$

Step 1. *Calculate the value of* Σmf.

To calculate the value of Σmf, first find the midpoint m of each class (see the third column in Table 3.12) and then multiply the corresponding class midpoints and class frequencies (see the fourth column). The value of Σmf is obtained by adding these products. Thus,

$$\Sigma mf = 535$$

Step 2. *Find the value of* Σm^2f.

To find the value of Σm^2f, square each m value and multiply this squared value of m by the corresponding frequency (see the fifth column in Table 3.12). The sum of these products (that is, the sum of the fifth column) gives Σm^2f. Hence,

$$\Sigma m^2f = 14{,}825$$

Step 3. *Calculate the variance.*

Because the data set includes all 25 employees of the company, it represents the population. Therefore, we use the formula for the population variance:

$$\sigma^2 = \frac{\Sigma m^2f - \frac{(\Sigma mf)^2}{N}}{N} = \frac{14{,}825 - \frac{(535)^2}{25}}{25} = \frac{3376}{25} = \mathbf{135.04}$$

Step 4. *Calculate the standard deviation.*

To obtain the standard deviation, take the (positive) square root of the variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{135.04} = \mathbf{11.62\ minutes}$$

Thus, the standard deviation of the daily commuting times for these employees is 11.62 minutes.

Note that the values of the variance and standard deviation calculated in Example 3–16 for grouped data are approximations. The exact values of the variance and standard deviation can be obtained only by using the ungrouped data on the daily commuting times of the 25 employees.

EXAMPLE 3–17

Calculating the sample variance and standard deviation for grouped data.

The following data, reproduced from Table 3.10 of Examle 3–15, give the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company.

Number of Orders	f
10–12	4
13–15	12
16–18	20
19–21	14

Calculate the variance and standard deviation.

Solution All the information required for the calculation of the variance and standard deviation appears in Table 3.13.

Table 3.13

Number of Orders	f	m	mf	m^2f
10–12	4	11	44	484
13–15	12	14	168	2352
16–18	20	17	340	5780
19–21	14	20	280	5600
	$n = 50$		$\Sigma mf = 832$	$\Sigma m^2f = 14{,}216$

Because the data set includes only 50 days, it represents a sample. Hence, we use the sample formulas to calculate the variance and standard deviation. By substituting the values into the formula for the sample variance, we obtain

$$s^2 = \frac{\Sigma m^2 f - \dfrac{(\Sigma mf)^2}{n}}{n - 1} = \frac{14{,}216 - \dfrac{(832)^2}{50}}{50 - 1} = \mathbf{7.5820}$$

Hence, the standard deviation is

$$s = \sqrt{s^2} = \sqrt{7.5820} = \mathbf{2.75\ orders}$$

Thus, the standard deviation of the number of orders received at the office of this mail-order company during the past 50 days is 2.75.

EXERCISES

CONCEPTS AND PROCEDURES

3.61 Are the values of the mean and standard deviation that are calculated using grouped data exact or approximate values of the mean and standard deviation, respectively? Explain.

3.62 Using the population formulas, calculate the mean, variance, and standard deviation for the following grouped data.

x	2–4	5–7	8–10	11–13	14–16
f	5	9	14	7	5

3.63 Using the sample formulas, find the mean, variance, and standard deviation for the grouped data displayed in the following table.

x	f
0 to less than 4	17
4 to less than 8	23
8 to less than 12	15
12 to less than 16	11
16 to less than 20	8
20 to less than 24	6

APPLICATIONS

3.64 The following table gives the frequency distribution of the amounts of telephone bills for October 2005 for a sample of 50 families.

Amount of Telephone Bill (dollars)	Number of Families
40 to less than 70	9
70 to less than 100	11
100 to less than 130	16
130 to less than 160	10
160 to less than 190	4

Calculate the mean, variance, and standard deviation.

3.65 The following table gives the frequency distribution of the number of hours spent per week playing video games by all 60 students of the eighth grade at a school.

Hours Per Week	Number of Students
0 to less than 5	7
5 to less than 10	12
10 to less than 15	15
15 to less than 20	13
20 to less than 25	8
25 to less than 30	5

Find the mean, variance, and standard deviation.

3.66 The following table gives the grouped data on the weights of all 100 babies born at a hospital in 2005.

Weight (pounds)	Number of Babies
3 to less than 5	5
5 to less than 7	30
7 to less than 9	40
9 to less than 11	20
11 to less than 13	5

Find the mean, variance, and standard deviation.

3.67 The following table gives the frequency distribution of the total miles driven during 2005 by 300 car owners.

Miles Driven in 2002 (in thousands)	Number of Car Owners
0 to less than 5	7
5 to less than 10	26
10 to less than 15	59
15 to less than 20	71
20 to less than 25	62
25 to less than 30	39
30 to less than 35	22
35 to less than 40	14

Find the mean, variance, and standard deviation. Give a brief interpretation of the values in the column labeled mf in your table of calculations. What does Σmf represent?

3.68 The following table gives information on the amounts (in dollars) of electric bills for August 2005 for a sample of 50 families.

Amount of Electric Bill (dollars)	Number of Families
0 to less than 20	5
20 to less than 40	16
40 to less than 60	11
60 to less than 80	10
80 to less than 100	8

Find the mean, variance, and standard deviation. Give a brief interpretation of the values in the column labeled mf in your table of calculations. What does Σmf represent?

3.69 For 50 airplanes that arrived late at an airport during a week, the time by which they were late was observed. In the following table, x denotes the time (in minutes) by which an airplane was late and f denotes the number of airplanes.

x	f
0 to less than 20	14
20 to less than 40	18
40 to less than 60	9
60 to less than 80	5
80 to less than 100	4

Find the mean, variance, and standard deviation.

3.70 The following table gives the frequency distribution of the number of errors committed by a college baseball team in all of the 45 games that it played during the 2005–2006 season.

Number of Errors	Number of Games
0	11
1	14
2	9
3	7
4	3
5	1

Find the mean, variance, and standard deviation. (*Hint:* The classes in this example are single-valued. These values of classes will be used as values of m in the formulas for the mean, variance, and standard deviation.)

1519T_c03 10/31/05 6:55 AM Page 101

EQA

3.71 During fall 2004, oil prices fluctuated a lot due to wars, political unrests, and storm damages in some oil-producing nations. The following data give the spot prices (in dollars) per barrel of crude oil for 15 business days from October 20 to November 9, 2004.

54.92	54.47	55.17	55.18	55.95	52.47	50.93	
51.74	50.14	49.63	50.89	48.83	49.62	49.09	47.38

a. Find the mean for these data.
b. Construct a frequency distribution table for these data using a class width of 2.00 and the lower boundary of the first class equal to 47.00.
c. Using the method of Section 3.3.1, find the mean of the grouped data of part b.
d. Compare your means from parts a and c. If the two means are not equal, then explain why they differ.

3.4 Use of Standard Deviation

By using the mean and standard deviation, we can find the proportion or percentage of the total observations that fall within a given interval about the mean. This section briefly discusses Chebyshev's theorem and the empirical rule, both of which demonstrate this use of the standard deviation.

3.4.1 Chebyshev's Theorem

Chebyshev's theorem gives a lower bound for the area under a curve between two points that are on opposite sides of the mean and at the same distance from the mean.

Definition

Chebyshev's Theorem For any number k greater than 1, at least $(1 - 1/k^2)$ of the data values lie within k standard deviations of the mean.

Figure 3.5 illustrates Chebyshev's theorem.

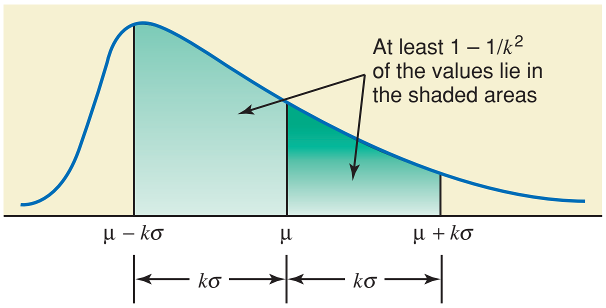


Figure 3.5 Chebyshev's theorem.

Thus, for example, if $k = 2$, then

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(2)^2} = 1 - \frac{1}{4} = 1 - .25 = .75 \text{ or } 75\%$$

Therefore, according to Chebyshev's theorem, at least .75 or 75% of the values of a data set lie within two standard deviations of the mean. This is shown in Figure 3.6 on the next page.

If $k = 3$, then,

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3)^2} = 1 - \frac{1}{9} = 1 - .11 = .89 \text{ or } 89\% \text{ approximately}$$

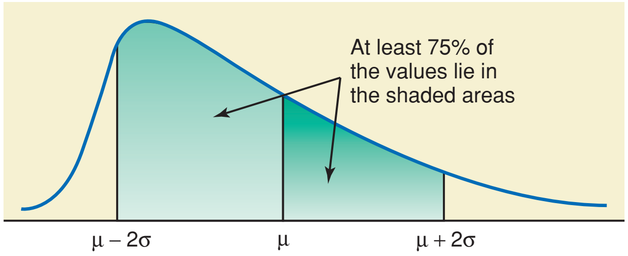


Figure 3.6 Percentage of values within two standard deviations of the mean for Chebyshev's theorem.

According to Chebyshev's theorem, at least .89 or 89% of the values fall within three standard deviations of the mean. This is shown in Figure 3.7.

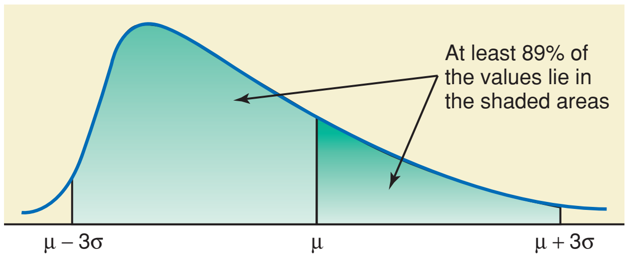


Figure 3.7 Percentage of values within three standard deviations of the mean for Chebyshev's theorem.

Although in Figures 3.5 through 3.7 we have used the population notation for the mean and standard deviation, the theorem applies to both sample and population data. Note that Chebyshev's theorem is applicable to a distribution of any shape. However, Chebyshev's theorem can be used only for $k > 1$. This is so because when $k = 1$, the value of $1 - 1/k^2$ is zero, and when $k < 1$, the value of $1 - 1/k^2$ is negative.

■ EXAMPLE 3–18

Applying Chebyshev's theorem.

The average systolic blood pressure for 4000 women who were screened for high blood pressure was found to be 187 with a standard deviation of 22. Using Chebyshev's theorem, find at least what percentage of women in this group have a systolic blood pressure between 143 and 231.

Solution Let μ and σ be the mean and the standard deviation, respectively, of the systolic blood pressures of these women. Then, from the given information,

$$\mu = 187 \quad \text{and} \quad \sigma = 22$$

To find the percentage of women whose systolic blood pressures are between 143 and 231, the first step is to determine k. As shown below, each of the two points, 143 and 231, is 44 units away from the mean.

$$|\leftarrow\ 143 - 187 = -44\ \rightarrow|\leftarrow\ 231 - 187 = 44\ \rightarrow|$$

$$143 \qquad\qquad \mu = 187 \qquad\qquad 231$$

The value of k is obtained by dividing the distance between the mean and each point by the standard deviation. Thus,

$$k = 44/22 = 2$$

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(2)^2} = 1 - \frac{1}{4} = 1 - .25 = .75 \text{ or } \mathbf{75\%}$$

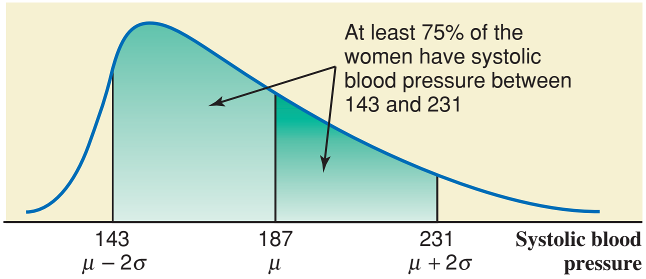


Figure 3.8 Percentage of women with systolic blood pressure between 143 and 231.

