CHAPTER 3

Descriptive Statistics

LEARNING OBJECTIVES

The focus of Chapter 3 is the use of statistical techniques to describe data, thereby enabling you to:

1. Apply various measures of central tendency—including the mean, median, and mode—to a set of ungrouped data
2. Apply various measures of variability—including the range, interquartile range, mean absolute deviation, variance, and standard deviation (using the empirical rule and Chebyshev's theorem)—to a set of ungrouped data
3. Compute the mean, median, mode, standard deviation, and variance of grouped data
4. Describe a data distribution statistically and graphically using skewness, kurtosis, and box-and-whisker plots
5. Use computer packages to compute various measures of central tendency, variation, and shape on a set of data, as well as to describe the data distribution graphically

Stewart Cohen/Stone/Getty Images

Laundry Statistics

Decision Dilemma

According to Procter & Gamble, 35 billion loads of laundry are run in the United States each year. Every second 1,100 loads are started. Statistics show that one person in the United States generates a quarter of a ton of dirty clothing each year. Americans appear to be spending more time doing laundry than they did 40 years ago. Today, the average American woman spends seven to nine hours a week on laundry. However, industry research shows that the result is dirtier laundry than in other developed countries. Various companies market new and improved versions of washers and detergents. Yet, Americans seem to be resistant to manufacturers' innovations in this area. In the United States, the average washing machine uses about 16 gallons of water. In Europe, the figure is about 4 gallons. The average wash cycle of an American wash is about 35 minutes compared to 90 minutes in Europe. Americans prefer top loading machines because they do not have to bend over, and the top loading machines are larger. Europeans use the smaller front-loading machines because of smaller living spaces.

Managerial and Statistical Questions

Virtually all of the statistics cited here are gleaned from studies or surveys.

1. Suppose a study of laundry usage is done in 50 U.S. households that contain washers and dryers. Water measurements are taken for the number of gallons of water used by each washing machine in completing a cycle. The following data are the number of gallons used by each washing machine during the washing cycle. Summarize the data so that study findings can be reported.

15	17	16	15	16	17	18	15	14	15
16	16	17	16	15	15	17	14	15	16
16	17	14	15	12	15	16	14	14	16
15	13	16	17	17	15	16	16	16	14
17	16	17	14	16	13	16	15	16	15

2. The average wash cycle for an American wash is 35 minutes. Suppose the standard deviation of a wash cycle for an American wash is 5 minutes. Within what range of time do most American wash cycles fall?

Source: Adapted from Emily Nelson, "In Doing Laundry, Americans Cling to Outmoded Ways," *The Wall Street Journal*, May 16, 2002, pp. A1 & A10.

Chapter 2 presented graphical techniques for organizing and displaying data. Even though such graphical techniques allow the researcher to make some general observations about the shape and spread of the data, a more complete understanding of the data can be attained by summarizing the data using statistics. This chapter presents such statistical measures, including measures of central tendency, measures of variability, and measures of shape. The computation of these measures is different for ungrouped and grouped data. Hence we present some measures for both ungrouped and grouped data.

3.1 MEASURES OF CENTRAL TENDENCY: UNGROUPED DATA

Interactive Applet

One type of measure that is used to describe a set of data is the **measure of central tendency**. Measures of central tendency *yield information about the center, or middle part, of a group of numbers.* Table 3.1 displays offer price for the 20 largest U.S. initial public offerings in a recent year according to Securities Data. For these data, measures of central tendency can yield such information as the average offer price, the middle offer price, and the most frequently occurring offer price. Measures of central tendency do not focus on the span of the data set or how far values are from the middle numbers. The measures of central tendency presented here for ungrouped data are the mode, the median, the mean, percentiles, and quartiles.

TABLE 3.1
Offer Prices for the 20 Largest U.S. Initial Public Offerings in a Recent Year

$14.25	$19.00	$11.00	$28.00
24.00	23.00	43.25	19.00
27.00	25.00	15.00	7.00
34.22	15.50	15.00	22.00
19.00	19.00	27.00	21.00

Mode

The **mode** is *the most frequently occurring value in a set of data.* For the data in Table 3.1 the mode is $19.00 because the offer price that recurred the most times (four) was $19.00. Organizing the data into an ordered array (an ordering of the numbers from smallest to largest) helps to locate the mode. The following is an ordered array of the values from Table 3.1.

7.00	11.00	14.25	15.00	15.00	15.50	19.00	19.00	19.00	19.00
21.00	22.00	23.00	24.00	25.00	27.00	27.00	28.00	34.22	43.25

This grouping makes it easier to see that 19.00 is the most frequently occurring number.

In the case of a tie for the most frequently occurring value, two modes are listed. Then the data are said to be **bimodal**. If a set of data is not exactly bimodal but contains two values that are more dominant than others, some researchers take the liberty of referring to the data set as bimodal even without an exact tie for the mode. Data sets with more than two modes are referred to as **multimodal**.

In the world of business, the concept of mode is often used in determining sizes. As an example, manufacturers who produce cheap rubber flip-flops that are sold for as little as $1.00 around the world might only produce them in one size in order to save on machine setup costs. In determining the one size to produce, the manufacturer would most likely produce flip-flops in the modal size. The mode is an appropriate measure of central tendency for nominal-level data.

Median

The **median** is *the middle value in an ordered array of numbers.* For an array with an odd number of terms, the median is the middle number. For an array with an even number of terms, the median is the average of the two middle numbers. The following steps are used to determine the median.

STEP 1. Arrange the observations in an ordered data array.

STEP 2. For an odd number of terms, find the middle term of the ordered array. It is the median.

STEP 3. For an even number of terms, find the average of the middle two terms. This average is the median.

Suppose a business researcher wants to determine the median for the following numbers.

15 11 14 3 21 17 22 16 19 16 5 7 19 8 9 20 4

The researcher arranges the numbers in an ordered array.

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21 22

Because the array contains 17 terms (an odd number of terms), the median is the middle number, or 15.

If the number 22 is eliminated from the list, the array would contain only 16 terms.

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21

Now, for an even number of terms, the statistician determines the median by averaging the two middle values, 14 and 15. The resulting median value is 14.5.

Another way to locate the median is by finding the $(n + 1)/2$ term in an ordered array. For example, if a data set contains 77 terms, the median is the 39th term. That is,

$$\frac{n+1}{2} = \frac{77+1}{2} = \frac{78}{2} = 39\text{th term}$$

This formula is helpful when a large number of terms must be manipulated.

Consider the offer price data in Table 3.1. Because this data set contains 20 values, or $n = 20$, the median for these data is located at the $(20 + 1)/2$ term, or the 10.5th term. This equation indicates that the median is located halfway between the 10th and 11th terms

or the average of 19.00 and 21.00. Thus, the median offer price for the largest 20 U.S. initial public offerings is $20.00.

The median is unaffected by the magnitude of extreme values. This characteristic is an advantage, because large and small values do not inordinately influence the median. For this reason, the median is often the best measure of location to use in the analysis of variables such as house costs, income, and age. Suppose, for example, that a real estate broker wants to determine the median selling price of 10 houses listed at the following prices.

$67,000	$105,000	$148,000	$5,250,000
91,000	116,000	167,000	
95,000	122,000	189,000	

The median is the average of the two middle terms, $116,000 and $122,000, or $119,000. This price is a reasonable representation of the prices of the 10 houses. Note that the house priced at $5,250,000 did not enter into the analysis other than to count as one of the 10 houses. If the price of the tenth house were $200,000, the results would be the same. However, if all the house prices were averaged, the resulting average price of the original 10 houses would be $635,000, higher than 9 of the 10 individual prices.

A disadvantage of the median is that not all the information from the numbers is used. For example, information about the specific asking price of the most expensive house does not really enter into the computation of the median. The level of data measurement must be at least ordinal for a median to be meaningful.

Mean

The **arithmetic mean** is *the average of a group of numbers* and is computed by summing all numbers and dividing by the number of numbers. Because the arithmetic mean is so widely used, most statisticians refer to it simply as the *mean*.

The population mean is represented by the Greek letter mu (μ). The sample mean is represented by $\bar{x}$. The formulas for computing the population mean and the sample mean are given in the boxes that follow.

POPULATION MEAN

$$\mu = \frac{\Sigma x}{N} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N}$$

SAMPLE MEAN

$$\bar{x} = \frac{\Sigma x}{n} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

The capital Greek letter sigma (Σ) is commonly used in mathematics to represent a summation of all the numbers in a grouping.* Also, N is the number of terms in the population, and n is the number of terms in the sample. The algorithm for computing a mean is to sum all the numbers in the population or sample and divide by the number of terms. It is inappropriate to use the mean to analyze data that are not at least interval level in measurement.

Suppose a company has five departments with 24, 13, 19, 26, and 11 workers each. The *population mean* number of workers in each department is 18.6 workers. The computations follow.

$$\begin{aligned} &24 \\ &13 \\ &19 \\ &26 \\ &\underline{11} \\ \Sigma x = \ &93 \end{aligned}$$

*The mathematics of summations is not discussed here. A more detailed explanation is given in WileyPLUS, Chapter 3.

and

$$\mu = \frac{\Sigma x}{N} = \frac{93}{5} = 18.6$$

The calculation of a sample mean uses the same algorithm as for a population mean and will produce the same answer if computed on the same data. However, it is inappropriate to compute a sample mean for a population or a population mean for a sample. Because both populations and samples are important in statistics, a separate symbol is necessary for the population mean and for the sample mean.

DEMONSTRATION PROBLEM 3.1

The number of U.S. cars in service by top car rental companies in a recent year according to *Auto Rental News* follows.

Company	Number of Cars in Service
Enterprise	643,000
Hertz	327,000
National/Alamo	233,000
Avis	204,000
Dollar/Thrifty	167,000
Budget	144,000
Advantage	20,000
U-Save	12,000
Payless	10,000
ACE	9,000
Fox	9,000
Rent-A-Wreck	7,000
Triangle	6,000

Compute the mode, the median, and the mean.

Solution

Mode: 9,000

Median: With 13 different companies in this group, $N = 13$. The median is located at the $(13 + 1)/2 = 7$th position. Because the data are already ordered, the 7th term is 20,000, which is the median.

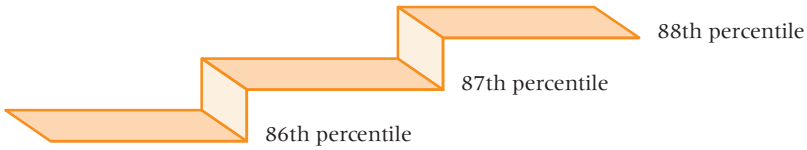
Mean: The total number of cars in service is 1,791,000 $= \Sigma x$

$$\mu = \frac{\Sigma x}{N} = \frac{1{,}791{,}000}{13} = 137{,}769.23$$

The mean is affected by each and every value, which is an advantage. The mean uses all the data, and each data item influences the mean. It is also a disadvantage because extremely large or small values can cause the mean to be pulled toward the extreme value. Recall the preceding discussion of the 10 house prices. If the mean is computed for the 10 houses, the mean price is higher than the prices of 9 of the houses because the $5,250,000 house is included in the calculation. The total price of the 10 houses is $6,350,000, and the mean price is $635,000.

The mean is the most commonly used measure of central tendency because it uses each data item in its computation, it is a familiar measure, and it has mathematical properties that make it attractive to use in inferential statistics analysis.

FIGURE 3.1
Stair-Step Percentiles



Percentiles

Percentiles are *measures of central tendency that divide a group of data into 100 parts.* There are 99 percentiles because it takes 99 dividers to separate a group of data into 100 parts. The nth percentile is the value such that at least n percent of the data are below that value and at most $(100 - n)$ percent are above that value. Specifically, the 87th percentile is a value such that at least 87% of the data are below the value and no more than 13% are above the value. Percentiles are "stair-step" values, as shown in Figure 3.1, because the 87th percentile and the 88th percentile have no percentile between. If a plant operator takes a safety examination and 87.6% of the safety exam scores are below that person's score, he or she still scores at only the 87th percentile, even though more than 87% of the scores are lower.

Percentiles are widely used in reporting test results. Almost all college or university students have taken the SAT, ACT, GRE, or GMAT examination. In most cases, the results for these examinations are reported in percentile form and also as raw scores. Shown next is a summary of the steps used in determining the location of a percentile.

Steps in Determining the Location of a Percentile

1. Organize the numbers into an ascending-order array.
2. Calculate the percentile location (i) by:

$$i = \frac{P}{100}(N)$$

where

P = the percentile of interest
i = percentile location
N = number in the data set

3. Determine the location by either (a) or (b).
   a. If i is a whole number, the Pth percentile is the average of the value at the ith location and the value at the $(i + 1)^{st}$ location.
   b. If i is not a whole number, the Pth percentile value is located at the whole number part of $i + 1$.

For example, suppose you want to determine the 80th percentile of 1240 numbers. P is 80 and N is 1240. First, order the numbers from lowest to highest. Next, calculate the location of the 80th percentile.

$$i = \frac{80}{100}(1240) = 992$$

Because $i = 992$ is a whole number, follow the directions in step 3(a). The 80th percentile is the average of the 992nd number and the 993rd number.

$$P_{80} = \frac{(992\text{nd number} + 993\text{rd number})}{2}$$

DEMONSTRATION PROBLEM 3.2

Determine the 30th percentile of the following eight numbers: 14, 12, 19, 23, 5, 13, 28, 17.

Solution

For these eight numbers, we want to find the value of the 30th percentile, so $N = 8$ and $P = 30$

First, organize the data into an ascending-order array.

5 12 13 14 17 19 23 28

Next, compute the value of *i*.

$$i = \frac{30}{100}(8) = 2.4$$

Because *i* is not a whole number, step 3(b) is used. The value of $i + 1$ is $2.4 + 1$, or 3.4. The whole-number part of 3.4 is 3. The 30th percentile is located at the third value. The third value is 13, so 13 is the 30th percentile. Note that a percentile may or may not be one of the data values.

Quartiles

Quartiles are *measures of central tendency that divide a group of data into four subgroups or parts*. The three quartiles are denoted as Q_1, Q_2, and Q_3. The first quartile, Q_1, separates the first, or lowest, one-fourth of the data from the upper three-fourths and is equal to the 25th percentile. The second quartile, Q_2, separates the second quarter of the data from the third quarter. Q_2 is located at the 50th percentile and equals the median of the data. The third quartile, Q_3, divides the first three-quarters of the data from the last quarter and is equal to the value of the 75th percentile. These three quartiles are shown in Figure 3.2.

Suppose we want to determine the values of Q_1, Q_2, and Q_3 for the following numbers.

106 109 114 116 121 122 125 129

The value of Q_1 is found at the 25th percentile, P_{25}, by:

$$\text{For } N = 8,\ i = \frac{25}{100}(8) = 2$$

Because *i* is a whole number, P_{25} is found as the average of the second and third numbers.

$$P_{25} = \frac{(109 + 114)}{2} = 111.5$$

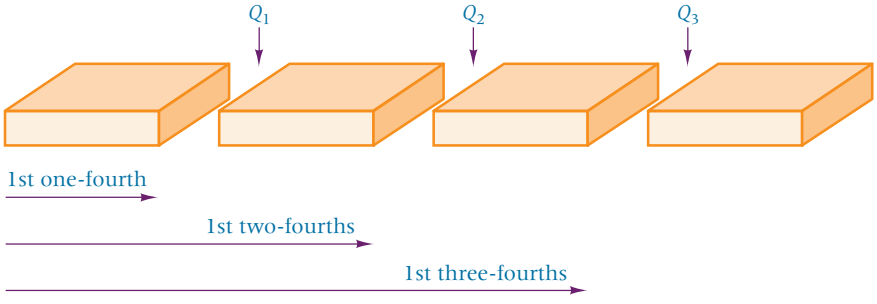
The value of Q_1 is $P_{25} = 111.5$. Notice that one-fourth, or two, of the values (106 and 109) are less than 111.5.

The value of Q_2 is equal to the median. Because the array contains an even number of terms, the median is the average of the two middle terms.

$$Q_2 = \text{median} = \frac{(116 + 121)}{2} = 118.5$$

Notice that exactly half of the terms are less than Q_2 and half are greater than Q_2.

FIGURE 3.2
Quartiles



The value of Q_3 is determined by P_{75} as follows.

$$i = \frac{75}{100}(8) = 6$$

Because i is a whole number, P_{75} is the average of the sixth and the seventh numbers.

$$P_{75} = \frac{(122 + 125)}{2} = 123.5$$

The value of Q_3 is $P_{75} = 123.5$. Notice that three-fourths, or six, of the values are less than 123.5 and two of the values are greater than 123.5.

DEMONSTRATION PROBLEM 3.3

The following shows the top 16 global marketing categories for advertising spending for a recent year according to *Advertising Age*. Spending is given in millions of U.S. dollars. Determine the first, the second, and the third quartiles for these data.

Category	Ad Spending
Automotive	$22,195
Personal Care	19,526
Entertainment & Media	9,538
Food	7,793
Drugs	7,707
Electronics	4,023
Soft Drinks	3,916
Retail	3,576
Cleaners	3,571
Restaurants	3,553
Computers	3,247
Telephone	2,488
Financial	2,433
Beer, Wine & Liquor	2,050
Candy	1,137
Toys	699

Solution

For 16 marketing organizations, $N = 16$. $Q_1 = P_{25}$ is found by

$$i = \frac{25}{100}(16) = 4$$

Because i is a whole number, Q_1 is found to be the average of the fourth and fifth values from the bottom.

$$Q_1 = \frac{2433 + 2488}{2} = 2460.5$$

$Q_2 = P_{50}$ = median; with 16 terms, the median is the average of the eighth and ninth terms.

$$Q_2 = \frac{3571 + 3576}{2} = 3573.5$$

$Q_3 = P_{75}$ is solved by

$$i = \frac{75}{100}(16) = 12$$

Q_3 is found by averaging the 12th and 13th terms.

$$Q_3 = \frac{7707 + 7793}{2} = 7750$$

3.1 PROBLEMS

3.1 Determine the mode for the following numbers.

2 4 8 4 6 2 7 8 4 3 8 9 4 3 5

3.2 Determine the median for the numbers in Problem 3.1.

3.3 Determine the median for the following numbers.

213 345 609 073 167 243 444 524 199 682

3.4 Compute the mean for the following numbers.

17.3 44.5 31.6 40.0 52.8 38.8 30.1 78.5

3.5 Compute the mean for the following numbers.

7 −2 5 9 0 −3 −6 −7 −4 −5 2 −8

3.6 Compute the 35th percentile, the 55th percentile, Q_1, Q_2, and Q_3 for the following data.

16 28 29 13 17 20 11 34 32 27 25 30 19 18 33

3.7 Compute P_{20}, P_{47}, P_{83}, Q_1, Q_2, and Q_3 for the following data.

120	138	97	118	172	144
138	107	94	119	139	145
162	127	112	150	143	80
105	116	142	128	116	171

3.8 The following list shows the 15 largest banks in the world by assets according to EagleTraders.com. Compute the median and the mean assets from this group. Which of these two measures do think is most appropriate for summarizing these data, and why? What is the value of Q_2? Determine the 63rd percentile for the data. Determine the 29th percentile for the data.

Bank	Assets ($ billions)
Deutsche Bank AG (Frankfurt)	842
BNP Paribas SA (Paris)	700
Bank of Tokyo-Mitsubishi Ltd (Tokyo)	700
UBS AG (Zurich)	687
Bank of America NA (Charlotte)	572
The Sumitomo Bank Ltd (Tokyo)	524
Bayerische Hypo-und Vereinsbank AG (Munich)	504
The Norinchukin Bank (Tokyo)	485
The Dai-Ichi Kangyo Bank Ltd (Tokyo)	481
The Sakura Bank Ltd (Tokyo)	473
ABN AMRO Holding NV (Amsterdam)	459
The Fuji Bank Ltd (Tokyo)	458
Credit Agricole (Paris)	441
Industrial & Commercial Bank of China (Beijing)	428
Societe Generale (Paris)	407

3.9 The following lists the 10 largest automakers in the world and the number of vehicles produced by each in a recent year. Compute the median, Q_3, P_{20}, P_{60}, P_{80}, and P_{93} on these data.

Auto Manufacturer	Production (millions)
Toyota Motor Corp.	9.37
General Motors	8.90
Volkswagen AG	6.19
Ford Motor Co.	5.96
Hyundai-Kia Automotive Group	3.96
Honda Motor Co. Ltd.	3.83
Nissan Motor Co.	3.68
PSA/Peugeot-Citreon SA	3.43
Chrysler LLC	2.68
Fiat S.p.A.	2.62

3.10 The following lists the number of fatal accidents by scheduled commercial airlines over a 17-year period according to the Air Transport Association of America. Using these data, compute the mean, median, and mode. What is the value of the third quartile? Determine P_{11}, P_{35}, P_{58}, and P_{67}.

4 4 4 1 4 2 4 3 8 6 4 4 1 4 2 3 3

3.2 MEASURES OF VARIABILITY: UNGROUPED DATA

Measures of central tendency yield information about the center or middle part of a data set. However, business researchers can use another group of analytic tools, **measures of variability**, to *describe the spread or the dispersion of a set of data.* Using measures of variability in conjunction with measures of central tendency makes possible a more complete numerical description of the data.

For example, a company has 25 salespeople in the field, and the median annual sales figure for these people is $1.2 million. Are the salespeople being successful as a group or not? The median provides information about the sales of the person in the middle, but what about the other salespeople? Are all of them selling $1.2 million annually, or do the sales figures vary widely, with one person selling $5 million annually and another selling only $150,000 annually? Measures of variability provide the additional information necessary to answer that question.

Figure 3.3 shows three distributions in which the mean of each distribution is the same ($\mu = 50$) but the variabilities differ. Observation of these distributions shows that a measure of variability is necessary to complement the mean value in describing the data. Methods of computing measures of variability differ for ungrouped data and grouped data. This section focuses on seven measures of variability for ungrouped data: range, interquartile range, mean absolute deviation, variance, standard deviation, z scores, and coefficient of variation.

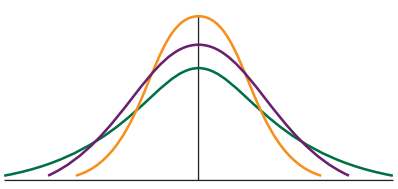
Range

The **range** is *the difference between the largest value of a data set and the smallest value of a set.* Although it is usually a single numeric value, some business researchers define the range of data as the ordered pair of smallest and largest numbers (smallest, largest). It is a crude measure of variability, describing the distance to the outer bounds of the data set. It reflects those extreme values because it is constructed from them. An advantage of the range is its ease of computation. One important use of the range is in quality assurance, where the range is used to construct control charts. A disadvantage of the range is that, because it is computed with the values that are on the extremes of the data, it is affected by extreme values, and its application as a measure of variability is limited.

The data in Table 3.1 represent the offer prices for the 20 largest U.S. initial public offerings in a recent year. The lowest offer price was $7.00 and the highest price was $43.25. The range of the offer prices can be computed as the difference of the highest and lowest values:

$$\text{Range} = \text{Highest} - \text{Lowest} = \$43.25 - \$7.00 = \$36.25$$

FIGURE 3.3
Three Distributions with the Same Mean but Different Dispersions



Interquartile Range

Another measure of variability is the **interquartile range**. The interquartile range is *the range of values between the first and third quartile*. Essentially, it is the range of the middle 50% of the data and is determined by computing the value of $Q_3 - Q_1$. The interquartile range is especially useful in situations where data users are more interested in values toward the middle and less interested in extremes. In describing a real estate housing market, Realtors might use the interquartile range as a measure of housing prices when describing the middle half of the market for buyers who are interested in houses in the midrange. In addition, the interquartile range is used in the construction of box-and-whisker plots.