Hence, according to Chebyshev's theorem, at least 75% of the women have systolic blood pressure between 143 and 231. This percentage is shown in Figure 3.8. ■

3.4.2 Empirical Rule

Whereas Chebyshev's theorem is applicable to any kind of distribution, the **empirical rule** applies only to a specific type of distribution called a *bell-shaped distribution*, as shown in Figure 3.9. More will be said about such a distribution in Chapter 6, where it is called a *normal curve.* In this section, only the following three rules for the curve are given.

Empirical Rule For a bell-shaped distribution, approximately

1. 68% of the observations lie within one standard deviation of the mean
2. 95% of the observations lie within two standard deviations of the mean
3. 99.7% of the observations lie within three standard deviations of the mean

Figure 3.9 illustrates the empirical rule. Again, the empirical rule applies to population data as well as to sample data.

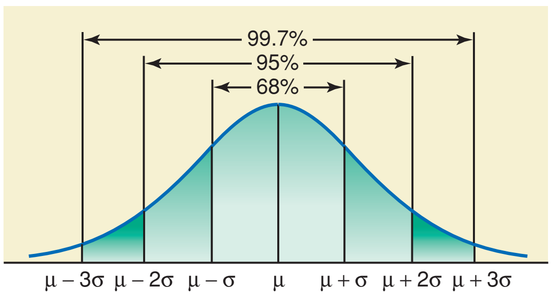


Figure 3.9 Illustration of the empirical rule.

■ EXAMPLE 3–19

Applying the empirical rule.

The age distribution of a sample of 5000 persons is bell-shaped with a mean of 40 years and a standard deviation of 12 years. Determine the approximate percentage of people who are 16 to 64 years old.

Solution We use the empirical rule to find the required percentage because the distribution of ages follows a bell-shaped curve. From the given information, for this distribution,

$$\bar{x} = 40 \text{ years} \quad \text{and} \quad s = 12 \text{ years}$$

Each of the two points, 16 and 64, is 24 units away from the mean. Dividing 24 by 12, we convert the distance between each of the two points and the mean in terms of standard deviations. Thus, the distance between 16 and 40 and between 40 and 64 is each equal to 2s.

CASE STUDY 3–3

HERE COMES THE SD

When your servant first became a *Fortune* writer several decades ago, it was hard doctrine that "several" meant three to eight, also that writers must not refer to "gross national product" without pausing to define this arcane term. GNP was in fact a relatively new concept at the time, having been introduced to the country only several years previously—in Roosevelt's 1944 budget message—so the presumption that readers had to be told repeatedly it was the "value of all goods and services produced by the economy" seemed entirely reasonable to this young writer, who personally had to look up the definition every time.

Numeracy lurches on. Nowadays the big question for editors is whether an average college-educated bloke needs a handhold when confronted with the term "standard deviation." The SD is suddenly onstage because the Securities and Exchange Commission is wondering aloud whether investment companies should be required to tell investors the standard deviation of their mutual funds' total returns over various past periods. Barry Barbash, SEC director of investment management, favors the requirement but confessed to the Washington *Post* that he worries about investors who will think a standard deviation is the dividing line on a highway or something.

The view around here is that the SEC is performing a noble service, but only partly because the requirement would enhance folks' insights into mutual funds. The commission's underlying idea is to give investors a better and more objective measure than is now available of the risk associated with different kinds of portfolios. The SD is a measure of variability, and funds with unusually variable returns—sometimes very high, sometimes very low—are presumed to be more risky.

What one really likes about the proposal, however, is the prospect that it will incentivize millions of greedy Americans to learn a little elementary statistics. One already has a list of issues that could be discussed much more thrillingly if only your average liberal arts graduate had a glimmer about the SD and the normal curve. The bell-shaped normal curve, or rather, the area underneath the curve, shows you how Providence arranged for things to be distributed in our world—with people's heights, or incomes, or IQs, or investment returns bunched around middling outcomes, and fewer and fewer cases as you move down and out toward the extremes. A line down the center of the curve represents the mean outcome, and deviations from the mean are measured by the SD.

An amazing property of the SD is that exactly 68.26% of all normally distributed data are within one SD of the mean. We once asked a professor of statistics a question that seemed to us quite profound, to wit, why that particular figure? The Prof answered dismissively that God had decided on 68.26% for exactly the same reason He had landed on 3.14 as the ratio between circumferences and diameters—because He just felt like it. The Almighty has also proclaimed that 95.44% of all data are within two SDs of the mean, and 99.73% within three SDs. When you know the mean and SD of some outcome, you can instantly establish the percentage probability of its occurrence. White men's heights in the U.S. average 69.2 inches, with an SD of 2.8 inches (according to the National Center for Health Statistics), which means that a 6-foot-5 chap is in the 99th percentile. In 1994, scores on the verbal portion of the Scholastic Assessment Test had a mean of 423 and an SD of 113, so if you scored 649—two SDs above the mean—you were in the 95th percentile.

As the SEC is heavily hinting, average outcomes are interesting but for many purposes inadequate; one also yearns to know the variability around that average. From 1926 through 1994, the S&P 500 had an average annual return of just about 10%. The SD accompanying that figure was just about 20%. Since returns will be within 1 SD some 68% of the time, they will be more than 1 SD from the mean 32% of the time. And since half these swings will be on the downside, we expect fund owners to lose more than 10% of their money about one year out of six and to lose more than 30% (two SDs below the mean) about one year out of 20. If your time horizon is short and you can't take losses like that, you arguably don't belong in stocks. If you think SDs are highway dividers, you arguably don't belong in cars.

Source: Daniel Seligman, "Here comes the SD," *Fortune*, May 15, 1995.

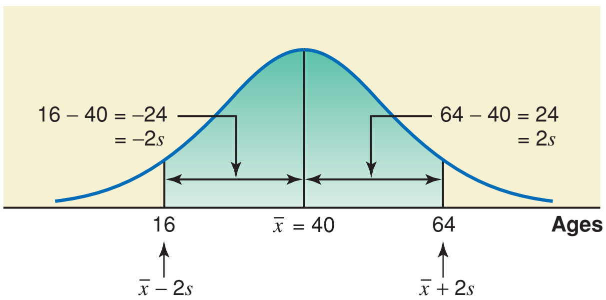


Figure 3.10 Percentage of people who are 16 to 64 years old.

Consequently, as shown in Figure 3.10, the area from 16 to 64 is the area from $\bar{x} - 2s$ to $\bar{x} + 2s$.

Because the area within two standard deviations of the mean is approximately 95% for a bell-shaped curve, approximately **95%** of the people in the sample are 16 to 64 years old. ■

EXERCISES

CONCEPTS AND PROCEDURES

3.72 Briefly explain Chebyshev's theorem and its applications.

3.73 Briefly explain the empirical rule. To what kind of distribution is it applied?

3.74 A sample of 2000 observations has a mean of 74 and a standard deviation of 12. Using Chebyshev's theorem, find at least what percentage of the observations fall in the intervals $\bar{x} \pm 2s$, $\bar{x} \pm 2.5s$, and $\bar{x} \pm 3s$. Note that here $\bar{x} \pm 2s$ represents the interval $\bar{x} - 2s$ to $\bar{x} + 2s$, and so on.

3.75 A large population has a mean of 230 and a standard deviation of 41. Using Chebyshev's theorem, find at least what percentage of the observations fall in the intervals $\mu \pm 2\sigma$, $\mu \pm 2.5\sigma$, and $\mu \pm 3\sigma$.

3.76 A large population has a mean of 310 and a standard deviation of 37. Using the empirical rule, find what percentage of the observations fall in the intervals $\mu \pm 1\sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$.

3.77 A sample of 3000 observations has a mean of 82 and a standard deviation of 16. Using the empirical rule, find what percentage of the observations fall in the intervals $\bar{x} \pm 1s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$.

APPLICATIONS

3.78 The mean time taken by all participants to run a road race was found to be 220 minutes with a standard deviation of 20 minutes. Using Chebyshev's theorem, find the percentage of runners who ran this road race in

a. 180 to 260 minutes **b.** 160 to 280 minutes **c.** 170 to 270 minutes

3.79 The 2005 gross sales of all firms in a large city have a mean of \$2.3 million and a standard deviation of \$.6 million. Using Chebyshev's theorem, find at least what percentage of firms in this city had 2005 gross sales of

a. \$1.1 to \$3.5 million **b.** \$.8 to \$3.8 million **c.** \$.5 to \$4.1 million

3.80 Suppose the average credit card debt for households currently is \$9500 with a standard deviation of \$2600.

a. Using Chebyshev's theorem, find at least what percentage of current credit card debts for all households are between

i. \$4300 and \$14,700 **ii.** \$3000 and \$16,000

***b.** Using Chebyshev's theorem, find the interval that contains credit card debts of at least 89% of all households.

3.81 The mean monthly mortgage paid by all home owners in a city is \$2365 with a standard deviation of \$340.

a. Using Chebyshev's theorem, find at least what percentage of all home owners in the city pay a monthly mortgage of

i. \$1685 to \$3045 **ii.** \$1345 to \$3385

***b.** Using Chebyshev's theorem, find the interval that contains the monthly mortgage payments of at least 84% of all home owners.

3.82 The mean life of a certain brand of auto batteries is 44 months with a standard deviation of 3 months. Assume that the lives of all auto batteries of this brand have a bell-shaped distribution. Using the empirical rule, find the percentage of auto batteries of this brand that have a life of

a. 41 to 47 months **b.** 38 to 50 months **c.** 35 to 53 months

3.83 According to Hewitt and Associates (a consulting firm in Lincolnshire, Illinois), the employee share of health insurance premiums at large U.S. companies was expected to be \$1481, on average, in 2005. Suppose the current payments by all such employees toward health insurance premiums have a bell-shaped distribution with a mean of \$1481 per year and a standard deviation of \$355. Using the empirical rule, find the percentage of employees whose annual payments toward such premiums are between

a. \$771 and \$2191 **b.** \$1126 and \$1836 **c.** \$416 and \$2546

3.84 The prices of all college textbooks follow a bell-shaped distribution with a mean of \$105 and a standard deviation of \$20.

a. Using the empirical rule, find the percentage of all college textbooks with their prices between

i. \$85 and \$125 **ii.** \$65 and \$145

***b.** Using the empirical rule, find the interval that contains the prices of 99.7% of college textbooks.

3.85 Suppose that on a certain section of I-95, with a posted speed limit of 65 miles per hour, the speeds of all vehicles have a bell-shaped distribution with a mean of 72 mph and a standard deviation of 3 mph.

a. Using the empirical rule, find the percentage of vehicles with the following speeds on this section of I-95.

i. 63 to 81 mph **ii.** 69 to 75 mph

***b.** Using the empirical rule, find the interval that contains the speeds of 95% of vehicles traveling on this section of I-95.

3.5 Measures of Position

A **measure of position** determines the position of a single value in relation to other values in a sample or a population data set. There are many measures of position; however, only quartiles, percentiles, and percentile rank are discussed in this section.

3.5.1 Quartiles and Interquartile Range

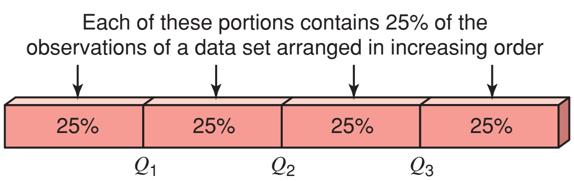
Quartiles are the summary measures that divide a ranked data set into four equal parts. Three measures will divide any data set into four equal parts. These three measures are the **first quartile** (denoted by Q_1), the **second quartile** (denoted by Q_2), and the **third quartile** (denoted by Q_3). The data should be ranked in increasing order before the quartiles are determined. The quartiles are defined as follows.

Definition

Quartiles *Quartiles* are three summary measures that divide a ranked data set into four equal parts. The second quartile is the same as the median of a data set. The first quartile is the value of the middle term among the observations that are less than the median, and the third quartile is the value of the middle term among the observations that are greater than the median.

Figure 3.11 describes the positions of the three quartiles.

Figure 3.11 Quartiles.



Approximately 25% of the values in a ranked data set are less than Q_1 and about 75% are greater than Q_1. The second quartile, Q_2, divides a ranked data set into two equal parts; hence, the second quartile and the median are the same. Approximately 75% of the data values are less than Q_3 and about 25% are greater than Q_3.