INTERQUARTILE RANGE	$Q_3 - Q_1$

The following data indicate the top 15 trading partners of the United States in exports in a recent year according to the U.S. Census Bureau.

Country	Exports ($ billions)
Canada	213.1
Mexico	119.4
China	61.0
Japan	58.1
United Kingdom	45.4
Germany	44.3
South Korea	33.0
Netherlands	30.5
France	25.8
Taiwan	24.8
Singapore	23.6
Belgium	23.0
Brazil	21.7
Australia	17.9
India	16.3

What is the interquartile range for these data? The process begins by computing the first and third quartiles as follows.

Solving for $Q_1 = P_{25}$ when $N = 15$:

$$i = \frac{25}{100}(15) = 3.75$$

Because i is not a whole number, P_{25} is found as the fourth term from the bottom.

$$Q_1 = P_{25} = 23.0$$

Solving for $Q_3 = P_{75}$:

$$i = \frac{75}{100}(15) = 11.25$$

Because i is not a whole number, P_{75} is found as the 12th term from the bottom.

$$Q_3 = P_{75} = 58.1$$

The interquartile range is:

$$Q_3 - Q_1 = 58.1 - 23.0 = 35.1$$

The middle 50% of the exports for the top 15 U.S. trading partners spans a range of 35.1 ($ billions).

STATISTICS IN BUSINESS TODAY

Recycling Statistics

There are many interesting statistics with regard to recycling. Recycling one aluminum can saves enough energy, the equivalent of a half gallon of gasoline, to run a television for three hours. Because Americans have done such a good job of recycling aluminum cans, they account for less than 1% of the total U.S. waste stream. Recycling 1 pound of steel saves enough energy to run a 60-watt light bulb for over a day. On average, one American uses seven trees a year in paper, wood, and other products made from trees. In addition, Americans use 680 pounds of paper per year. Each ton of recycled paper saves about 17 trees, 380 gallons of oil, three cubic yards of landfill space, 4000 kilowatts of energy, and 7000 gallons of water. Americans use 2.5 million plastic bottles every hour and throw way 25 billion Styrofoam cups every year. The energy saved from recycling one glass bottle can run a 100-watt light bulb for four hours. Every year, each American throws out about 1200 pounds of organic garbage that could be composted. The U.S. is number one in the world in producing trash, with an average of 1609 pounds per person per year.

Sources: http://www.recycling-revolution.com/recycling-facts.html, National Recycling Coalition, the Environmental Protection Agency, Earth911.org

Mean Absolute Deviation, Variance, and Standard Deviation

Three other measures of variability are the variance, the standard deviation, and the mean absolute deviation. They are obtained through similar processes and are, therefore, presented together. These measures are not meaningful unless the data are at least interval-level data. The variance and standard deviation are widely used in statistics. Although the standard deviation has some stand-alone potential, the importance of variance and standard deviation lies mainly in their role as tools used in conjunction with other statistical devices.

Suppose a small company started a production line to build computers. During the first five weeks of production, the output is 5, 9, 16, 17, and 18 computers, respectively. Which descriptive statistics could the owner use to measure the early progress of production? In an attempt to summarize these figures, the owner could compute a mean.

$$\begin{array}{r} x \\ \hline 5 \\ 9 \\ 16 \\ 17 \\ 18 \end{array}$$

$$\Sigma x = 65 \qquad \mu = \frac{\Sigma x}{N} = \frac{65}{5} = 13$$

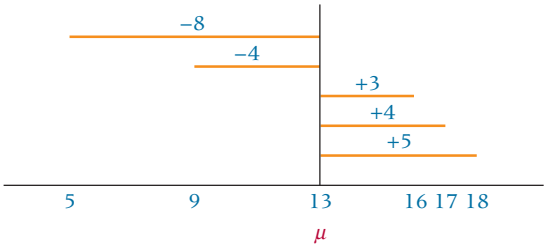
What is the variability in these five weeks of data? One way for the owner to begin to look at the spread of the data is to subtract the mean from each data value. *Subtracting the mean from each value of data* yields the **deviation from the mean** $(x - \mu)$. Table 3.2 shows these deviations for the computer company production. Note that some deviations from

TABLE 3.2
Deviations from the Mean for Computer Production

Number (x)	Deviations from the Mean ($x - \mu$)
5	$5 - 13 = -8$
9	$9 - 13 = -4$
16	$16 - 13 = +3$
17	$17 - 13 = +4$
18	$18 - 13 = +5$
$\Sigma x = 65$	$\Sigma(x - \mu) = 0$

FIGURE 3.4

Geometric Distances from the Mean (from Table 3.2)



the mean are positive and some are negative. Figure 3.4 shows that geometrically the negative deviations represent values that are below (to the left of) the mean and positive deviations represent values that are above (to the right of) the mean.

An examination of deviations from the mean can reveal information about the variability of data. However, the deviations are used mostly as a tool to compute other measures of variability. Note that in both Table 3.2 and Figure 3.4 these deviations total zero. This phenomenon applies to all cases. For a given set of data, the sum of all deviations from the arithmetic mean is always zero.

SUM OF DEVIATIONS FROM THE ARITHMETIC MEAN IS ALWAYS ZERO

$$\Sigma(x - \mu) = 0$$

This property requires considering alternative ways to obtain measures of variability.

One obvious way to force the sum of deviations to have a nonzero total is to take the absolute value of each deviation around the mean. Utilizing the absolute value of the deviations about the mean makes solving for the mean absolute deviation possible.

Mean Absolute Deviation

The **mean absolute deviation (MAD)** is *the average of the absolute values of the deviations around the mean for a set of numbers.*

MEAN ABSOLUTE DEVIATION

$$\text{MAD} = \frac{\Sigma|x - \mu|}{N}$$

Using the data from Table 3.2, the computer company owner can compute a mean absolute deviation by taking the absolute values of the deviations and averaging them, as shown in Table 3.3. The mean absolute deviation for the computer production data is 4.8.

TABLE 3.3

MAD for Computer Production Data

x	$x - \mu$	$\lvert x - \mu \rvert$
5	−8	+8
9	−4	+4
16	+3	+3
17	+4	+4
18	+5	+5
$\Sigma x = 65$	$\Sigma(x - \mu) = 0$	$\Sigma\lvert x - \mu \rvert = 24$

$$\text{MAD} = \frac{\Sigma|x - \mu|}{N} = \frac{24}{5} = 4.8$$

Because it is computed by using absolute values, the mean absolute deviation is less useful in statistics than other measures of dispersion. However, in the field of forecasting, it is used occasionally as a measure of error.

Variance

Because absolute values are not conducive to easy manipulation, mathematicians developed an alternative mechanism for overcoming the zero-sum property of deviations from the mean. This approach utilizes the square of the deviations from the mean. The result is the variance, an important measure of variability.

The **variance** is *the average of the squared deviations about the arithmetic mean for a set of numbers.* The population variance is denoted by σ^2.

POPULATION VARIANCE

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

Table 3.4 shows the original production numbers for the computer company, the deviations from the mean, and the squared deviations from the mean.

The sum of the squared deviations about the mean of a set of values—called the **sum of squares of *x*** and sometimes abbreviated as SS_x—is used throughout statistics. For the computer company, this value is 130. Dividing it by the number of data values (5 weeks) yields the variance for computer production.

$$\sigma^2 = \frac{130}{5} = 26.0$$

Because the variance is computed from squared deviations, the final result is expressed in terms of squared units of measurement. Statistics measured in squared units are problematic to interpret. Consider, for example, Mattel Toys attempting to interpret production costs in terms of squared dollars or Troy-Bilt measuring production output variation in terms of squared lawn mowers. Therefore, when used as a descriptive measure, variance can be considered as an intermediate calculation in the process of obtaining the standard deviation.

TABLE 3.4
Computing a Variance and a Standard Deviation from the Computer Production Data

x	$x - \mu$	$(x - \mu)^2$
5	−8	64
9	−4	16
16	+3	9
17	+4	16
18	+5	25
$\Sigma x = 65$	$\Sigma(x-\mu) = 0$	$\Sigma(x-\mu)^2 = 130$

$$SS_x = \Sigma(x - \mu)^2 = 130$$

$$\text{Variance} = \sigma^2 = \frac{SS_x}{N} = \frac{\Sigma(x - \mu)^2}{N} = \frac{130}{5} = 26.0$$

$$\text{Standard Deviation} = \sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}} = \sqrt{\frac{130}{5}} = 5.1$$

Standard Deviation

The standard deviation is a popular measure of variability. It is used both as a separate entity and as a part of other analyses, such as computing confidence intervals and in hypothesis testing (see Chapters 8, 9, and 10).

POPULATION STANDARD DEVIATION

$$\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$$

The **standard deviation** is *the square root of the variance*. The population standard deviation is denoted by σ.

Like the variance, the standard deviation utilizes the sum of the squared deviations about the mean (SS_x). It is computed by averaging these squared deviations (SS_x/N) and taking the square root of that average. One feature of the standard deviation that distinguishes it from a variance is that the standard deviation is expressed in the same units as the raw data, whereas the variance is expressed in those units squared. Table 3.4 shows the standard deviation for the computer production company: $\sqrt{26}$, or 5.1.

What does a standard deviation of 5.1 mean? The meaning of standard deviation is more readily understood from its use, which is explored in the next section. Although the standard deviation and the variance are closely related and can be computed from each other, differentiating between them is important, because both are widely used in statistics.

Meaning of Standard Deviation

What is a standard deviation? What does it do, and what does it mean? The most precise way to define standard deviation is by reciting the formula used to compute it. However, insight into the concept of standard deviation can be gleaned by viewing the manner in which it is applied. Two ways of applying the standard deviation are the **empirical rule** and **Chebyshev's theorem**.

Empirical Rule

The empirical rule is an important rule of thumb that *is used to state the approximate percentage of values that lie within a given number of standard deviations from the mean of a set of data if the data are normally distributed.*

The empirical rule is used only for three numbers of standard deviations: lσ, 2σ, and 3σ. More detailed analysis of other numbers of σ values is presented in Chapter 6. Also discussed in further detail in Chapter 6 is the normal distribution, a unimodal, symmetrical distribution that is bell (or mound) shaped. The requirement that the data be normally distributed contains some tolerance, and the empirical rule generally applies as long as the data are approximately mound shaped.

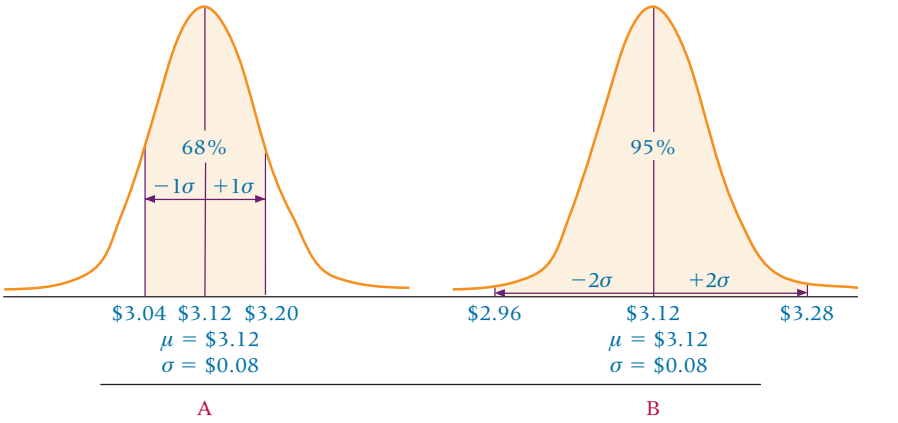
EMPIRICAL RULE*

Distance from the Mean	Values Within Distance
$\mu \pm 1\sigma$	68%
$\mu \pm 2\sigma$	95%
$\mu \pm 3\sigma$	99.7%

*Based on the assumption that the data are approximately normally distributed.