The difference between the third quartile and the first quartile for a data set is called the **interquartile range (IQR)**.

Calculating Interquartile Range The difference between the third and the first quartiles gives the *interquartile range*; that is,

$$\text{IQR} = \text{Interquartile range} = Q_3 - Q_1$$

Examples 3–20 and 3–21 show the calculation of the quartiles and the interquartile range.

EXAMPLE 3–20

Finding quartiles and the interquartile range.

Refer to Table 3.3 in Example 3–5 that lists the number of car thefts during 2003 in 12 cities. That table is reproduced below.

City	Number of Car Thefts
Phoenix-Mesa, Arizona	40,769
Washington, D.C.	33,956
Miami, Florida	21,088
Atlanta, Georgia	29,920
Chicago, Illinois	42,082
Kansas City, Kansas	11,669
Baltimore, Maryland	13,435
Detroit, Michigan	40,197
St. Louis, Missouri	18,215
Las Vegas, Nevada	18,103
Newark, New Jersey	14,413
Dallas, Texas	26,343

Source: National Insurance Crime Bureau.

(a) Find the values of the three quartiles. Where does the number of car thefts of 40,197 fall in relation to these quartiles?

(b) Find the interquartile range.

Solution

Finding quartiles for an even number of data values.

(a) First we rank the given data in increasing order. Then we calculate the three quartiles as follows:

Values less than the median						Values greater than the median					
11,669	13,435	14,413	18,103	18,215	21,088	26,343	29,920	33,956	40,197	40,769	42,082

$$Q_1 = \frac{14{,}413 + 18{,}103}{2} = \mathbf{16{,}258} \qquad Q_2 = \frac{21{,}088 + 26{,}343}{2} = \mathbf{23{,}715.50} \qquad Q_3 = \frac{33{,}956 + 40{,}197}{2} = \mathbf{37{,}076.50}$$

Also the median

The value of Q_2, which is also the median, is given by the value of the middle term in the ranked data set. For the data of this example, this value is the average of the sixth and seventh terms. Consequently, Q_2 is 23,715.50 car thefts. The value of Q_1 is given by the value of the middle term of the six values that fall below the median (or Q_2). Thus, it is obtained by taking the average of the third and fourth terms. So, Q_1 is 16,258 car thefts. The value of Q_3 is given by the value of the middle term of the six values that fall above the median. For the data of this example, Q_3 is obtained by taking the average of the ninth and tenth terms, and it is 37,076.50 car thefts.

The value of $Q_1 = 16{,}258$ indicates that the number of car thefts in (approximately) 25% of these cities were less than 16,258 in 2003 and those in (approximately) 75% of the cities were greater than this value. Similarly, we can state that the car thefts in about half of these cities were less than 23,715.50 (which is Q_2) in 2003 and those in the other half were greater than this value. The value of $Q_3 = 37{,}076.50$ indicates that the car thefts in (approximately) 75% of the cities in this sample were less than 37,076.50 in 2003 and those in (approximately) 25% of the cities were greater than this value.

By looking at the position of 40,197, we can state that this value lies in the **top 25%** of the car thefts.

Finding the interquartile range.

(b) The interquartile range is given by the difference between the values of the third and the first quartiles. Thus,

$$\text{IQR} = \text{Interquartile range} = Q_3 - Q_1 = 37{,}076.50 - 16{,}258 = \mathbf{20{,}818.50 \text{ car thefts}}$$

EXAMPLE 3–21

Finding quartiles and the interquartile range.

The following are the ages of nine employees of an insurance company:

47 28 39 51 33 37 59 24 33

(a) Find the values of the three quartiles. Where does the age of 28 fall in relation to the ages of these employees?

(b) Find the interquartile range.

Solution

Finding quartiles for an odd number of data values.

(a) First we rank the given data in increasing order. Then we calculate the three quartiles as follows:

$$\underbrace{24 \quad 28 \quad 33 \quad 33}_{\text{Values less than the median}} \quad 37 \quad \underbrace{39 \quad 47 \quad 51 \quad 59}_{\text{Values greater than the median}}$$

$$Q_1 = \frac{28+33}{2} = 30.5 \qquad Q_2 = 37 \qquad Q_3 = \frac{47+51}{2} = 49$$

Also the median

Thus the values of the three quartiles are

$$Q_1 = \mathbf{30.5 \text{ years}}, \quad Q_2 = \mathbf{37 \text{ years}}, \quad \text{and} \quad Q_3 = \mathbf{49 \text{ years}}$$

The age of 28 falls in the **lowest 25%** of the ages.

Finding the interquartile range.

(b) The interquartile range is

$$\text{IQR} = \text{Interquartile range} = Q_3 - Q_1 = 49 - 30.5 = \mathbf{18.5 \text{ years}}$$

3.5.2 Percentiles and Percentile Rank

Percentiles are the summary measures that divide a ranked data set into 100 equal parts. Each (ranked) data set has 99 percentiles that divide it into 100 equal parts. The data should be ranked in increasing order to compute percentiles. The kth percentile is denoted by P_k, where k is an integer in the range 1 to 99. For instance, the 25th percentile is denoted by P_{25}. Figure 3.12 shows the positions of the 99 percentiles.

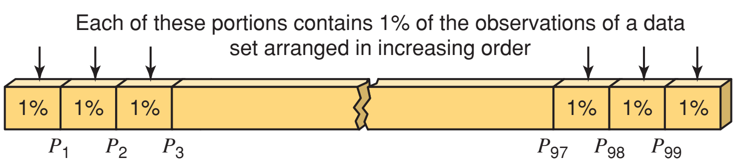


Figure 3.12 Percentiles.

Thus, the kth percentile, P_k, can be defined as a value in a data set such that about k% of the measurements are smaller than the value of P_k and about $(100 - k)$% of the measurements are greater than the value of P_k.

The approximate value of the kth percentile is determined as explained next.

Calculating Percentiles The (approximate) value of the kth *percentile*, denoted by P_k, is

$$P_k = \text{Value of the} \left(\frac{kn}{100}\right)\text{th term in a ranked data set}$$

where k denotes the number of the percentile and n represents the sample size.

Example 3–22 describes the procedure to calculate the percentiles.

■ EXAMPLE 3–22

Finding the percentile for a data set.

Refer to the data on 2003 car thefts in 12 cities given in Example 3–20. Find the value of the 42nd percentile. Give a brief interpretation of the 42nd percentile.

Solution From Example 3–20, the data arranged in increasing order are as follows:

11,669 13,435 14,413 18,103 18,215 21,088 26,343 29,920 33,956 40,197 40,769 42,082

The position of the 42nd percentile is

$$\frac{kn}{100} = \frac{42(12)}{100} = 5.04\text{th term}$$

The value of the 5.04th term can be approximated by the value of the fifth term in the ranked data. Therefore,

$$P_{42} = \text{42nd percentile} = \mathbf{18{,}215\ car\ thefts}$$

Thus, approximately 42% of these 12 cities had 18,215 or fewer car thefts in 2003 and 58% had higher than 18,215 car thefts. ■

We can also calculate the **percentile rank** for a particular value x_i of a data set by using the formula given below. The percentile rank of x_i gives the percentage of values in the data set that are less than x_i.

Finding Percentile Rank of a Value

$$\text{Percentile rank of } x_i = \frac{\text{Number of values less than } x_i}{\text{Total number of values in the data set}} \times 100$$

Example 3–23 shows how the percentile rank is calculated for a data value.

■ EXAMPLE 3–23

Finding the percentile rank for a data value.

Refer to the data on 2003 car thefts in 12 cities given in Example 3–20. Find the percentile rank for 29,920 car thefts. Give a brief interpretation of this percentile rank.

Solution From Example 3–20, the data arranged in increasing order are as follows:

11,669 13,435 14,413 18,103 18,215 21,088 26,343 29,920 33,956 40,197 40,769 42,082

In this data set, 7 of the 12 values are less than 29,920. Hence,

$$\text{Percentile rank of } 29{,}920 = \frac{7}{12} \times 100 = \mathbf{58.33}\%$$

Rounding this answer to the nearest integral value, we can state that about 58% of the cities in these 12 cities had less than 29,920 car thefts in 2003. Hence, about 42% of the 12 cities had 29,920 or higher car thefts in 2003. ■

EXERCISES

CONCEPTS AND PROCEDURES

3.86 Briefly describe how the three quartiles are calculated for a data set. Illustrate by calculating the three quartiles for two examples, the first with an odd number of observations and the second with an even number of observations.

3.87 Explain how the interquartile range is calculated. Give one example.

3.88 Briefly describe how the percentiles are calculated for a data set.

3.89 Explain the concept of the percentile rank for an observation of a data set.

APPLICATIONS

3.90 The following data give the weights (in pounds) lost by 15 members of a health club at the end of two months after joining the club.

5	10	8	7	25	12	5	14
11	10	21	9	8	11	18	

a. Compute the values of the three quartiles and the interquartile range.
b. Calculate the (approximate) value of the 82nd percentile.
c. Find the percentile rank of 10.

3.91 The following data give the speeds of 13 cars, measured by radar, traveling on I-84.

73	75	69	68	78	69	74
76	72	79	68	77	71	

a. Find the values of the three quartiles and the interquartile range.
b. Calculate the (approximate) value of the 35th percentile.
c. Compute the percentile rank of 71.

3.92 The following data give the numbers of computer keyboards assembled at the Twentieth Century Electronics Company for a sample of 25 days.

45	52	48	41	56	46	44	42	48	53
51	53	51	48	46	43	52	50	54	47
44	47	50	49	52					

a. Calculate the values of the three quartiles and the interquartile range.
b. Determine the (approximate) value of the 53rd percentile.
c. Find the percentile rank of 50.

3.93 The following data give the number of runners left on bases by each of the 30 Major League Baseball teams in the games played on August 12, 2004.

6	6	6	7	6	10	6	3	6	8	10	7	18	11	6
9	4	8	9	5	5	4	8	8	8	5	5	5	13	8

a. Calculate the values of the three quartiles and the interquartile range.
b. Find the (approximate) value of the 63rd percentile.
c. Find the percentile rank of 10.

3.94 Refer to Exercise 3.22. The following data give the number of students suspended for bringing weapons to schools in the Tri-City School District for each of the past 12 weeks.

15	9	12	11	7	6	9	10	14	3	6	5

a. Determine the values of the three quartiles and the interquartile range. Where does the value of 10 fall in relation to these quartiles?
b. Calculate the (approximate) value of the 55th percentile.
c. Find the percentile rank of 7.

3.95 Nixon Corporation manufactures computer monitors. The following data give the numbers of computer monitors produced at the company for a sample of 30 days.

24	32	27	23	33	33	29	25	23	36
26	26	31	20	27	33	27	23	28	29
31	35	34	22	37	28	23	35	31	43

a. Calculate the values of the three quartiles and the interquartile range. Where does the value of 31 lie in relation to these quartiles?
b. Find the (approximate) value of the 65th percentile. Give a brief interpretation of this percentile.
c. For what percentage of the days was the number of computer monitors produced 32 or higher? Answer by finding the percentile rank of 32.

3.96 The following data give the numbers of new cars sold at a dealership during a 20-day period.

8	5	12	3	9	10	6	12	8	8
4	16	10	11	7	7	3	5	9	11

a. Calculate the values of the three quartiles and the interquartile range. Where does the value of 4 lie in relation to these quartiles?
b. Find the (approximate) value of the 25th percentile. Give a brief interpretation of this percentile.
c. Find the percentile rank of 10. Give a brief interpretation of this percentile rank.

3.97 According to the National Association of Realtors, the median home price in San Diego for the second quarter of 2003 was $559,700 (*USA TODAY*, August 27, 2004). Suppose the following data give the sale prices (in thousands of dollars) of a random sample of 20 recently sold homes in San Diego.

605	789	550	881	499	675	700	543	910	808
1016	929	544	397	649	752	698	710	495	509

a. Calculate the values of the three quartiles and the interquartile range. Where does the value of 649 fall in relation to these quartiles?
b. Calculate the (approximate) value of the 77th percentile. Give a brief interpretation of this percentile.
c. Find the percentile rank of 700. Give a brief interpretation of this percentile rank.

3.6 Box-and-Whisker Plot

A **box-and-whisker plot** gives a graphic presentation of data using five measures: the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences. (The inner fences are explained in Example 3–24 below.) A box-and-whisker plot can help us visualize the center, the spread, and the skewness of a data set. It also helps detect outliers. We can compare different distributions by making box-and-whisker plots for each of them.

Definition

Box-and-Whisker Plot A plot that shows the center, spread, and skewness of a data set. It is constructed by drawing a box and two whiskers that use the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences.

Example 3–24 explains all the steps needed to make a box-and-whisker plot.

■ EXAMPLE 3–24

The following data are the incomes (in thousands of dollars) for a sample of 12 households.

Constructing a box-and-whisker plot.

35	29	44	72	34	64	41	50	54	104	39	58

Construct a box-and-whisker plot for these data.

Solution The following five steps are performed to construct a box-and-whisker plot.

Step 1. First, rank the data in increasing order and calculate the values of the median, the first quartile, the third quartile, and the interquartile range. The ranked data are

29 34 35 39 41 44 50 54 58 64 72 104

For these data,

$$\text{Median} = (44 + 50)/2 = 47$$
$$Q_1 = (35 + 39)/2 = 37$$
$$Q_3 = (58 + 64)/2 = 61$$
$$\text{IQR} = Q_3 - Q_1 = 61 - 37 = 24$$

Step 2. Find the points that are $1.5 \times \text{IQR}$ below Q_1 and $1.5 \times \text{IQR}$ above Q_3. These two points are called the **lower** and the **upper inner fences**, respectively.

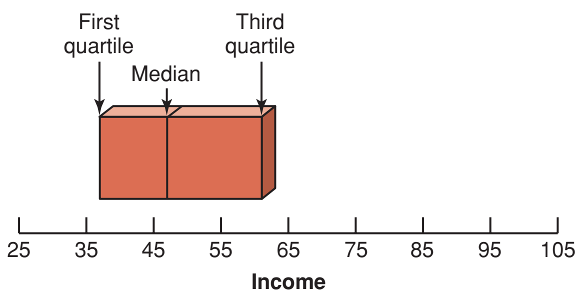
$$1.5 \times \text{IQR} = 1.5 \times 24 = 36$$
$$\text{Lower inner fence} = Q_1 - 36 = 37 - 36 = 1$$
$$\text{Upper inner fence} = Q_3 + 36 = 61 + 36 = 97$$

Step 3. Determine the smallest and the largest values in the given data set within the two inner fences. These two values for our example are as follows:

$$\text{Smallest value within the two inner fences} = 29$$
$$\text{Largest value within the two inner fences} = 72$$

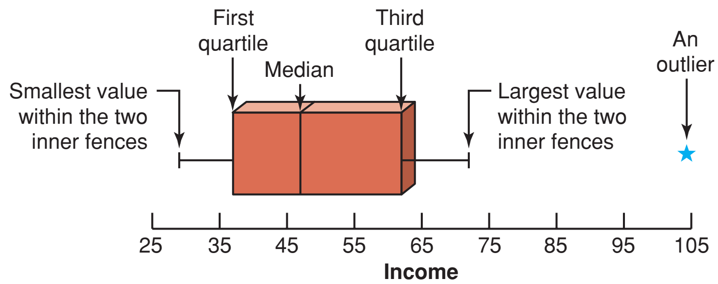
Step 4. Draw a horizontal line and mark the income levels on it such that all the values in the given data set are covered. Above the horizontal line, draw a box with its left side at the position of the first quartile and the right side at the position of the third quartile. Inside the box, draw a vertical line at the position of the median. The result of this step is shown in Figure 3.13.

Figure 3.13



Step 5. By drawing two lines, join the points of the smallest and the largest values within the two inner fences to the box. These values are 29 and 72 in this example as listed in Step 3. The two lines that join the box to these two values are called **whiskers**. A value that falls outside the two inner fences is shown by marking an asterisk and is called an outlier. This completes the box-and-whisker plot, as shown in Figure 3.14.

Figure 3.14



In Figure 3.14, about 50% of the data values fall within the box, about 25% of the values fall on the left side of the box, and about 25% fall on the right side of the box. Also, 50% of the values fall on the left side of the median and 50% lie on the right side of the median. The data of this example are skewed to the right because the lower 50% of the values are spread over a smaller range than the upper 50% of the values. ■

The observations that fall outside the two inner fences are called outliers. These outliers can be classified into two kinds of outliers—mild and extreme outliers. To do so, we define two outer fences—a **lower outer fence** at 3.0 × IQR below the first quartile and an **upper outer fence** at 3.0 × IQR above the third quartile. If an observation is outside either of the two inner fences but within either of the two outer fences, it is called a *mild outlier.* An observation that is outside either of the two outer fences is called an *extreme outlier.* For the previous example, the outer fences are at −35 and 133. Because 104 is outside the upper inner fence but inside the upper outer fence, it is a mild outlier.

For a symmetric data set, the line representing the median will be in the middle of the box and the spread of the values will be over almost the same range on both sides of the box.

EXERCISES

CONCEPTS AND PROCEDURES

3.98 Briefly explain what summary measures are used to construct a box-and-whisker plot.

3.99 Prepare a box-and-whisker plot for the following data:

36	43	28	52	41	59	47	61
24	55	63	73	32	25	35	49
31	22	61	42	58	65	98	34

Does this data set contain any outliers?

3.100 Prepare a box-and-whisker plot for the following data:

11	8	26	31	62	19	7	3	14	75
33	30	42	15	18	23	29	13	16	6

Does this data set contain any outliers?

APPLICATIONS

3.101 The following data give the time (in minutes) that each of 20 students selected from a university waited in line at their bookstore to pay for their textbooks in the beginning of the Spring 2006 semester.

15	8	23	21	5	17	31	22	34	6
5	10	14	17	16	25	30	3	31	19

Prepare a box-and-whisker plot. Comment on the skewness of these data.

3.102 Refer to Exercise 3.97. The following data give the sale prices (in thousands of dollars) of a random sample of 20 recently sold homes in San Diego.

605	789	550	881	499	675	700	543	910	808
1016	929	544	397	649	752	698	710	495	509

Prepare a box-and-whisker plot. Are the data skewed in any direction?

3.103 The following data give the crude oil reserves (in billions of barrels) of Saudi Arabia, Iraq, Kuwait, Iran, United Arab Emirates, Venezuela, Russia, Libya, Nigeria, China, Mexico, and the United States (*USA TODAY*, June 7, 2004). The reserves for these countries are listed in that order.

261.7	112.0	97.7	94.4	80.3	64.0
51.2	29.8	27.0	26.8	25.0	22.5

Prepare a box-and-whisker plot. Are the data symmetric or skewed?

3.104 The following data give the numbers of computer keyboards assembled at the Twentieth Century Electronics Company for a sample of 25 days.

45	52	48	41	56	46	44	42	48	53
51	53	51	48	46	43	52	50	54	47
44	47	50	49	52					

Prepare a box-and-whisker plot. Comment on the skewness of these data.

3.105 Refer to Exercise 3.93. The following data give the number of runners left on bases by each of the 30 Major League Baseball teams in the games played on August 12, 2004.

6	6	6	7	6	10	6	3	6	8	10	7	18	11	6
9	4	8	9	5	5	4	8	8	8	5	5	5	13	8

Prepare a box-and-whisker plot. Are the data symmetric or skewed?

3.106 Refer to Exercise 3.22. The following data give the number of students suspended for bringing weapons to schools in the Tri-City School District for each of the past 12 weeks.

15	9	12	11	7	6	9	10	14	3	6	5

Make a box-and-whisker plot. Comment on the skewness of these data.

3.107 Nixon Corporation manufactures computer monitors. The following are the numbers of computer monitors produced at the company for a sample of 30 days:

24	32	27	23	33	33	29	25	23	28
21	26	31	20	27	33	27	23	28	29
31	35	34	22	26	28	23	35	31	27

Prepare a box-and-whisker plot. Comment on the skewness of these data.

3.108 The following data give the numbers of new cars sold at a dealership during a 20-day period.

8	5	12	3	9	10	6	12	8	8
4	16	10	11	7	7	3	5	9	11

Make a box-and-whisker plot. Comment on the skewness of these data.

USES AND MISUSES... BEATING THE CURVE

Instructors often grade exams on a curve. The *curve* is a very loose way of saying that a set of exam scores is compared to a bell-shaped curve and grades are assigned by the relationship of any particular score to quantitative statistical measures, such as the mean, standard deviation, quartiles, or quintiles. Knowing a little bit about statistics can help your instructor make an honest assessment.

Quite often, assumptions about the curve itself are flawed. A set of exam scores, when classified and plotted as a histogram, may not resemble a symmetric distribution at all, and in many cases is actually skewed to the left or right. If the distribution of scores is skewed to the left, most of the scores are clustered around the high scores; if the distribution of scores is skewed to the right, most of the scores are clustered around the low scores. Every college or university has several famously difficult classes in which the average exam score is very low: 25 points out of 100, for example. You might hear that a fellow student "beat the curve" on an exam. He or she is the one who scored a 95 when most of the class scored somewhere near 20. This is a way of saying that the student's score is an outlier.

There is nothing wrong with comparing a set of exam scores to a bell curve. A problem can arise when there is an attempt to assign a grade, ultimately a measure of student performance, based on the characteristics of the distribution of scores. If the distribution of scores represents a symmetric curve, the teacher may choose to give the mean (and median, and mode, in this case) a suitably "average" grade of B−. Those students with scores within one standard deviation above the mean get Bs; those students with scores between one and two standard deviations above the mean get As. Similarly, Cs and Ds are given out for scores appropriately below the mean. Students with exceptional scores get A+ while some students fail. Another strategy is to divide the scores into quintiles: the students in the top quintile receive As, those in the second quintile receive Bs, and so on (note that quintiles divide an arranged data set into five equal parts). This technique, called quantitative partitioning of the data, seems to be an objective and reasonable method for assigning grades, but when you as a student are affected by the partitioning technique, issues of fairness arise. For example, is it right that only one-fifth of the class can receive a particular grade?

Glossary

Bimodal distribution A distribution that has two modes.

Box-and-whisker plot A plot that shows the center, spread, and skewness of a data set with a box and two whiskers using the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences.

Chebyshev's theorem For any number k greater than 1, at least $(1 - 1/k^2)$ of the values for any distribution lie within k standard deviations of the mean.

Coefficient of variation A measure of relative variability that expresses standard deviation as a percentage of the mean.

Empirical rule For a specific bell-shaped distribution, about 68% of the observations fall in the interval $(\mu - \sigma)$ to $(\mu + \sigma)$, about 95% fall in the interval $(\mu - 2\sigma)$ to $(\mu + 2\sigma)$, and about 99.7% fall in the interval $(\mu - 3\sigma)$ to $(\mu + 3\sigma)$.

First quartile The value in a ranked data set such that about 25% of the measurements are smaller than this value and about 75% are larger. It is the median of the values that are smaller than the median of the whole data set.

Geometric mean Calculated by taking the nth root of the product of all values in a data set.

Interquartile range (IQR) The difference between the third and the first quartiles.

Lower inner fence The value in a data set that is 1.5 × IQR below the first quartile.

Lower outer fence The value in a data set that is 3.0 × IQR below the first quartile.

Mean A measure of central tendency calculated by dividing the sum of all values by the number of values in the data set.

Measures of central tendency Measures that describe the center of a distribution. The mean, median, and mode are three of the measures of central tendency.

Measures of dispersion Measures that give the spread of a distribution. The range, variance, and standard deviation are three such measures.

Measures of position Measures that determine the position of a single value in relation to other values in a data set. Quartiles, percentiles, and percentile rank are examples of measures of position.

Median The value of the middle term in a ranked data set. The median divides a ranked data set into two equal parts.

Mode The value (or values) that occurs with highest frequency in a data set.

Multimodal distribution A distribution that has more than two modes.

Parameter A summary measure calculated for population data.

Percentile rank The percentile rank of a value gives the percentage of values in the data set that are smaller than this value.

Percentiles Ninety-nine values that divide a ranked data set into 100 equal parts.

Quartiles Three summary measures that divide a ranked data set into four equal parts.

Range A measure of spread obtained by taking the difference between the largest and the smallest values in a data set.

Second quartile Middle or second of the three quartiles that divide a ranked data set into four equal parts. About 50% of the values in the data set are smaller and about 50% are larger than the second quartile. The second quartile is the same as the median.