FIGURE 3.5

Empirical Rule for One and Two Standard Deviations of Gasoline Prices



If a set of data is normally distributed, or bell shaped, approximately 68% of the data values are within one standard deviation of the mean, 95% are within two standard deviations, and almost 100% are within three standard deviations.

Suppose a recent report states that for California, the average statewide price of a gallon of regular gasoline is $3.12. Suppose regular gasoline prices vary across the state with a standard deviation of $0.08 and are normally distributed. According to the empirical rule, approximately 68% of the prices should fall within $\mu \pm 1\sigma$, or $3.12 ± 1 ($0.08). Approximately 68% of the prices should be between $3.04 and $3.20, as shown in Figure 3.5A. Approximately 95% should fall within $\mu \pm 2\sigma$ or $3.12 ± 2 ($0.08) = $3.12 ± $0.16, or between $2.96 and $3.28, as shown in Figure 3.5B. Nearly all regular gasoline prices (99.7%) should fall between $2.88 and $3.36 ($\mu \pm 3\sigma$).

Note that with 68% of the gasoline prices falling within one standard deviation of the mean, approximately 32% are outside this range. Because the normal distribution is symmetrical, the 32% can be split in half such that 16% lie in each tail of the distribution. Thus, approximately 16% of the gasoline prices should be less than $3.04 and approximately 16% of the prices should be greater than $3.20.

Many phenomena are distributed approximately in a bell shape, including most human characteristics such as height and weight; therefore the empirical rule applies in many situations and is widely used.

DEMONSTRATION PROBLEM 3.4

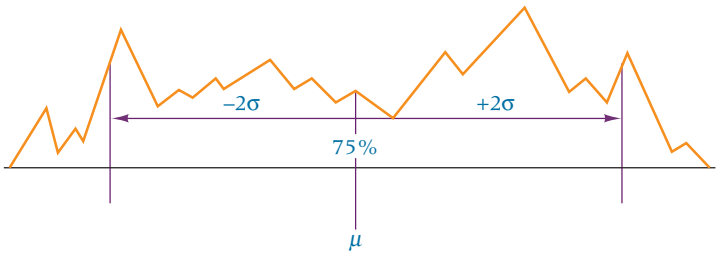
A company produces a lightweight valve that is specified to weigh 1365 grams. Unfortunately, because of imperfections in the manufacturing process not all of the valves produced weigh exactly 1365 grams. In fact, the weights of the valves produced are normally distributed with a mean weight of 1365 grams and a standard deviation of 294 grams. Within what range of weights would approximately 95% of the valve weights fall? Approximately 16% of the weights would be more than what value? Approximately 0.15% of the weights would be less than what value?

Solution

Because the valve weights are normally distributed, the empirical rule applies. According to the empirical rule, approximately 95% of the weights should fall within $\mu \pm 2\sigma = 1365 \pm 2(294) = 1365 \pm 588$. Thus, approximately 95% should fall between 777 and 1953. Approximately 68% of the weights should fall within $\mu \pm 1\sigma$, and 32% should fall outside this interval. Because the normal distribution is symmetrical, approximately 16% should lie above $\mu + 1\sigma = 1365 + 294 = 1659$. Approximately 99.7% of the weights should fall within $\mu \pm 3\sigma$, and .3% should fall outside this interval. Half of these, or .15%, should lie below $\mu - 3\sigma = 1365 - 3(294) = 1365 - 882 = 483$.

FIGURE 3.6

Application of Chebyshev's Theorem for Two Standard Deviations



Chebyshev's Theorem

The empirical rule applies only when data are known to be approximately normally distributed. What do researchers use when data are not normally distributed or when the shape of the distribution is unknown? Chebyshev's theorem applies to all distributions regardless of their shape and thus can be used whenever the data distribution shape is unknown or is nonnormal. Even though Chebyshev's theorem can in theory be applied to data that are normally distributed, the empirical rule is more widely known and is preferred whenever appropriate. Chebyshev's theorem is not a rule of thumb, as is the empirical rule, but rather it is presented in formula format and therefore can be more widely applied. Chebyshev's theorem states that *at least* $1-1/k^2$ *values will fall within* $\pm$ *k standard deviations of the mean regardless of the shape of the distribution.*

CHEBYSHEV'S THEOREM

Within k standard deviations of the mean, $\mu \pm k\sigma$, lie at least

$$1 - \frac{1}{k^2}$$

proportion of the values.
Assumption: $k > 1$

Specifically, Chebyshev's theorem says that at least 75% of all values are within $\pm 2\sigma$ of the mean regardless of the shape of a distribution because if $k = 2$, then $1 - 1/k^2 = 1 - 1/2^2 = 3/4 = .75$. Figure 3.6 provides a graphic illustration. In contrast, the empirical rule states that if the data are normally distributed 95% of all values are within $\mu \pm 2\sigma$. According to Chebyshev's theorem, the percentage of values within three standard deviations of the mean is at least 89%, in contrast to 99.7% for the empirical rule. Because a formula is used to compute proportions with Chebyshev's theorem, any value of k greater than 1 ($k > 1$) can be used. For example, if $k = 2.5$, at least .84 of all values are within $\mu \pm 2.5\sigma$, because $1 - 1/k^2 = 1 - 1/(2.5)^2 = .84$.

DEMONSTRATION PROBLEM 3.5

In the computing industry the average age of professional employees tends to be younger than in many other business professions. Suppose the average age of a professional employed by a particular computer firm is 28 with a standard deviation of 6 years. A histogram of professional employee ages with this firm reveals that the data are not normally distributed but rather are amassed in the 20s and that few workers are over 40. Apply Chebyshev's theorem to determine within what range of ages would at least 80% of the workers' ages fall.

Solution

Because the ages are not normally distributed, it is not appropriate to apply the empirical rule; and therefore Chebyshev's theorem must be applied to answer the question.

Chebyshev's theorem states that at least $1 - 1/k^2$ proportion of the values are within $\mu \pm k\sigma$. Because 80% of the values are within this range, let

$$1 - \frac{1}{k^2} = .80$$

Solving for k yields

$$.20 = \frac{1}{k^2}$$
$$k^2 = 5.000$$
$$k = 2.24$$

Chebyshev's theorem says that at least .80 of the values are within $\pm$ 2.24 of the mean.

For $\mu = 28$ and $\sigma = 6$, at least .80, or 80%, of the values are within $28 \pm 2.24(6) = 28 \pm 13.4$ years of age or between 14.6 and 41.4 years old.

Population Versus Sample Variance and Standard Deviation

The sample variance is denoted by s^2 and the sample standard deviation by s. The main use for sample variances and standard deviations is as estimators of population variances and standard deviations. Because of this, computation of the sample variance and standard deviation differs slightly from computation of the population variance and standard deviation. Both the sample variance and sample standard deviation use $n - 1$ in the denominator instead of n because using n in the denominator of a sample variance results in a statistic that tends to underestimate the population variance. While discussion of the properties of *good estimators* is beyond the scope of this text, one of the properties of a good estimator is being *unbiased*. Whereas using n in the denominator of the sample variance makes it a *biased* estimator, using $n - 1$ allows it to be an *unbiased* estimator, which is a desirable property in inferential statistics.

SAMPLE VARIANCE

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

SAMPLE STANDARD DEVIATION

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

Shown here is a sample of six of the largest accounting firms in the United States and the number of partners associated with each firm as reported by the *Public Accounting Report.*

Firm	Number of Partners
Deloitte & Touche	2654
Ernst & Young	2108
PricewaterhouseCoopers	2069
KPMG	1664
RSM McGladrey	720
Grant Thornton	309

The sample variance and sample standard deviation can be computed by:

x	$(x - \bar{x})^2$
2654	1,137,784.89
2108	271,097.25
2069	232,005.99
1664	5,878.29
720	752,261.33
309	1,634,127.59
$\Sigma x = 9524$	$\Sigma(x - \bar{x})^2 = 4{,}033{,}155.34$

$$\bar{x} = \frac{9524}{6} = 1587.33$$

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{4{,}033{,}155.34}{5} = 806{,}631.07$$

$$s = \sqrt{s^2} = \sqrt{806{,}631.07} = 898.13$$

The sample variance is 806,631.07, and the sample standard deviation is 898.13.

Computational Formulas for Variance and Standard Deviation

An alternative method of computing variance and standard deviation, sometimes referred to as the computational method or shortcut method, is available. Algebraically,

$$\Sigma(x - \mu)^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{N}$$

and

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

Substituting these equivalent expressions into the original formulas for variance and standard deviation yields the following computational formulas.

COMPUTATIONAL FORMULA FOR POPULATION VARIANCE AND STANDARD DEVIATION

$$\sigma^2 = \frac{\Sigma x^2 - \dfrac{(\Sigma x)^2}{N}}{N}$$

$$\sigma = \sqrt{\sigma^2}$$

COMPUTATIONAL FORMULA FOR SAMPLE VARIANCE AND STANDARD DEVIATION

$$s^2 = \frac{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}{n - 1}$$

$$s = \sqrt{s^2}$$

These computational formulas utilize the sum of the x values and the sum of the x^2 values instead of the difference between the mean and each value and computed deviations. In the precalculator/computer era, this method usually was faster and easier than using the original formulas.

TABLE 3.5
Computational Formula Calculations of Variance and Standard Deviation for Computer Production Data

x	x^2
5	25
9	81
16	256
17	289
18	324
$\Sigma x = 65$	$\Sigma x^2 = 975$

$$\sigma^2 = \frac{975 - \frac{(65)^2}{5}}{5} = \frac{975 - 845}{5} = \frac{130}{5} = 26$$

$$\sigma = \sqrt{26} = 5.1$$

For situations in which the mean is already computed or is given, alternative forms of these formulas are

$$\sigma^2 = \frac{\Sigma x^2 - N\mu^2}{N}$$

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

Using the computational method, the owner of the start-up computer production company can compute a population variance and standard deviation for the production data, as shown in Table 3.5. (Compare these results with those in Table 3.4.)

DEMONSTRATION PROBLEM 3.6

The effectiveness of district attorneys can be measured by several variables, including the number of convictions per month, the number of cases handled per month, and the total number of years of conviction per month. A researcher uses a sample of five district attorneys in a city and determines the total number of years of conviction that each attorney won against defendants during the past month, as reported in the first column in the following tabulations. Compute the mean absolute deviation, the variance, and the standard deviation for these figures.

Solution

The researcher computes the mean absolute deviation, the variance, and the standard deviation for these data in the following manner.

x	$\lvert x - \bar{x} \rvert$	$(x - \bar{x})^2$
55	41	1,681
100	4	16
125	29	841
140	44	1,936
60	36	1,296
$\Sigma x = 480$	$\Sigma \lvert x - \bar{x} \rvert = 154$	$\Sigma (x - \bar{x})^2 = 5{,}770$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{480}{5} = 96$$

$$MAD = \frac{154}{5} = 30.8$$

$$s^2 = \frac{5{,}770}{4} = 1{,}442.5 \text{ and } s = \sqrt{s^2} = 37.98$$

She then uses computational formulas to solve for s^2 and s and compares the results.

x	x^2
55	3,025
100	10,000
125	15,625
140	19,600
60	3,600
$\Sigma x = 480$	$\Sigma x^2 = 51{,}850$

$$s^2 = \frac{51{,}850 - \frac{(480)^2}{5}}{4} = \frac{51{,}850 - 46{,}080}{4} = \frac{5{,}770}{4} = 1{,}442.5$$

$$s = \sqrt{1{,}442.5} = 37.98$$

The results are the same. The sample standard deviation obtained by both methods is 37.98, or 38, years.

z Scores

A ***z* score** represents the number of standard deviations a value (x) is above or below the mean of a set of numbers when the data are normally distributed. Using z scores allows translation of a value's raw distance from the mean into units of standard deviations.

***z* SCORE**

$$z = \frac{x - \mu}{\sigma}$$

For samples,

$$z = \frac{x - \bar{x}}{s}$$

If a z score is negative, the raw value (x) is below the mean. If the z score is positive, the raw value (x) is above the mean.

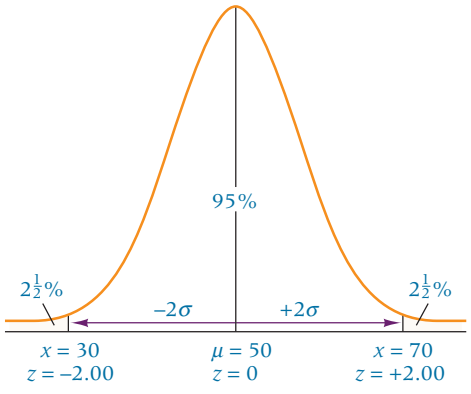
For example, for a data set that is normally distributed with a mean of 50 and a standard deviation of 10, suppose a statistician wants to determine the z score for a value of 70. This value ($x = 70$) is 20 units above the mean, so the z value is

$$z = \frac{70 - 50}{10} = +2.00$$

This z score signifies that the raw score of 70 is two standard deviations above the mean. How is this z score interpreted? The empirical rule states that 95% of all values are within two standard deviations of the mean if the data are approximately normally distributed. Figure 3.7 shows that because the value of 70 is two standard deviations above the mean ($z = +2.00$), 95% of the values are between 70 and the value ($x = 30$), that is two standard

FIGURE 3.7

Percentage Breakdown of Scores Two Standard Deviations from the Mean



deviations below the mean, or $z = (30 - 50)/10 = -2.00$. Because 5% of the values are outside the range of two standard deviations from the mean and the normal distribution is symmetrical, 2½% ($\frac{1}{2}$ of the 5%) are below the value of 30. Thus 97½% of the values are below the value of 70. Because a z score is the number of standard deviations an individual data value is from the mean, the empirical rule can be restated in terms of z scores.

Between $z = -1.00$ and $z = +1.00$ are approximately 68% of the values.
Between $z = -2.00$ and $z = +2.00$ are approximately 95% of the values.
Between $z = -3.00$ and $z = +3.00$ are approximately 99.7% of the values.

The topic of z scores is discussed more extensively in Chapter 6.

Coefficient of Variation

The **coefficient of variation** is a statistic that *is the ratio of the standard deviation to the mean expressed in percentage* and is denoted CV.

COEFFICIENT OF VARIATION

$$\text{CV} = \frac{\sigma}{\mu}(100)$$

The coefficient of variation essentially is a relative comparison of a standard deviation to its mean. The coefficient of variation can be useful in comparing standard deviations that have been computed from data with different means.

Suppose five weeks of average prices for stock A are 57, 68, 64, 71, and 62. To compute a coefficient of variation for these prices, first determine the mean and standard deviation: $\mu = 64.40$ and $\sigma = 4.84$. The coefficient of variation is:

$$\text{CV}_\text{A} = \frac{\sigma_A}{\mu_A}(100) = \frac{4.84}{64.40}(100) = .075 = 7.5\%$$

The standard deviation is 7.5% of the mean.

Sometimes financial investors use the coefficient of variation or the standard deviation or both as measures of risk. Imagine a stock with a price that never changes. An investor bears no risk of losing money from the price going down because no variability occurs in the price. Suppose, in contrast, that the price of the stock fluctuates wildly. An investor who buys at a low price and sells for a high price can make a nice profit. However, if the price drops below what the investor buys it for, the stock owner is subject to a potential loss. The greater the variability is, the more the potential for loss. Hence, investors use measures of variability such as standard deviation or coefficient of variation to determine the risk of a stock. What does the coefficient of variation tell us about the risk of a stock that the standard deviation does not?

Suppose the average prices for a second stock, B, over these same five weeks are 12, 17, 8, 15, and 13. The mean for stock B is 13.00 with a standard deviation of 3.03. The coefficient of variation can be computed for stock B as:

$$\text{CV}_\text{B} = \frac{\sigma_B}{\mu_B}(100) = \frac{3.03}{13}(100) = .233 = 23.3\%$$

The standard deviation for stock B is 23.3% of the mean.

With the standard deviation as the measure of risk, stock A is more risky over this period of time because it has a larger standard deviation. However, the average price of stock A is almost five times as much as that of stock B. Relative to the amount invested in stock A, the standard deviation of $4.84 may not represent as much risk as the standard deviation of $3.03 for stock B, which has an average price of only $13.00. The coefficient of variation reveals the risk of a stock in terms of the size of standard deviation relative to the size of the mean (in percentage). Stock B has a coefficient of variation that is nearly three times as much as the coefficient of variation for stock A. Using coefficient of variation as a measure of risk indicates that stock B is riskier.

The choice of whether to use a coefficient of variation or raw standard deviations to compare multiple standard deviations is a matter of preference. The coefficient of variation also provides an optional method of interpreting the value of a standard deviation.

STATISTICS IN BUSINESS TODAY

Business Travel

Findings from the Bureau of Transportation Statistics' National Household Travel Survey revealed that more than 405 million long-distance business trips are taken each year in the United States. Over 80% of these business trips are taken by personal vehicle. Almost three out of four business trips are for less than 250 miles, and only about 7% are for more than 1000 miles. The mean one-way distance for a business trip in the United States is 123 miles. Air travel accounts for 16% of all business travel. The average per diem cost of business travel to New York City is about $450, to Beijing is about $282, to Moscow is about $376, and to Paris is about $305. Seventy-seven percent of all business travelers are men, and 55% of business trips are taken by people in the 30-to-49-year-old age bracket. Forty-five percent of business trips are taken by people who have a household income of more than $75,000.

Sources: U.S. Department of Transportation site at http://www.dot.gov/affairs/bts2503.htm and Expansion Management.com site at http://www.expansionmanagement.com/cmd/articledetail/articleid/15602/default.asp

3.2 PROBLEMS

3.11 A data set contains the following seven values.

6 2 4 9 1 3 5

a. Find the range.
b. Find the mean absolute deviation.
c. Find the population variance.
d. Find the population standard deviation.
e. Find the interquartile range.
f. Find the *z* score for each value.

3.12 A data set contains the following eight values.

4 3 0 5 2 9 4 5

a. Find the range.
b. Find the mean absolute deviation.
c. Find the sample variance.
d. Find the sample standard deviation.
e. Find the interquartile range.

3.13 A data set contains the following six values.

12 23 19 26 24 23

a. Find the population standard deviation using the formula containing the mean (the original formula).
b. Find the population standard deviation using the computational formula.
c. Compare the results. Which formula was faster to use? Which formula do you prefer? Why do you think the computational formula is sometimes referred to as the "shortcut" formula?

3.14 Use your calculator or computer to find the sample variance and sample standard deviation for the following data.

57	88	68	43	93
63	51	37	77	83
66	60	38	52	28
34	52	60	57	29
92	37	38	17	67

3.15 Use your calculator or computer to find the population variance and population standard deviation for the following data.

123	090	546	378
392	280	179	601
572	953	749	075
303	468	531	646

3.16 Determine the interquartile range on the following data.

44	18	39	40	59
46	59	37	15	73
23	19	90	58	35
82	14	38	27	24
71	25	39	84	70

3.17 According to Chebyshev's theorem, at least what proportion of the data will be within $\mu \pm k\sigma$ for each value of k?

a. $k = 2$

b. $k = 2.5$

c. $k = 1.6$

d. $k = 3.2$

3.18 Compare the variability of the following two sets of data by using both the population standard deviation and the population coefficient of variation.

Data Set 1	Data Set 2
49	159
82	121
77	138
54	152

3.19 A sample of 12 small accounting firms reveals the following numbers of professionals per office.

7	10	9	14	11	8
5	12	8	3	13	6

a. Determine the mean absolute deviation.

b. Determine the variance.

c. Determine the standard deviation.

d. Determine the interquartile range.

e. What is the z score for the firm that has six professionals?

f. What is the coefficient of variation for this sample?

3.20 Shown below are the top food and drug stores in the United States in a recent year according to *Fortune* magazine.

Company	Revenues ($ billions)
Kroger	66.11
Walgreen	47.41
CVS/Caremark	43.81
Safeway	40.19
Publix Super Markets	21.82
Supervalu	19.86
Rite Aid	17.27
Winn-Dixie Stores	7.88

Assume that the data represent a population.

a. Find the range.

b. Find the mean absolute deviation.

c. Find the population variance.

d. Find the population standard deviation.

e. Find the interquartile range.

f. Find the z score for Walgreen.

g. Find the coefficient of variation.

3.21 A distribution of numbers is approximately bell shaped. If the mean of the numbers is 125 and the standard deviation is 12, between what two numbers would approximately 68% of the values fall? Between what two numbers would 95% of the values fall? Between what two values would 99.7% of the values fall?

3.22 Some numbers are not normally distributed. If the mean of the numbers is 38 and the standard deviation is 6, what proportion of values would fall between 26 and 50? What proportion of values would fall between 14 and 62? Between what two values would 89% of the values fall?

3.23 According to Chebyshev's theorem, how many standard deviations from the mean would include at least 80% of the values?

3.24 The time needed to assemble a particular piece of furniture with experience is normally distributed with a mean time of 43 minutes. If 68% of the assembly times are between 40 and 46 minutes, what is the value of the standard deviation? Suppose 99.7% of the assembly times are between 35 and 51 minutes and the mean is still 43 minutes. What would the value of the standard deviation be now? Suppose the time needed to assemble another piece of furniture is not normally distributed and that the mean assembly time is 28 minutes. What is the value of the standard deviation if at least 77% of the assembly times are between 24 and 32 minutes?

3.25 Environmentalists are concerned about emissions of sulfur dioxide into the air. The average number of days per year in which sulfur dioxide levels exceed 150 milligrams per cubic meter in Milan, Italy, is 29. The number of days per year in which emission limits are exceeded is normally distributed with a standard deviation of 4.0 days. What percentage of the years would average between 21 and 37 days of excess emissions of sulfur dioxide? What percentage of the years would exceed 37 days? What percentage of the years would exceed 41 days? In what percentage of the years would there be fewer than 25 days with excess sulfur dioxide emissions?

3.26 Shown below are the per diem business travel expenses listed by Runzheimer International for 11 selected cities around the world. Use this list to calculate the z scores for Moscow, Beijing, Rio de Janeiro, and London. Treat the list as a sample.

City	Per Diem Expense ($)
Beijing	282
Hong Kong	361
London	430
Los Angeles	259
Mexico City	302
Moscow	376
New York (Manhattan)	457
Paris	305
Rio de Janeiro	343
Rome	297
Sydney	188

3.3 MEASURES OF CENTRAL TENDENCY AND VARIABILITY: GROUPED DATA

Grouped data do not provide information about individual values. Hence, measures of central tendency and variability for grouped data must be computed differently from those for ungrouped or raw data.

Measures of Central Tendency

Three measures of central tendency are presented here for grouped data: the mean, the median, and the mode.

Mean

For ungrouped data, the mean is computed by summing the data values and dividing by the number of values. With grouped data, the specific values are unknown. What can be used to represent the data values? The midpoint of each class interval is used to represent all the values in a class interval. This midpoint is weighted by the frequency of values in that class interval. The mean for grouped data is then computed by summing the products of the class midpoint and the class frequency for each class and dividing that sum by the total number of frequencies. The formula for the mean of grouped data follows.