Standard deviation A measure of spread that is given by the positive square root of the variance.

Statistic A summary measure calculated for sample data.

Third quartile Third of the three quartiles that divide a ranked data set into four equal parts. About 75% of the values in a data set are smaller than the value of the third quartile and about 25% are larger. It is the median of the values that are greater than the median of the whole data set.

Trimmed mean The k% trimmed mean is obtained by dropping k% of the smallest values and k% of the largest values from the given data and then calculating the mean of the remaining $(100 - 2k)$% of the values.

Unimodal distribution A distribution that has only one mode.

Upper inner fence The value in a data set that is 1.5 × IQR above the third quartile.

Upper outer fence The value in a data set that is 3.0 × IQR above the third quartile.

Variance A measure of spread.

Weighted mean Mean of a data set whose values are assigned different weights before the mean is calculated.

Supplementary Exercises

3.109 Each year the faculty at Metro Business College chooses 10 members from the current graduating class that they feel are most likely to succeed. The data below give the current annual incomes (in thousands of dollars) of the 10 members of the class of 2005 who were voted most likely to succeed.

59	68	44	68	57	104	56	44	47	40

a. Calculate the mean and median.
b. Does this data set contain any outlier(s)? If yes, drop the outlier(s) and recalculate the mean and median. Which of these measures changes by a greater amount when you drop the outlier(s)?
c. Is the mean or the median a better summary measure for these data? Explain.

3.110 The following data give the weights (in pounds) of the nine running backs selected for *PARADE* magazine's 42nd annual All-America High School Football Team (*PARADE*, January 23, 2005). Note that because this All-America team included only nine running backs, it can be considered the population of running backs for this team.

225 225 210 234 218 188 190 195 185

a. Calculate the mean and the median. Do these data have a mode? Why or why not? Explain.
b. Find the range, variance, and standard deviation.

3.111 The following table gives the total yards gained by each of the top 10 NFL pass receivers in a single game during the 2004 regular National Football League season. Note that the games included in this data set are the ones with the highest total yards for each pass receiver during that season.

Player	Yards Gained
D. Bennett	233
R. Smith	208
J. Walker	200
R. Wayne	184
M. Muhammad	179
T. J. Houshmandzadeh	171
I. Bruce	170
A. Johnson	170
R. Gardner	167
J. Horn	167

a. Calculate the mean and median. Do these data have a mode(s)? Why or why not? Explain.
b. Find the range, variance, and standard deviation.

3.112 The following data give the numbers of driving citations received by 12 drivers.

4 8 0 3 11 7 4 14 8 13 7 9

a. Find the mean, median, and mode for these data.
b. Calculate the range, variance, and standard deviation.
c. Are the values of the summary measures in parts a and b population parameters or sample statistics?

3.113 The following table gives the distribution of the amounts of rainfall (in inches) for July 2005 for 50 cities.

Rainfall	Number of Cities
0 to less than 2	6
2 to less than 4	10
4 to less than 6	20
6 to less than 8	7
8 to less than 10	4
10 to less than 12	3

Find the mean, variance, and standard deviation. Are the values of these summary measures population parameters or sample statistics?

3.114 The following table gives the frequency distribution of the times (in minutes) that 50 commuter students at a large university spent looking for parking spaces on the first day of classes in the Spring semester of 2006.

Time	Number of Students
0 to less than 4	1
4 to less than 8	7
8 to less than 12	15
12 to less than 16	18
16 to less than 20	6
20 to less than 24	3

Find the mean, variance, and standard deviation. Are the values of these summary measures population parameters or sample statistics?

3.115 The mean time taken to learn the basics of a word processor by all students is 200 minutes with a standard deviation of 20 minutes.

- **a.** Using Chebyshev's theorem, find at least what percentage of students will learn the basics of this word processor in
  i. 160 to 240 minutes **ii.** 140 to 260 minutes
- ***b.** Using Chebyshev's theorem, find the interval that contains the time taken by at least 75% of all students to learn this word processor.

3.116 According to the *Statistical Abstract of the United States,* Americans were expected to spend an average of 1669 hours watching television in 2004 (*USA TODAY*, March 30, 2004). Assume that the average time spent watching television by Americans this year will have a distribution that is skewed to the right with a mean of 1750 hours and a standard deviation of 450 hours.

- **a.** Using Chebyshev's theorem, find at least what percentage of Americans will watch television this year for
  i. 850 to 2650 hours **ii.** 400 to 3100 hours
- ***b.** Using Chebyshev's theorem, find the interval that will contain the television viewing times of at least 84% of all Americans.

3.117 Refer to Exercise 3.115. Suppose the times taken to learn the basics of this word processor by all students have a bell-shaped distribution with a mean of 200 minutes and a standard deviation of 20 minutes.

- **a.** Using the empirical rule, find the percentage of students who learn the basics of this word processor in
  i. 180 to 220 minutes **ii.** 160 to 240 minutes
- ***b.** Using the empirical rule, find the interval that contains the time taken by 99.7% of all students to learn this word processor.

3.118 Assume that the annual earnings of all employees with CPA certification and 12 years of experience and working for large firms have a bell-shaped distribution with a mean of $134,000 and a standard deviation of $12,000.

- **a.** Using the empirical rule, find the percentage of all such employees whose annual earnings are between
  i. $98,000 and $170,000 **ii.** $110,000 and $158,000
- ***b.** Using the empirical rule, find the interval that contains the annual earnings of 68% of all such employees.

3.119 Refer to the data of Exercise 3.109 on the current annual incomes (in thousands of dollars) of the 10 members of the class of 2005 of the Metro Business College who were voted most likely to succeed.

59 68 44 68 57 104 56 44 47 40

- **a.** Determine the values of the three quartiles and the interquartile range. Where does the value of 40 fall in relation to these quartiles?
- **b.** Calculate the (approximate) value of the 70th percentile. Give a brief interpretation of this percentile.
- **c.** Find the percentile rank of 47. Give a brief interpretation of this percentile rank.

3.120 Refer to the data given in Exercise 3.111 on the total yards gained by the top 10 NFL pass receivers in single games during the 2004 regular National Football League season.

a. Determine the values of the three quartiles and the interquartile range. Where does the value of 179 lie in relation to these quartiles?

b. Calculate the (approximate) value of the 70th percentile. Give a brief interpretation of this percentile.

c. Find the percentile rank of 171. Give a brief interpretation of this percentile rank.

3.121 A student washes her clothes at a laundromat once a week. The data below give the time (in minutes) she spent in the laundromat for each of 15 randomly selected weeks. Here, time spent in the laundromat includes the time spent waiting for a machine to become available.

75	62	84	73	107	81	93	72
135	77	85	67	90	83	112	

Prepare a box-and-whisker plot. Is the data set skewed in any direction? If yes, is it skewed to the right or to the left? Does this data set contain any outliers?

3.122 The following data give the lengths of time (in weeks) taken to find a full-time job by 18 computer science majors who graduated in 2005 from a small college.

10	3	12	21	15	8	4	2	16
8	9	14	33	7	24	11	42	15

Make a box-and-whisker plot. Comment on the skewness of this data set. Does this data set contain any outliers?

Advanced Exercises

3.123 Melissa's grade in her math class is determined by three 100-point tests and a 200-point final exam. To determine the grade for a student in this class, the instructor will add the four scores together and divide this sum by 5 to obtain a percentage. This percentage must be at least 80 for a grade of B. If Melissa's three test scores are 75, 69, and 87, what is the minimum score she needs on the final exam to obtain a B grade?

3.124 Jeffrey is serving on a six-person jury for a personal-injury lawsuit. All six jurors want to award damages to the plaintiff but cannot agree on the amount of the award. The jurors have decided that each of them will suggest an amount that he or she thinks should be awarded; then they will use the mean of these six numbers as the award to recommend to the plaintiff.

a. Jeffrey thinks the plaintiff should receive \$20,000, but he thinks the mean of the other five jurors' recommendations will be about \$12,000. He decides to suggest an inflated amount so that the mean for all six jurors is \$20,000. What amount would Jeffrey have to suggest?

b. How might this jury revise its procedure to prevent a juror like Jeffrey from having an undue influence on the amount of damages to be awarded to the plaintiff?

3.125 The heights of five starting players on a basketball team have a mean of 76 inches, a median of 78 inches, and a range of 11 inches.

a. If the tallest of these five players is replaced by a substitute who is two inches taller, find the new mean, median, and range.

b. If the tallest player is replaced by a substitute who is four inches shorter, which of the new values (mean, median, range) could you determine, and what would their new values be?

3.126 On a 300-mile auto trip, Lisa averaged 52 miles per hour for the first 100 miles, 65 mph for the second 100 miles, and 58 mph for the last 100 miles.

a. How long did the 300-mile trip take?

b. Could you find Lisa's average speed for the 300-mile trip by calculating $(52 + 65 + 58)/3$? If not, find the correct average speed for the trip.

3.127 A small country bought oil from three different sources in one week, as shown in the following table.

Source	Barrels Purchased	Price Per Barrel
Mexico	1000	\$51
Kuwait	200	64
Spot Market	100	70

Find the mean price per barrel for all 1300 barrels of oil purchased in that week.

3.128 During the 2004 winter season, a homeowner received four deliveries of heating oil, as shown in the following table.

Gallons Purchased	Price Per Gallon
198	$1.10
173	1.25
130	1.28
124	1.33

The homeowner claimed that the mean price he paid for oil during the season was $(1.10 + 1.25 + 1.28 + 1.33)/4 = \$1.24$ per gallon. Do you agree with this claim? If not, explain why this method of calculating the mean is not appropriate in this case. Find the correct value of the mean price.

3.129 In the Olympic Games, when events require a subjective judgment of an athlete's performance, the highest and lowest of the judges' scores may be dropped. Consider a gymnast whose performance is judged by seven judges and the highest and the lowest of the seven scores are dropped.

- **a.** Gymnast A's scores in this event are 9.4, 9.7, 9.5, 9.5, 9.4, 9.6, and 9.5. Find this gymnast's mean score after dropping the highest and the lowest scores.
- **b.** The answer to part a is an example of what percentage of trimmed mean?
- **c.** Write another set of scores for a gymnast B so that gymnast A has a higher mean score than gymnast B based on the trimmed mean, but gymnast B would win if all seven scores were counted. Do not use any scores lower than 9.0.

3.130 A survey of young people's shopping habits in a small city during the summer months of 2005 showed the following: Shoppers aged 12–14 took an average of 8 shopping trips per month and spent an average of $14 per trip. Shoppers aged 15–17 took an average of 11 trips per month and spent an average of $18 per trip. Assume that this city has 1100 shoppers aged 12–14 and 900 shoppers aged 15–17.

- **a.** Find the total amount spent per month by all these 2000 shoppers in both age groups.
- **b.** Find the mean number of shopping trips per person per month for these 2000 shoppers.
- **c.** Find the mean amount spent per person per month by shoppers aged 12–17 in this city.

3.131 The following table shows the total population and the number of deaths (in thousands) due to heart attack for two age groups in Countries A and B for 2005.

	Age 30 and Under		Age 31 and Over	
	A	**B**	**A**	**B**
Population	40,000	25,000	20,000	35,000
Deaths due to heart attack	1000	500	2000	3000

- **a.** Calculate the death rate due to heart attack per 1000 population for the 30 and under age group for each of the two countries. Which country has the lower death rate in this age group?
- **b.** Calculate the death rates due to heart attack for the two countries for the 31 and over age group. Which country has the lower death rate in this age group?
- **c.** Calculate the death rate due to heart attack for the entire population of Country A; then do the same for Country B. Which country has the lower overall death rate?
- **d.** How can the country with lower death rate in both age groups have the higher overall death rate? (This phenomenon is known as Simpson's paradox.)

3.132 In a study of distances traveled to a college by commuting students, data from 100 commuters yielded a mean of 8.73 miles. After the mean was calculated, data came in late from three students, with distances of 11.5, 7.6, and 10.0 miles. Calculate the mean distance for all 103 students.

3.133 The test scores for a large statistics class have an unknown distribution with a mean of 70 and a standard deviation of 10.

- **a.** Find k so that at least 50% of the scores are within k standard deviations of the mean.
- **b.** Find k so that at most 10% of the scores are more than k standard deviations above the mean.