MEAN OF GROUPED DATA

$$\mu_{\text{grouped}} = \frac{\Sigma fM}{N} = \frac{\Sigma fM}{\Sigma f} = \frac{f_1M_1 + f_2M_2 + \cdots + f_iM_i}{f_1 + f_2 + \cdots + f_i}$$

where

i = the number of classes
f = class frequency
N = total frequencies

Table 3.6 gives the frequency distribution of the unemployment rates of Canada from Table 2.2. To find the mean of these data, we need Σf and ΣfM. The value of Σf can be determined by summing the values in the frequency column. To calculate ΣfM, we must first determine the values of M, or the class midpoints. Next we multiply each of these class midpoints by the frequency in that class interval, f, resulting in fM. Summing these values of fM yields the value of ΣfM.

Table 3.7 contains the calculations needed to determine the group mean. The group mean for the unemployment data is 6.93. Remember that because each class interval was represented by its class midpoint rather than by actual values, the group mean is only approximate.

Median

The median for ungrouped or raw data is the middle value of an ordered array of numbers. For grouped data, solving for the median is considerably more complicated. The calculation of the median for grouped data is done by using the following formula.

MEDIAN OF GROUPED DATA

$$\text{Median} = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W)$$

where:

L = the lower limit of the median class interval
cf_p = a cumulative total of the frequencies up to but not including the frequency of the median class
f_{med} = the frequency of the median class
W = the width of the median class interval
N = total number of frequencies

The first step in calculating a grouped median is to determine the value of $N/2$, which is the location of the median term. Suppose we want to calculate the median for the frequency distribution data in Table 3.6. Since there are 60 values (N), the value of $N/2$ is $60/2 = 30$. The median is the 30th term. The question to ask is where does the 30th term fall? This can be answered by determining the cumulative frequencies for the data, as shown in Table 3.6.

An examination of these cumulative frequencies reveals that the 30th term falls in the fourth class interval because there are only 29 values in the first three class intervals. Thus, the median value is in the fourth class interval somewhere between 7 and 9. The class interval containing the median value is referred to as the *median class interval.*

Since the 30th value is between 7 and 9, the value of the median must be at least 7. How much more than 7 is the median? The difference between the location of the median value, $N/2 = 30$, and the cumulative frequencies up to but not including the median class interval, $cf_p = 29$, tells how many values into the median class interval lies the value of the median. This is determined by solving for $N/2 - cf_p = 30 - 29 = 1$. The median value is located one

TABLE 3.6
Frequency Distribution of 60 Years of Unemployment Data for Canada (Grouped Data)

Class Interval	Frequency	Cumulative Frequency
1–under 3	4	4
3–under 5	12	16
5–under 7	13	29
7–under 9	19	48
9–under 11	7	55
11–under 13	5	60

TABLE 3.7
Calculation of Grouped Mean

Class Interval	Frequency (f)	Class Midpoint (M)	fM
1–under 3	4	2	8
3–under 5	12	4	48
5–under 7	13	6	78
7–under 9	19	8	152
9–under 11	7	10	70
11–under 13	5	12	60
	$\Sigma f = N = 60$		$\Sigma fM = 416$

$$\mu = \frac{\Sigma fM}{\Sigma f} = \frac{416}{60} = 6.93$$

value into the median class interval. However, there are 19 values in the median interval (denoted in the formula as f_{med}). The median value is 1/19 of the way through this interval.

$$\frac{\frac{N}{2} - cf_p}{f_{med}} = \frac{30 - 29}{19} = \frac{1}{19}$$

Thus, the median value is at least 7– the value of L– and is 1/19 of the way across the median interval. How far is it across the median interval? Each class interval is 2 units wide (w). Taking 1/19 of this distance tells us how far the median value is into the class interval.

$$\frac{\frac{N}{2} - cf_p}{f_{med}}(W) = \frac{\frac{60}{2} - 29}{19}(2) = \frac{1}{19}(2) = .105$$

Adding this distance to the lower endpoint of the median class interval yields the value of the median.

$$\text{Median} = 7 + \frac{\frac{60}{2} - 29}{19}(2) = 7 + \frac{1}{19}(2) = 7 + .105 = 7.105$$

The median value of unemployment rates for Canada is 7.105. Keep in mind that like the grouped mean, this median value is merely approximate. The assumption made in these calculations is that the actual values fall uniformly across the median class interval—which may or may not be the case.

Mode

The *mode* for grouped data is *the class midpoint of the modal class. The modal class is the class interval with the greatest frequency.* Using the data from Table 3.7, the 7–under 9 class interval contains the greatest frequency, 19. Thus, the modal class is 7–under 9. The class midpoint of this modal class is 8. Therefore, the mode for the frequency distribution shown in Table 3.7 is 8. The modal unemployment rate is 8%.

Measures of Variability

Two measures of variability for grouped data are presented here: the variance and the standard deviation. Again, the standard deviation is the square root of the variance. Both measures have original and computational formulas.

FORMULAS FOR POPULATION VARIANCE AND STANDARD DEVIATION OF GROUPED DATA

Original Formula	Computational Version
$\sigma^2 = \frac{\Sigma f(M - \mu)^2}{N}$	$\sigma^2 = \frac{\Sigma fM^2 - \frac{(\Sigma fM)^2}{N}}{N}$
$\sigma = \sqrt{\sigma^2}$	

where:

f = frequency
M = class midpoint
$N = \Sigma f$, or total frequencies of the population
μ = grouped mean for the population

TABLE 3.8

Calculating Grouped Variance and Standard Deviation with the Original Formula

Class Interval	f	M	fM	$(M-\mu)$	$(M-\mu)^2$	$f(M-\mu)^2$
1–under 3	4	2	8	−4.93	24.305	97.220
3–under 5	12	4	48	−2.93	8.585	103.020
5–under 7	13	6	78	−0.93	0.865	11.245
7–under 9	19	8	152	1.07	1.145	21.755
9–under 11	7	10	70	3.07	9.425	65.975
11–under 13	5	12	60	5.07	25.705	128.525
	$\Sigma f = N = 60$		$\Sigma fM = 416$			$\Sigma f(M-\mu)^2 = 427.740$

$$\mu = \frac{\Sigma fM}{\Sigma f} = \frac{416}{60} = 6.93$$

$$\sigma^2 = \frac{\Sigma f(M-\mu)^2}{N} = \frac{427.74}{60} = 7.129$$

$$\sigma = \sqrt{7.129} = 2.670$$

FORMULAS FOR SAMPLE VARIANCE AND STANDARD DEVIATION OF GROUPED DATA

Original Formula	Computational Version
$s^2 = \dfrac{\Sigma f(M-\bar{x})^2}{n-1}$	$s^2 = \dfrac{\Sigma fM^2 - \dfrac{(\Sigma fM)^2}{n}}{n-1}$
$s = \sqrt{s^2}$	

where:

f = frequency

M = class midpoint

$n = \Sigma f$, or total of the frequencies of the sample

$\bar{x}$ = grouped mean for the sample

For example, let us calculate the variance and standard deviation of the Canadian unemployment data grouped as a frequency distribution in Table 3.6. If the data are treated as a population, the computations are as follows.

For the original formula, the computations are given in Table 3.8. The method of determining σ^2 and σ by using the computational formula is shown in Table 3.9. In either case, the variance of the unemployment data is 7.129 (squared percent), and the standard deviation is 2.67%. As with the computation of the grouped mean, the class midpoint is used to represent all values in a class interval. This approach may or may not be appropriate, depending on whether the average value in a class is at the midpoint. If this situation does not occur, then the variance and the standard deviation are only approximations. Because grouped statistics are usually computed without knowledge of the actual data, the statistics computed potentially may be only approximations.

TABLE 3.9

Calculating Grouped Variance and Standard Deviation with the Computational Formula

Class Interval	f	M	fM	fM^2
1–under 3	4	2	8	16
3–under 5	12	4	48	192
5–under 7	13	6	78	468
7–under 9	19	8	152	1216
9–under 11	7	10	70	700
11–under 13	5	12	60	720
	$\Sigma f = N = 60$		$\Sigma fM = 416$	$\Sigma fM^2 = 3312$

$$\sigma^2 = \frac{\Sigma fM^2 - \dfrac{(\Sigma fM)^2}{N}}{N} = \frac{3312 - \dfrac{416^2}{60}}{60} = \frac{3312 - 2884.27}{60} = \frac{427.73}{60} = 7.129$$

$$\sigma = \sqrt{7.129} = 2.670$$

DEMONSTRATION PROBLEM 3.7

Compute the mean, median, mode, variance, and standard deviation on the following sample data.

Class Interval	Frequency	Cumulative Frequency
10–under 15	6	6
15–under 20	22	28
20–under 25	35	63
25–under 30	29	92
30–under 35	16	108
35–under 40	8	116
40–under 45	4	120
45–under 50	2	122

Solution

The mean is computed as follows.

Class	f	M	fM
10–under 15	6	12.5	75.0
15–under 20	22	17.5	385.0
20–under 25	35	22.5	787.5
25–under 30	29	27.5	797.5
30–under 35	16	32.5	520.0
35–under 40	8	37.5	300.0
40–under 45	4	42.5	170.0
45–under 50	2	47.5	95.0
	$\Sigma f = n = 122$		$\Sigma fM = 3130.0$

$$\bar{x} = \frac{\Sigma fM}{\Sigma f} = \frac{3130}{122} = 25.66$$

The grouped mean is 25.66.

The grouped median is located at the 61st value (122/2). Observing the cumulative frequencies, the 61st value falls in the 20-under 25 class, making it the median class interval; and thus, the grouped median is at least 20. Since there are 28 cumulative values before the median class interval, 33 more (61 − 28) are needed to reach the grouped median. However, there are 35 values in the median class. The grouped median is located 33/35 of the way across the class interval which has a width of 5. The grouped median is $20 + \frac{33}{35}(5) = 20 + 4.71 = 24.71$.

The grouped mode can be determined by finding the class midpoint of the class interval with the greatest frequency. The class with the greatest frequency is 20–under 25 with a frequency of 35. The midpoint of this class is 22.5, which is the grouped mode.

The variance and standard deviation can be found as shown next. First, use the original formula.

Class	f	M	$M - \bar{x}$	$(M - \bar{x})^2$	$f(M - \bar{x})^2$
10–under 15	6	12.5	−13.16	173.19	1039.14
15–under 20	22	17.5	−8.16	66.59	1464.98
20–under 25	35	22.5	−3.16	9.99	349.65
25–under 30	29	27.5	1.84	3.39	98.31
30–under 35	16	32.5	6.84	46.79	748.64
35–under 40	8	37.5	11.84	140.19	1121.52
40–under 45	4	42.5	16.84	283.59	1134.36
45–under 50	2	47.5	21.84	476.99	953.98
	$\Sigma f = n = 122$				$\Sigma f(M - \bar{x})^2 = 6910.58$

$$s^2 = \frac{\Sigma f(M - \bar{x})^2}{n - 1} = \frac{6910.58}{121} = 57.11$$

$$s = \sqrt{57.11} = 7.56$$

Next, use the computational formula.

Class	f	M	fM	fM^2
10–under 15	6	12.5	75.0	937.50
15–under 20	22	17.5	385.0	6,737.50
20–under 25	35	22.5	787.5	17,718.75
25–under 30	29	27.5	797.5	21,931.25
30–under 35	16	32.5	520.0	16,900.00
35–under 40	8	37.5	300.0	11,250.00
40–under 45	4	42.5	170.0	7,225.00
45–under 50	2	47.5	95.0	4,512.50
	$\Sigma f = n = 122$		$\Sigma fM = 3{,}130.0$	$\Sigma fM^2 = 87{,}212.50$

$$s^2 = \frac{\Sigma fM^2 - \frac{(\Sigma fM)^2}{n}}{n-1} = \frac{87{,}212.5 - \frac{(3{,}130)^2}{122}}{121} = \frac{6{,}910.04}{121} = 57.11$$

$$s = \sqrt{57.11} = 7.56$$

The sample variance is 57.11 and the standard deviation is 7.56.

3.3 PROBLEMS

3.27 Compute the mean, the median, and the mode for the following data.

Class	f
0–under 2	39
2–under 4	27
4–under 6	16
6–under 8	15
8–under 10	10
10–under 12	8
12–under 14	6

3.28 Compute the mean, the median, and the mode for the following data.

Class	f
1.2–under 1.6	220
1.6–under 2.0	150
2.0–under 2.4	90
2.4–under 2.8	110
2.8–under 3.2	280

3.29 Determine the population variance and standard deviation for the following data by using the original formula.

Class	f
20–under 30	7
30–under 40	11
40–under 50	18
50–under 60	13
60–under 70	6
70–under 80	4

3.30 Determine the sample variance and standard deviation for the following data by using the computational formula.

Class	f
5–under 9	20
9–under 13	18
13–under 17	8
17–under 21	6
21–under 25	2

3.31 A random sample of voters in Nashville, Tennessee, is classified by age group, as shown by the following data.

Age Group	Frequency
18–under 24	17
24–under 30	22
30–under 36	26
36–under 42	35
42–under 48	33
48–under 54	30
54–under 60	32
60–under 66	21
66–under 72	15

a. Calculate the mean of the data.
b. Calculate the mode.
c. Calculate the median.
d. Calculate the variance.
e. Calculate the standard deviation.

3.32 The following data represent the number of appointments made per 15-minute interval by telephone solicitation for a lawn-care company. Assume these are population data.

Number of Appointments	Frequency of Occurrence
0–under 1	31
1–under 2	57
2–under 3	26
3–under 4	14
4–under 5	6
5–under 6	3

a. Calculate the mean of the data.
b. Calculate the mode.
c. Calculate the median.
d. Calculate the variance.
e. Calculate the standard deviation.

3.33 The Air Transport Association of America publishes figures on the busiest airports in the United States. The following frequency distribution has been constructed from these figures for a recent year. Assume these are population data.

Number of Passengers Arriving and Departing (millions)	Number of Airports
20–under 30	8
30–under 40	7
40–under 50	1
50–under 60	0
60–under 70	3
70–under 80	1

a. Calculate the mean of these data.
b. Calculate the mode.
c. Calculate the median.
d. Calculate the variance.
e. Calculate the standard deviation.

3.34 The frequency distribution shown represents the number of farms per state for the 50 United States, based on information from the U.S. Department of Agriculture. Determine the average number of farms per state from these data. The mean computed

from the original ungrouped data was 41,796 and the standard deviation was 38,856. How do your answers for these grouped data compare? Why might they differ?

Number of Farms per State	*f*
0–under 20,000	16
20,000–under 40,000	11
40,000–under 60,000	11
60,000–under 80,000	6
80,000–under 100,000	4
100,000–under 120,000	2

MEASURES OF SHAPE

Measures of shape are *tools that can be used to describe the shape of a distribution of data.* In this section, we examine two measures of shape, skewness and kurtosis. We also look at box-and-whisker plots.

Skewness

A distribution of data in which the right half is a mirror image of the left half is said to be *symmetrical*. One example of a symmetrical distribution is the normal distribution, or bell curve, shown in Figure 3.8 and presented in more detail in Chapter 6.

Skewness is when *a distribution is asymmetrical or lacks symmetry*. The distribution in Figure 3.8 has no skewness because it is symmetric. Figure 3.9 shows a distribution that is skewed left, or negatively skewed, and Figure 3.10 shows a distribution that is skewed right, or positively skewed.

The skewed portion is the long, thin part of the curve. Many researchers use skewed distribution to denote that the data are sparse at one end of the distribution and piled up at the other end. Instructors sometimes refer to a grade distribution as skewed, meaning that few students scored at one end of the grading scale, and many students scored at the other end.

Skewness and the Relationship of the Mean, Median, and Mode

The concept of skewness helps to understand the relationship of the mean, median, and mode. In a unimodal distribution (distribution with a single peak or mode) that is skewed, the mode is the apex (high point) of the curve and the median is the middle value. The mean tends to be located toward the tail of the distribution, because the mean is particularly affected by the extreme values. A bell-shaped or normal distribution with the mean, median, and mode all at the center of the distribution has no skewness. Figure 3.11 displays the relationship of the mean, median, and mode for different types of skewness.

Coefficient of Skewness

Statistician Karl Pearson is credited with developing at least two coefficients of skewness that can be used to determine the degree of skewness in a distribution. We present one of these coefficients here, referred to as a Pearsonian **coefficient of skewness**. This coefficient

FIGURE 3.8
Symmetrical Distribution

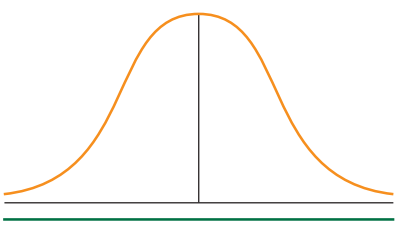


FIGURE 3.9
Distribution Skewed Left, or Negatively Skewed

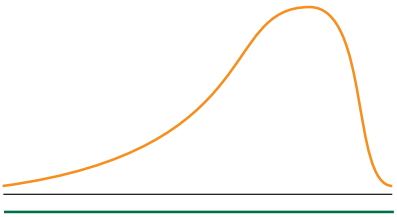


FIGURE 3.10
Distribution Skewed Right, or Positively Skewed

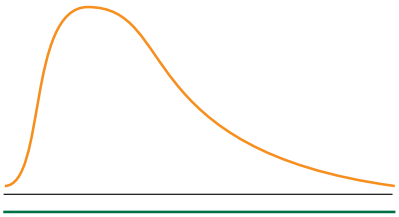
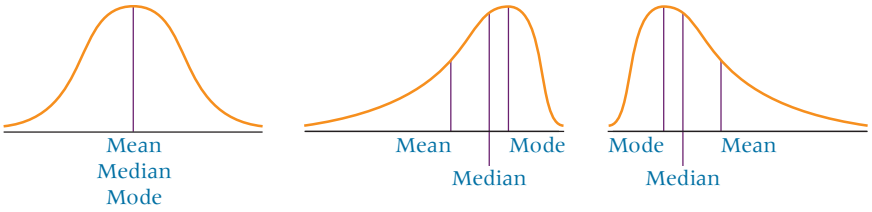


FIGURE 3.11
Relationship of Mean, Median, and Mode



(a) Symmetric distribution (no skewness) **(b) Negatively skewed** **(c) Positively skewed**

compares the mean and median in light of the magnitude of the standard deviation. Note that if the distribution is symmetrical, the mean and median are the same value and hence the coefficient of skewness is equal to zero.

COEFFICIENT OF SKEWNESS

$$S_k = \frac{3(\mu - M_d)}{\sigma}$$

where

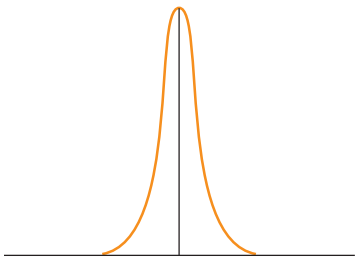
S_k = coefficient of skewness
M_d = median

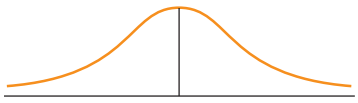
Suppose, for example, that a distribution has a mean of 29, a median of 26, and a standard deviation of 12.3. The coefficient of skewness is computed as

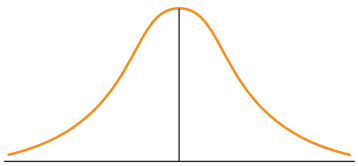
$$S_k = \frac{3(29 - 26)}{12.3} = +0.73$$

Because the value of S_k is positive, the distribution is positively skewed. If the value of S_k is negative, the distribution is negatively skewed. The greater the magnitude of S_k, the more skewed is the distribution.

FIGURE 3.12
Types of Kurtosis


Leptokurtic distribution


Platykurtic distribution


Mesokurtic distribution

Kurtosis

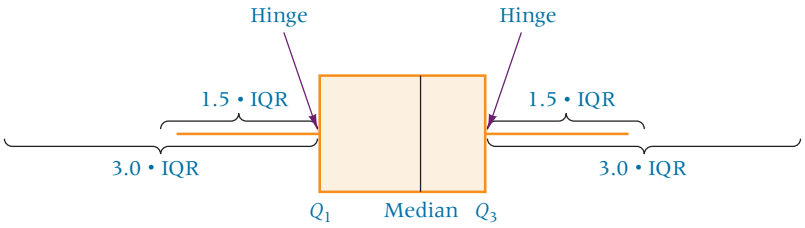
Kurtosis *describes the amount of peakedness of a distribution.* Distributions that are high and thin are referred to as **leptokurtic** distributions. Distributions that are flat and spread out are referred to as **platykurtic** distributions. Between these two types are distributions that are more "normal" in shape, referred to as **mesokurtic** distributions. These three types of kurtosis are illustrated in Figure 3.12.

Box-and-Whisker Plots

Another way to describe a distribution of data is by using a box and whisker plot. A **box-and-whisker plot**, sometimes called a *box plot*, is *a diagram that utilizes the upper and lower quartiles along with the median and the two most extreme values to depict a distribution graphically*. The plot is constructed by using a box to enclose the median. This *box* is extended outward from the median along a continuum to the lower and upper quartiles, enclosing not only the median but also the middle 50% of the data. From the lower and upper quartiles, lines referred to as *whiskers* are extended out from the box toward the outermost data values. The box-and-whisker plot is determined from five specific numbers.

1. The median (Q_2)
2. The lower quartile (Q_1)
3. The upper quartile (Q_3)

FIGURE 3.13
Box-and-Whisker Plot



4. The smallest value in the distribution
5. The largest value in the distribution

The box of the plot is determined by locating the median and the lower and upper quartiles on a continuum. A box is drawn around the median with the lower and upper quartiles (Q_1 and Q_3) as the box endpoints. These box endpoints (Q_1 and Q_3) are referred to as the *hinges* of the box.

Next the value of the interquartile range (IQR) is computed by $Q_3 - Q_1$. The interquartile range includes the middle 50% of the data and should equal the length of the box. However, here the interquartile range is used outside of the box also. At a distance of $1.5 \cdot \text{IQR}$ outward from the lower and upper quartiles are what are referred to as *inner fences*. A *whisker*, a line segment, is drawn from the lower hinge of the box outward to the smallest data value. A second whisker is drawn from the upper hinge of the box outward to the largest data value. The inner fences are established as follows.

$$Q_1 - 1.5 \cdot \text{IQR}$$
$$Q_3 + 1.5 \cdot \text{IQR}$$

If data fall beyond the inner fences, then *outer* fences can be constructed:

$$Q_1 - 3.0 \cdot \text{IQR}$$
$$Q_3 + 3.0 \cdot \text{IQR}$$

Figure 3.13 shows the features of a box-and-whisker plot.

Data values outside the mainstream of values in a distribution are viewed as *outliers*. Outliers can be merely the more extreme values of a data set. However, sometimes outliers occur due to measurement or recording errors. Other times they are values so unlike the other values that they should not be considered in the same analysis as the rest of the distribution. Values in the data distribution that are outside the inner fences but within the outer fences are referred to as *mild outliers*. Values that are outside the outer fences are called *extreme outliers*. Thus, one of the main uses of a box-and-whisker plot is to identify outliers. In some computer-produced box-and-whisker plots (such as in Minitab), the whiskers are drawn to the largest and smallest data values within the inner fences. An asterisk is then printed for each data value located between the inner and outer fences to indicate a mild outlier. Values outside the outer fences are indicated by a zero on the graph. These values are extreme outliers.