3.134 The test scores for a very large statistics class have a bell-shaped distribution with a mean of 70 points.

a. If 16% of all students in the class scored above 85, what is the standard deviation of the scores?

b. If 95% of the scores are between 60 and 80, what is the standard deviation?

3.135 How much does the typical American family spend to go away on vacation each year? Twenty-five randomly selected households reported the following vacation expenditures (rounded to the nearest hundred dollars) during the past year:

2500	500	800	0	100
0	200	2200	0	200
0	1000	900	321,500	400
500	100	0	8200	900
0	1700	1100	600	3400

a. Using both graphical and numerical methods, organize and interpret these data.

b. What measure of central tendency best answers the original question?

3.136 Actuaries at an insurance company must determine a premium for a new type of insurance. A random sample of 40 potential purchasers of this type of insurance were found to have suffered the following values of losses during the past year. These losses would have been covered by the insurance if it were available.

100	32	0	0	470	50	0	14,589	212	93
0	0	1127	421	0	87	135	420	0	250
12	0	309	0	177	295	501	0	143	0
167	398	54	0	141	0	3709	122	0	0

a. Find the mean, median, and mode of these 40 losses.

b. Which of the mean, median, or mode is largest?

c. Draw a box-and-whisker plot for these data, and describe the skewness, if any.

d. Which measure of central tendency should the actuaries use to determine the premium for this insurance?

3.137 A local golf club has men's and women's summer leagues. The following data give the scores for a round of 18 holes of golf for 17 men and 15 women randomly selected from their respective leagues.

Men	87	68	92	79	83	67	71	92	112
	75	77	102	79	78	85	75	72	
Women	101	100	87	95	98	81	117	107	103
	97	90	100	99	94	94			

a. Make a box-and-whisker plot for each of the data sets and use them to discuss the similarities and differences between the scores of the men and women golfers.

b. Compute the various descriptive measures you have learned for each sample. How do they compare?

3.138 Answer the following questions.

a. The total weight of all pieces of luggage loaded onto an airplane is 12,372 pounds, which works out to be an average of 51.55 pounds per piece. How many pieces of luggage are on the plane?

b. A group of seven friends, having just gotten back a chemistry exam, discuss their scores. Six of the students reveal that they received grades of 81, 75, 93, 88, 82, and 85, but the seventh student is reluctant to say what grade she received. After some calculation she announces that the group averaged 81 on the exam. What is her score?

3.139 Suppose that there are 150 freshmen engineering majors at a college and each of them will take the same five courses next semester. Four of these courses will be taught in small sections of 25 students each, whereas the fifth course will be taught in one section containing all 150 freshmen. To accommodate all 150 students, there must be six sections of each of the four courses taught in 25-student sections. Thus, there are 24 classes of 25 students each and one class of 150 students.

a. Find the mean size of these 25 classes.
b. Find the mean class size from a student's point of view, noting that each student has five classes containing 25, 25, 25, 25, and 150 students.

Are the means in parts a and b equal? If not, why not?

3.140 The following data give the weights (in pounds) of a random sample of 44 college students. (Here F and M indicate female and male, respectively.)

123 F	195 M	138 M	115 F	179 M	119 F
148 F	147 F	180 M	146 F	179 M	189 M
175 M	108 F	193 M	114 F	179 M	147 M
108 F	128 F	164 F	174 M	128 F	159 M
193 M	204 M	125 F	133 F	115 F	168 M
123 F	183 M	116 F	182 M	174 M	102 F
123 F	99 F	161 M	162 M	155 F	202 M
110 F	132 M				

Compute the mean, median, and standard deviation for the weights of all students, of men only, and of women only. Of the mean and median, which is the more informative measure of central tendency? Write a brief note comparing the three measures for all students, men only, and women only.

3.141 The distribution of the lengths of fish in a certain lake is not known, but it is definitely not bell-shaped. It is estimated that the mean length is 6 inches with a standard deviation of 2 inches.
a. At least what proportion of fish in the lake are between 3 inches and 9 inches long?
b. What is the smallest interval that will contain the lengths of at least 84% of the fish?
c. Find an interval so that fewer than 36% of the fish have lengths outside this interval.

3.142 The following stem-and-leaf diagram gives the distances (in thousands of miles) driven during the past year by a sample of drivers in a city.

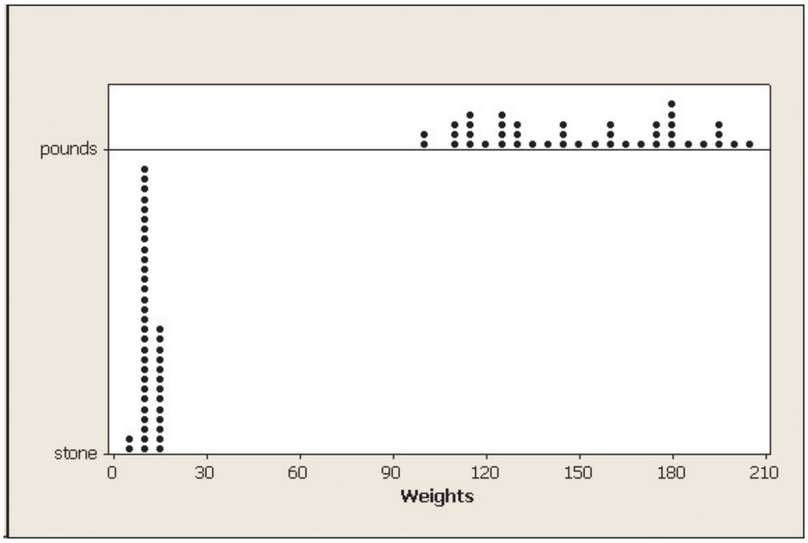
0	3 6 9
1	2 8 5 1 0 5
2	5 1 6
3	8
4	1
5	
6	2

a. Compute the sample mean, median, and mode for the data on distances driven.
b. Compute the range, variance, and standard deviation for these data.
c. Compute the first and third quartiles.
d. Compute the interquartile range. Describe what properties the interquartile range has. When would it be preferable to using the standard deviation when measuring variation?

3.143 Refer to the data in Problem 3.140. Two individuals, one from Canada and one from England, are interested in your analysis of these data but they need your results in different units. The Canadian individual wants the results in grams (1 pound = 435.59 grams). while the English individual wants the results in stone (1 stone = 14 pounds).
a. Convert the data on weights from pounds to grams, and then recalculate the mean, median, and standard deviation of weight for males and females separately. Repeat the procedure, changing the unit from pounds to stones.
b. Convert your answers from Problem 3.140 to grams and stone. What do you notice about these answers and your answers from part a?
c. What happens to the values of the mean, median, and standard deviation when you convert from a larger unit to a smaller unit (e.g., from pounds to grams)? Does the same thing happen if you convert from a smaller unit (e.g., pounds) to a larger unit (e.g., stone)?
d. Figure 3.15 on the next page gives a stacked dotplot of these weights in pounds and stone. Which of these two distributions has more variability? Use your results from parts a to c to explain why this is the case.
e. Now consider the weights in pounds and grams. Make a stacked dotplot for these data and answer part d.

3.144 Although the standard workweek is 40 hours a week, many people work a lot more than 40 hours a week. The data on the next page give the numbers of hours worked last week by 50 people.

Figure 3.15 Stacked Dotplot of Weights in Stone and Pounds.



40.5	41.3	41.4	41.5	42.0	42.2	42.4	42.4	42.6	43.3
43.7	43.9	45.0	45.0	45.2	45.8	45.9	46.2	47.2	47.5
47.8	48.2	48.3	48.8	49.0	49.2	49.9	50.1	50.6	50.6
50.8	51.5	51.5	52.3	52.3	52.6	52.7	52.7	53.4	53.9
54.4	54.8	55.0	55.4	55.4	55.4	56.2	56.3	57.8	58.7

a. The sample mean and sample standard deviation for this data set are 49.012 and 5.080, respectively. Using the Chebyshev's theorem, calculate the intervals that contain at least 75%, 88.89%, and 93.75% of the data.

b. Determine the actual percentages of the given data values that fall in each of the intervals that you calculated in part a. Also calculate the percentage of the data values that fall within one standard deviation of the mean.

c. Do you think the lower endpoints provided by Chebyshev's Theorem in part a are useful for this problem? Explain your answer.

d. Suppose that the individual with the first number (54.4) in the fifth row of the data is a workaholic who actually worked 84.4 hours last week, and not 54.4 hours. With this change now $\bar{x} = 49.61$ and $s = 7.10$. Recalculate the intervals for part a and the actual percentages for part b. Did your percentages change a lot or a little?

e. How many standard deviations above the mean would you have to go to capture all 50 data values? What is the lower bound for the percentage of the data that should fall in the interval, according to Chebyshev?

3.145 Refer to the women's golf scores in Exercise 3.137. It turns out that 117 was mistakenly entered. Although this person still had the highest score among the 15 women, her score was not a mild or extreme outlier according to the box-and-whisker plot, nor was she tied for the highest score. What are the possible scores that she could have shot?

APPENDIX 3.1

A3.1.1 BASIC FORMULAS FOR THE VARIANCE AND STANDARD DEVIATION FOR UNGROUPED DATA

Example 3–25 illustrates how to use the basic formulas to calculate the variance and standard deviation for ungrouped data. From Section 3.2.2, the basic formulas for variance for ungrouped data are

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

where σ^2 is the population variance and s^2 is the sample variance.

In either case, the standard deviation is obtained by taking the square root of the variance.

EXAMPLE 3–25 Refer to Example 3–12, where we used the short-cut formulas to compute the variance and standard deviation for the data on the total wealth (in billions of dollars) of five persons. Calculate the variance and standard deviation for those data using the basic formula.

Calculating the variance and standard deviation for ungrouped data using basic formulas.

Solution Let x denote the total wealth (in billions of dollars) of a person. Table 3.14 shows all the required calculations to find the variance and standard deviation.

Table 3.14

x	$(x - \bar{x})$	$(x - \bar{x})^2$
46.5	46.5 − 19.1 = 27.4	750.76
18.0	18.0 − 19.1 = −1.1	1.21
16.0	16.0 − 19.1 = −3.1	9.61
7.8	7.8 − 19.1 = −11.3	127.69
7.2	7.2 − 19.1 = −11.9	141.61
$\Sigma x = 95.5$		$\Sigma(x - \bar{x})^2 = 1030.88$

The following steps are performed to compute the variance and standard deviation.

Step 1. Find the mean as follows:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{95.5}{5} = 19.1$$

Step 2. Calculate $x - \bar{x}$, the deviation of each value of x from the mean. The results are shown in the second column of Table 3.14.

Step 3. Square each of the deviations of x from $\bar{x}$; that is, calculate each of the $(x - \bar{x})^2$ values. These values are called the *squared deviations*, and they are recorded in the third column.

Step 4. Add all the squared deviations to obtain $\Sigma(x - \bar{x})^2$; that is, sum all the values given in the third column of Table 3.14. This gives

$$\Sigma(x - \bar{x})^2 = 1030.88$$

Step 5. Obtain the sample variance by dividing the sum of the squared deviations by $n - 1$. Thus

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{1030.88}{5 - 1} = 257.72$$

Step 6. Obtain the sample standard deviation by taking the positive square root of the variance. Hence,

$$s = \sqrt{257.72} = \mathbf{16.05366 = \$16.05\ billion}$$

A3.1.2 BASIC FORMULAS FOR THE VARIANCE AND STANDARD DEVIATION FOR GROUPED DATA

Example 3–26 demonstrates how to use the basic formulas to calculate the variance and standard deviation for grouped data. The basic formulas for these calculations are

$$\sigma^2 = \frac{\Sigma f(m - \mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\Sigma f(m - \bar{x})^2}{n - 1}$$

where σ^2 is the population variance, s^2 is the sample variance, m is the midpoint of a class, and f is the frequency of a class.

In either case, the standard deviation is obtained by taking the square root of the variance.

EXAMPLE 3–26 In Example 3–17, we used the short-cut formula to compute the variance and standard deviation for the data on the numbers of orders received each day during the past 50 days at the office of a mail-order company. Calculate the variance and standard deviation for those data using the basic formula.

Calculating the variance and standard deviation for grouped data using basic formulas.

Solution All the required calculations to find the variance and standard deviation appear in Table 3.15.

Table 3.15

Number of Orders	f	m	mf	$m - \bar{x}$	$(m - \bar{x})^2$	$f(m - \bar{x})^2$
10–12	4	11	44	−5.64	31.8096	127.2384
13–15	12	14	168	−2.64	6.9696	83.6352
16–18	20	17	340	.36	.1296	2.5920
19–21	14	20	280	3.36	11.2896	158.0544
	$n = 50$		$\Sigma mf = 832$			$\Sigma f(m - \bar{x})^2 = 371.5200$

The following steps are performed to compute the variance and standard deviation using the basic formula.

Step 1. Find the midpoint of each class. Multiply the corresponding values of m and f. Find Σmf. From Table 3.15, $\Sigma mf = 832$.

Step 2. Find the mean as follows:

$$\bar{x} = \Sigma mf/n = 832/50 = 16.64$$

Step 3. Calculate $m - \bar{x}$, the deviation of each value of m from the mean. These calculations are done in the fifth column of Table 3.15.

Step 4. Square each of the deviations $m - \bar{x}$; that is, calculate each of the $(m - \bar{x})^2$ values. These are called *squared deviations*, and they are recorded in the sixth column.

Step 5. Multiply the squared deviations by the corresponding frequencies (see the seventh column of Table 3.15). Adding the values of the seventh column, we obtain

$$\Sigma f(m - \bar{x})^2 = 371.5200$$

Step 6. Obtain the sample variance by dividing $\Sigma f(m - \bar{x})^2$ by $n - 1$. Thus,

$$s^2 = \frac{\Sigma f(m - \bar{x})^2}{n - 1} = \frac{371.5200}{50 - 1} = \mathbf{7.5820}$$

Step 7. Obtain the standard deviation by taking the positive square root of the variance.

$$s = \sqrt{s^2} = \sqrt{7.5820} = \mathbf{2.75\ orders}$$

Self-Review Test

1. The value of the middle term in a ranked data set is called the
   a. mean **b.** median **c.** mode
2. Which of the following summary measures is/are influenced by extreme values?
   a. mean **b.** median **c.** mode **d.** range
3. Which of the following summary measures can be calculated for qualitative data?
   a. mean **b.** median **c.** mode
4. Which of the following can have more than one value?
   a. mean **b.** median **c.** mode
5. Which of the following is obtained by taking the difference between the largest and the smallest values of a data set?
   a. variance **b.** range **c.** mean
6. Which of the following is the mean of the squared deviations of x values from the mean?
   a. standard deviation **b.** population variance **c.** sample variance
7. The values of the variance and standard deviation are
   a. never negative **b.** always positive **c.** never zero

8. A summary measure calculated for the population data is called
a. a population parameter **b.** a sample statistic **c.** an outlier

9. A summary measure calculated for the sample data is called a
a. population parameter **b.** sample statistic **c.** box-and-whisker plot

10. Chebyshev's theorem can be applied to
a. any distribution **b.** bell-shaped distributions only **c.** skewed distributions only

11. The empirical rule can be applied to
a. any distribution **b.** bell-shaped distributions only **c.** skewed distributions only

12. The first quartile is a value in a ranked data set such that about
a. 75% of the values are smaller and about 25% are larger than this value
b. 50% of the values are smaller and about 50% are larger than this value
c. 25% of the values are smaller and about 75% are larger than this value

13. The third quartile is a value in a ranked data set such that about
a. 75% of the values are smaller and about 25% are larger than this value
b. 50% of the values are smaller and about 50% are larger than this value
c. 25% of the values are smaller and about 75% are larger than this value

14. The 75th percentile is a value in a ranked data set such that about
a. 75% of the values are smaller and about 25% are larger than this value
b. 25% of the values are smaller and about 75% are larger than this value

15. The following data give the numbers of times 10 persons used their credit cards during the past three months.

9 6 28 14 2 18 7 3 16 6

Calculate the mean, median, mode, range, variance, and standard deviation.

16. The mean, as a measure of central tendency, has the disadvantage of being influenced by extreme values. Illustrate this point with an example.

17. The range, as a measure of spread, has the disadvantage of being influenced by extreme values. Illustrate this point with an example.

18. When is the value of the standard deviation for a data set zero? Give one example of such a data set. Calculate the standard deviation for that data set to show that it is zero.

19. The following table gives the frequency distribution of the numbers of computers sold during the past 25 weeks at a computer store.

Computers Sold	Frequency
4 to 9	2
10 to 15	4
16 to 21	10
22 to 27	6
28 to 33	3

a. What does the frequency column in the table represent?
b. Calculate the mean, variance, and standard deviation.

20. The cars owned by all people living in a city are, on average, 7.3 years old with a standard deviation of 2.2 years.
a. Using Chebyshev's theorem, find at least what percentage of the cars in this city are
i. 1.8 to 12.8 years old **ii.** .7 to 13.9 years old
b. Using Chebyshev's theorem, find the interval that contains the ages of at least 75% of the cars owned by all people in this city.

21. The ages of cars owned by all people living in a city have a bell-shaped distribution with a mean of 7.3 years and a standard deviation of 2.2 years.
a. Using the empirical rule, find the percentage of cars in this city that are
i. 5.1 to 9.5 years old **ii.** .7 to 13.9 years old
b. Using the empirical rule, find the interval that contains the ages of 95% of the cars owned by all people in this city.

22. The following data give the number of times the metal detector was set off by passengers at a small airport during 15 consecutive half-hour periods on February 1, 2006.

7	2	12	13	0	8	10	
15	3	5	14	20	1	11	4

- **a.** Calculate the three quartiles and the interquartile range. Where does the value of 4 lie in relation to these quartiles?
- **b.** Find the (approximate) value of the 60th percentile. Give a brief interpretation of this value.
- **c.** Calculate the percentile rank of 12. Give a brief interpretation of this value.

23. Make a box-and-whisker plot for the data on the number of times passengers set off the airport metal detector given in Problem 22. Comment on the skewness of this data set.

***24.** The mean weekly wages of a sample of 15 employees of a company are \$435. The mean weekly wages of a sample of 20 employees of another company are \$490. Find the combined mean for these 35 employees.

***25.** The mean GPA of five students is 3.21. The GPAs of four of these five students are 3.85, 2.67, 3.45, and 2.91. Find the GPA of the fifth student.

***26.** The following are the prices (in thousands of dollars) of 10 houses sold recently in a city:

179	166	58	207	287	149	193	2534	163	238

Calculate the 10% trimmed mean for this data set. Do you think the 10% trimmed mean is a better summary measure than the (simple) mean (i.e., the mean of all 10 values) for these data? Briefly explain why or why not.

***27.** Consider the following two data sets.

Data Set I:	8	16	20	35
Data Set II:	5	13	17	32

Note that each value of the second data set is obtained by subtracting 3 from the corresponding value of the first data set.

- **a.** Calculate the mean for each of these two data sets. Comment on the relationship between the two means.
- **b.** Calculate the standard deviation for each of these two data sets. Comment on the relationship between the two standard deviations.

Mini-Projects

■ MINI-PROJECT 3–1

Refer to the data you collected for Mini-Project 1–1 of Chapter 1 and analyzed graphically in Mini-Project 2–1 of Chapter 2. Write a report summarizing those data. This report should include answers to at least the following questions.

- **a.** Calculate the summary measures (mean, standard deviation, five-number summary, interquartile range) for the variables you graphed in Mini-Project 2–1. Do this for the entire data set, as well as for the different groups formed by the categorical variable that you used to divide the data set in Mini-Project 2–1.
- **b.** Are the summary measures for the various groups similar to those for the entire data set? If not, which ones differ and how do they differ? Make the same comparisons among the summary measures for various groups. Do the groups have similar levels of variability? Explain how you can determine this from the graphs that you created in Mini-Project 2–1.
- **c.** Draw a box-and-whisker plot for the entire data set. Also draw side-by-side box-and-whisker plots for the various groups. Are there any outliers? If so, are there any values that are outliers in any of the groups but not in the entire data set? Does the plot show any skewness?
- **d.** Discuss which measures for the center and spread would be more appropriate to use to describe your data set. Also, discuss your reasons for using those measures.

■ MINI-PROJECT 3–2

You are employed as a statistician for a company that makes household products, which are sold by part-time salespersons who work during their spare time. The company has four salespersons employed in a small town. Let us denote these salespersons by A, B, C, and D. The sales records (in dollars) for the past six weeks for these four salespersons are shown in the following table.

Week	A	B	C	D
1	1774	2205	1330	1402
2	1808	1507	1295	1665
3	1890	2352	1502	1530
4	1932	1939	1104	1826
5	1855	2052	1189	1703
6	1726	1630	1441	1498

Your supervisor has asked you to prepare a brief report comparing the sales volumes and the consistency of sales of these four salespersons. Use the mean sales for each salesperson to compare the sales volumes, and then choose an appropriate statistical measure to compare the consistency of sales. Make the calculations and write a report.

DECIDE FOR YOURSELF

Deciding Where to Live

By the time you get to college, you must have heard it over and over again: "A picture is worth a thousand words." Now we have pictures and numbers discussed in Chapters 2 and 3, respectively. Why both? Well, although each one of them acts as a summary of a data set, it is a combination of the pictures and numbers that tells a big part of the story without having to look at the entire data set. Suppose that you ask a realtor for information on the prices of homes in two different but comparable suburbs. Let us call these Suburbs A and B. The realtor provides you with the following information that is obtained from a random sample of 40 houses in each suburb:

a. The average price of homes in each of the two suburbs
b. The five-number summary of prices of homes in each neighborhood
c. The histogram of the distribution of home prices for each suburb

All the information provided by the realtor is given in the following two tables and two histograms shown in Figures 3.16 and 3.17. Note that the second table gives the minimum and maximum prices of homes (in thousands of dollars) for each suburb along with the values of Q_1, median, and Q_3 (in thousands of dollars).

Suburb	A	B
Average Price (in thousands of dollars)	221.9	220.03

	Minimum	Q_1	Median	Q_3	Maximum
Suburb A	151.0	175.5	188.0	199.5	587.0
Suburb B	187.0	210.0	222.5	228.0	250.0

Before you decide which suburb you should buy the house in, answer the following questions:

1. Examine the summary statistics and graphs given here.

2. Explain how the information given here can help you to make a decision about the suburb where you should look for a house to buy.

3. Explain how and why you might be misled by simply looking at the average prices if you are looking to spend less money to buy a house.

4. Is there any information about the suburbs not given here that you will like to obtain before making a decision about the suburb where you should buy a house?

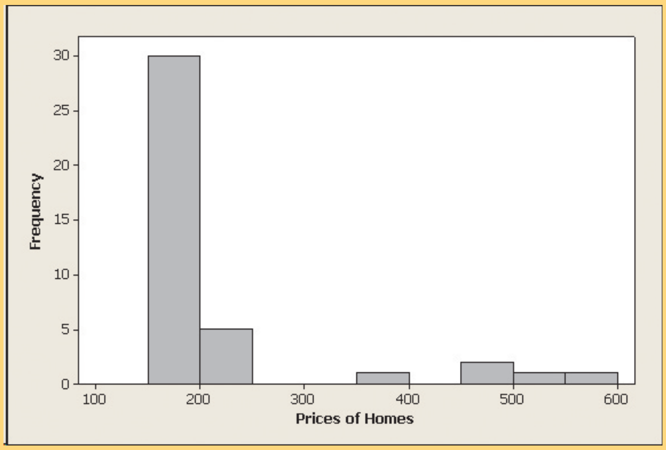


Figure 3.16 Histogram of Prices of Homes in Suburb A.

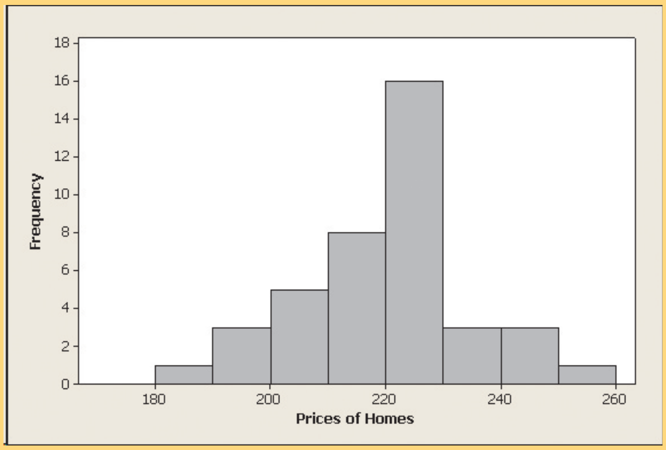
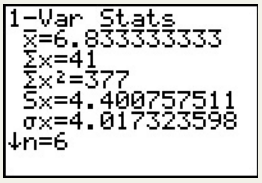


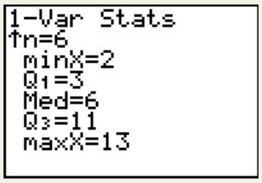
Figure 3.17 Histogram of Prices of Homes in Suburb B.

TECHNOLOGY INSTRUCTION — Numerical Descriptive Measures

TI-84



Screen 3.1

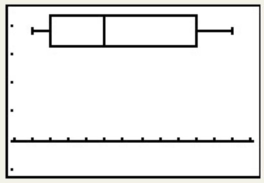


Screen 3.2

1. To calculate the **sample statistics** (e.g., mean, standard deviation, and five-number summary), first enter your data into a list such as L1, then select **STAT>CALC>1-Var Stats**, and press **Enter**. Access the name of your list by pressing **2nd>STAT** and scrolling through the list of names until you get to your list name. Press **Enter**. You will obtain the output shown in Screens 3.1 and 3.2.

Screen 3.1 shows, in this order, the sample mean, the sum of the data values, the sum of the squared data values, the sample standard deviation, the value of the population standard deviation (you will use this only when your data constitute a census instead of a sample), and the number of data values (e.g., the sample or population size). Pressing the downward arrow key will show the five-number summary, which is shown in Screen 3.2.

2. Constructing a box-and-whisker plot is similar to constructing a histogram. First enter your data into a list such as L1, then select **STAT PLOT** and go into one of the three plots. Make sure the plot is turned on. For the type, select the second row, first column (this boxplot will display outliers, if there are any). Enter the name of your list for **XList**. Select **ZOOM>9** to display the plot as shown in Screen 3.3.

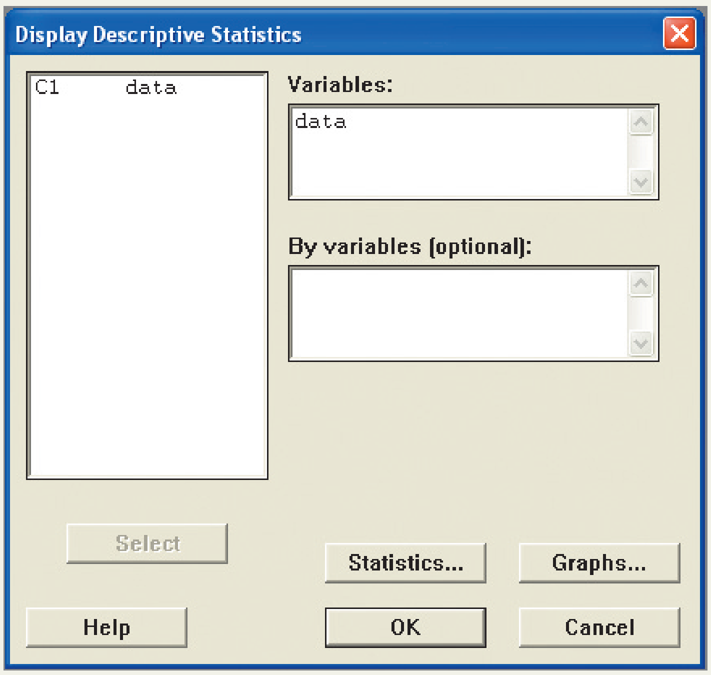


Screen 3.3

MINITAB

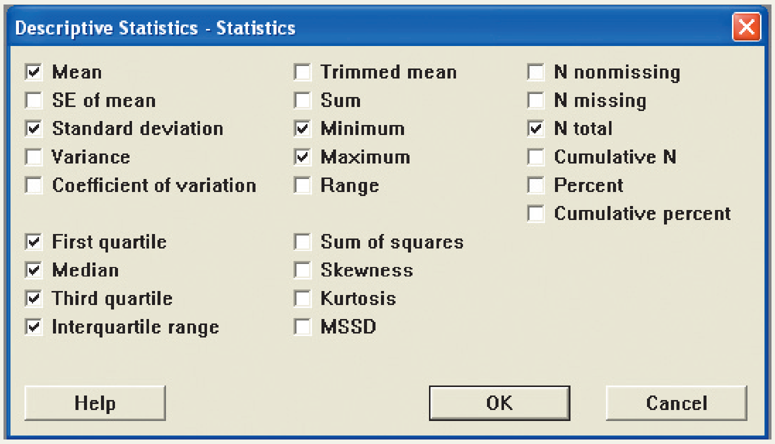
1. To find the sample statistics (e.g., the mean, standard deviation, and five-number summary), first enter the given data in a column such as C1, and then select **Stat>Basic Statistics> Display Descriptive Statistics**. In the dialog box you obtain, enter the name of the column where your data are stored in the **Variables** box as shown in Screen 3.4. Click the **Statistics** button in this dialog box and choose the summary measures you want to

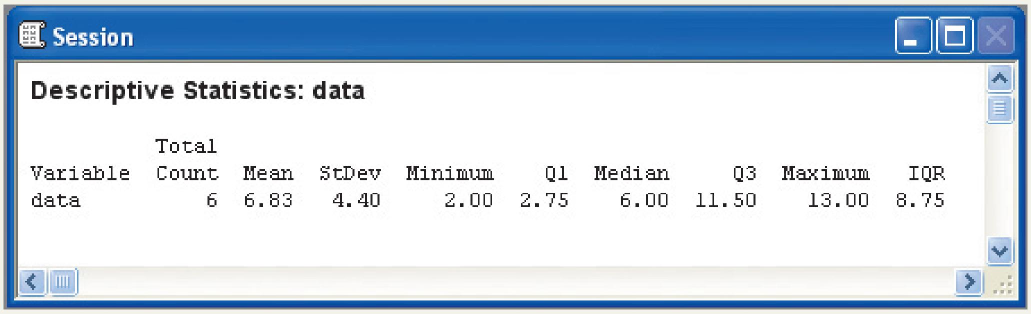
Screen 3.4



calculate in the new dialog box as shown in Screen 3.5. Click **OK** in both dialog boxes. The output will appear in the **Session** window, which is shown in Screen 3.6 here.

Screen 3.5

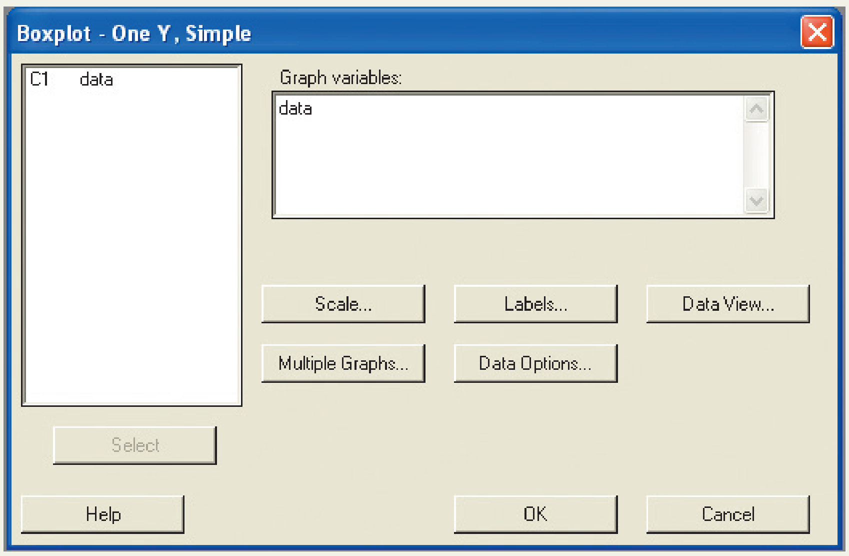




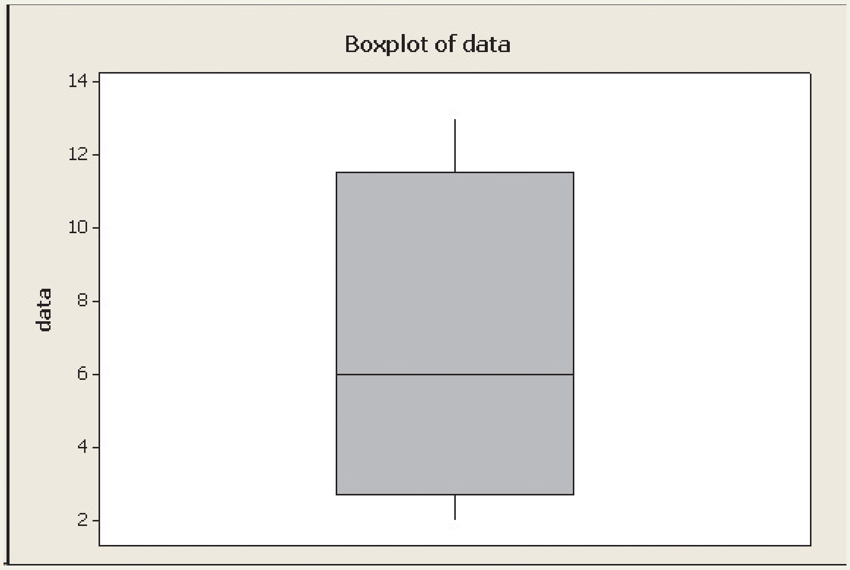
Screen 3.6

2. To create a box-and-whisker plot, enter the given data in a column such as C1, select **Graph>Boxplot>Simple**, and click **OK**. In the dialog box you obtain, enter the name of the column with data in the **Graph Variables** box (see Screen 3.7) and click **OK**. The boxplot shown in Screen 3.8 will appear.

Screen 3.7



Screen 3.8



EXCEL

1. For each of the commands in **Excel**,
  a. Type **=command(**
  b. Select the range of data
  c. Type a right parenthesis, and then press **Enter**.

2. To find the mean, use the command **average**. (See Screens 3.9 and 3.10)

3. To find the median, use the command **median**.

4. To find the mode, use the command **mode**.

	A	B	C
1	Data	Average	
2			
3	2	=average(A3:A8)	
4	3		
5	5		
6	7		
7	11		
8	13		

Screen 3.9

	A	B	C
1	Data	Average	
2			
3	2	6.833333	
4	3		
5	5		
6	7		
7	11		
8	13		

Screen 3.10

5. To find the standard deviation, use the command **stdev**.

6. To find the first or third quartiles:
  a. Type **=quartile(**
  b. Select the range of data and then type a comma
  c. Type **1** for the first quartile, **3** for the third quartile
  d. Type a right parenthesis, and then press **Enter**.

7. To find the kth percentile:
  a. Type **=percentile(**
  b. Select the range of data and then type a comma
  c. Type the value of **k** followed by a right parenthesis, and then press **Enter**.

TECHNOLOGY ASSIGNMENTS

TA3.1 Refer to the subsample taken in the Computer Assignment TA2.3 of Chapter 2 from the sample data on the time taken to run the Manchester Road Race. Find the mean, median, range, and standard deviation for those data.

TA3.2 Refer to the data on phone charges given in Data Set I. From that data set select the 4th value and then select every 10th value after that (i.e., select the 4th, 14th, 24th, 34th . . . values). Such a sample taken from a population is called a *systematic random sample.* Find the mean, median, standard deviation, first quartile, and third quartile for the phone charges for this subsample.

TA3.3 Refer to Data Set I on the prices of various products in different cities across the country. Select a subsample of the prices of regular unleaded gas for 40 cities. Find the mean, median, and standard deviation for the data of this subsample.

TA3.4 Refer to Data of TA3.3. Make a box-and-whisker plot for those data.

TA3.5 Refer to Data Set I on the prices of various products in different cities across the country. Make a box-and-whisker plot for the data on the monthly telephone charges.

TA3.6 Refer to the data on the numbers of computer keyboards assembled at the Twentieth Century Electronics Company for a sample of 25 days given in Exercise 3.104. Prepare a box-and-whisker plot for those data.