Another use of box-and-whisker plots is to determine whether a distribution is skewed. The location of the median in the box can relate information about the skewness of the middle 50% of the data. If the median is located on the right side of the box, then the middle 50% are skewed to the left. If the median is located on the left side of the box, then the middle 50% are skewed to the right. By examining the length of the whiskers on each side of the box, a business researcher can make a judgment about the skewness of the outer values. If the longest whisker is to the right of the box, then the outer data are skewed to the right and vice versa. We shall use the data given in Table 3.10 to construct a box-and-whisker plot.

TABLE 3.10
Data for Box-and-Whisker Plot

71	87	82	64	72	75	81	69
76	79	65	68	80	73	85	71
70	79	63	62	81	84	77	73
82	74	74	73	84	72	81	65
74	62	64	68	73	82	69	71

After organizing the data into an ordered array, as shown in Table 3.11, it is relatively easy to determine the values of the lower quartile (Q_1), the median, and the upper quartile (Q_3). From these, the value of the interquartile range can be computed.

The hinges of the box are located at the lower and upper quartiles, 69 and 80.5. The median is located within the box at distances of 4 from the lower quartile and 6.5 from

TABLE 3.11

Data in Ordered Array with Quartiles and Median

87	85	84	84	82	82	82	81	81	81
80	79	79	77	76	75	74	74	74	73
73	73	73	72	72	71	71	71	70	69
69	68	68	65	65	64	64	63	62	62

$Q_1 = 69$

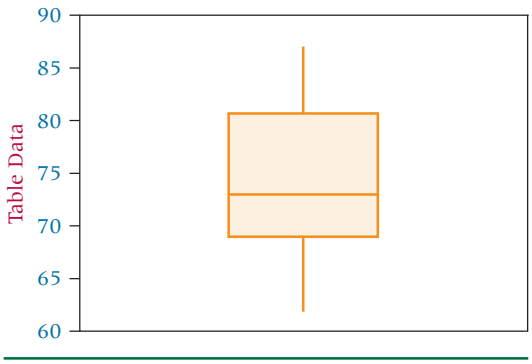
$Q_2 = \text{median} = 73$

$Q_3 = 80.5$

$\text{IQR} = Q_3 - Q_1 = 80.5 - 69 = 11.5$

FIGURE 3.14

Minitab Box-and-Whisker Plot



the upper quartile. The distribution of the middle 50% of the data is skewed right, because the median is nearer to the lower or left hinge. The inner fence is constructed by

$$Q_1 - 1.5 \cdot \text{IQR} = 69 - 1.5(11.5) = 69 - 17.25 = 51.75$$

and

$$Q_3 + 1.5 \cdot \text{IQR} = 80.5 + 1.5(11.5) = 80.5 + 17.25 = 97.75$$

The whiskers are constructed by drawing a line segment from the lower hinge outward to the smallest data value and a line segment from the upper hinge outward to the largest data value. An examination of the data reveals that no data values in this set of numbers are outside the inner fence. The whiskers are constructed outward to the lowest value, which is 62, and to the highest value, which is 87.

To construct an outer fence, we calculate $Q_1 - 3 \cdot \text{IQR}$ and $Q_3 + 3 \cdot \text{IQR}$, as follows.

$$Q_1 - 3 \cdot \text{IQR} = 69 - 3(11.5) = 69 - 34.5 = 34.5$$

$$Q_3 + 3 \cdot \text{IQR} = 80.5 + 3(11.5) = 80.5 + 34.5 = 115.0$$

Figure 3.14 is the Minitab computer printout for this box-and-whisker plot.

3.4 PROBLEMS

3.35 On a certain day the average closing price of a group of stocks on the New York Stock Exchange is $35 (to the nearest dollar). If the median value is $33 and the mode is $21, is the distribution of these stock prices skewed? If so, how?

3.36 A local hotel offers ballroom dancing on Friday nights. A researcher observes the customers and estimates their ages. Discuss the skewness of the distribution of ages if the mean age is 51, the median age is 54, and the modal age is 59.

3.37 The sales volumes for the top real estate brokerage firms in the United States for a recent year were analyzed using descriptive statistics. The mean annual dollar volume for these firms was $5.51 billion, the median was $3.19 billion, and the standard deviation was $9.59 billion. Compute the value of the Pearsonian coefficient of skewness and discuss the meaning of it. Is the distribution skewed? If so, to what extent?

3.38 Suppose the following data are the ages of Internet users obtained from a sample. Use these data to compute a Pearsonian coefficient of skewness. What is the meaning of the coefficient?

41	15	31	25	24
23	21	22	22	18
30	20	19	19	16
23	27	38	34	24
19	20	29	17	23

3.39 Construct a box-and-whisker plot on the following data. Do the data contain any outliers? Is the distribution of data skewed?

540	690	503	558	490	609
379	601	559	495	562	580
510	623	477	574	588	497
527	570	495	590	602	541

3.40 Suppose a consumer group asked 18 consumers to keep a yearly log of their shopping practices and that the following data represent the number of coupons used by each consumer over the yearly period. Use the data to construct a box-and-whisker plot. List the median, Q_1, Q_3, the endpoints for the inner fences, and the endpoints for the outer fences. Discuss the skewness of the distribution of these data and point out any outliers.

81	68	70	100	94	47	66	70	82
110	105	60	21	70	66	90	78	85

3.5 DESCRIPTIVE STATISTICS ON THE COMPUTER

Both Minitab and Excel yield extensive descriptive statistics. Even though each computer package can compute individual statistics such as a mean or a standard deviation, they can also produce multiple descriptive statistics at one time. Figure 3.15 displays a Minitab output for the descriptive statistics associated with the computer production data presented earlier in this section. The Minitab output contains, among other things, the mean, the median, the sample standard deviation, the minimum and maximum (which can then be used to compute the range), and Q_1 and Q_3 (from which the interquartile range can be computed). Excel's descriptive statistics output for the same computer production data is displayed in Figure 3.16. The Excel output contains the mean, the median, the mode, the sample standard deviation, the sample variance, and the range. The descriptive statistics feature on either of these computer packages yields a lot of useful information about a data set.

FIGURE 3.15

Minitab Output for the Computer Production Problem

DESCRIPTIVE STATISTICS

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q_1
Computers Produced	5	0	13.00	2.55	5.70	5.00	7.00

Variable	Median	Q_3	Maximum
Computers Produced	16.00	17.50	18.00

FIGURE 3.16

Excel Output for the Computer Production Problem

COMPUTER PRODUCTION DATA

Mean	13
Standard error	2.54951
Median	16
Mode	#N/A
Standard deviation	5.700877
Sample variance	32.5
Kurtosis	–1.71124
Skewness	–0.80959
Range	13
Minimum	5
Maximum	18
Sum	65
Count	5

Laundry Statistics

The descriptive statistics presented in this chapter are excellent for summarizing and presenting data sets in more concise formats. For example, question 1 of the managerial and statistical questions in the Decision Dilemma reports water measurements for 50 U.S. households. Using Excel and/or Minitab, many of the descriptive statistics presented in this chapter can be applied to these data. The results are shown in Figures 3.17 and 3.18.

Decision Dilemma Solved

These computer outputs show that the average water usage is 15.48 gallons with a standard deviation of about 1.233 gallons. The median is 16 gallons with a range of 6 gallons (12 to 18). The first quartile is 15 gallons and the third quartile is 16 gallons. The mode is also 16 gallons. The Minitab graph and the skewness measures show that the data are slightly skewed to the left. Applying Chebyshev's theorem to the mean and standard deviation shows that at least 88.9% of the measurements should fall between 11.78 gallons and 19.18 gallons. An examination of the data and the minimum and maximum reveals that 100% of the data actually fall within these limits.

According to the Decision Dilemma, the mean wash cycle time is 35 minutes with a standard deviation of 5 minutes. If the wash cycle times are approximately normally distributed, we can apply the empirical rule. According to the empirical rule, 68% of the times would fall within 30 and 40 minutes, 95% of the times would fall within 25 and 45 minutes, and 99.7% of the wash times would fall within 20 and 50 minutes. If the data are not normally distributed, Chebyshev's theorem reveals that at least 75% of the times should fall between 25 and 45 minutes and 88.9% should fall between 20 and 50 minutes.

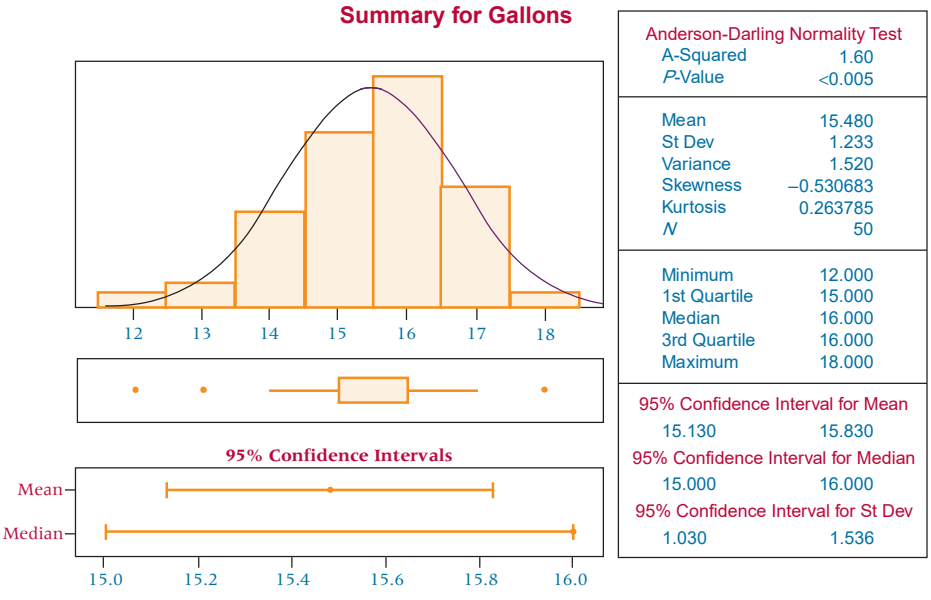
FIGURE 3.17

Excel Descriptive Statistics

GALLONS OF WATER	
Mean	15.48
Standard error	0.174356
Median	16
Mode	16
Standard deviation	1.232883
Sample variance	1.52
Kurtosis	0.263785
Skewness	–0.53068
Range	6
Minimum	12
Maximum	18
Sum	774
Count	50

FIGURE 3.18

Minitab Descriptive Statistics



Anderson-Darling Normality Test	
A-Squared	1.60
P-Value	<0.005
Mean	15.480
St Dev	1.233
Variance	1.520
Skewness	–0.530683
Kurtosis	0.263785
N	50
Minimum	12.000
1st Quartile	15.000
Median	16.000
3rd Quartile	16.000
Maximum	18.000
95% Confidence Interval for Mean	
15.130	15.830
95% Confidence Interval for Median	
15.000	16.000
95% Confidence Interval for St Dev	
1.030	1.536

ETHICAL CONSIDERATIONS

In describing a body of data to an audience, it is best to use whatever measures it takes to present a "full" picture of the data. By limiting the descriptive measures used, the business researcher may give the audience only part of the picture and can skew the way the receiver understands the data. For example, if a researcher presents only the mean, the audience will have no insight into the variability of the data; in addition, the mean might be inordinately large or small because of extreme values. Likewise, the choice of the median precludes a picture that includes the extreme values. Using the mode can cause the receiver of the information to focus only on values that occur often.

At least one measure of variability is usually needed with at least one measure of central tendency for the audience to begin to understand what the data look like. Unethical researchers might be tempted to present only the descriptive measure that will convey the picture of the data that they want the audience to see. Ethical researchers will instead use any and all methods that will present the fullest, most informative picture possible from the data.

Former governor of Colorado Richard Lamm has been quoted as having said that "Demographers are academics who can statistically prove that the average person in Miami is born Cuban and dies Jewish . . . ."* People are more likely to reach this type of conclusion if incomplete or misleading descriptive statistics are provided by researchers.

*Alan L. Otten."People Patterns/Odds and Ends," *The Wall Street Journal*, June 29, 1992, p. B1. Reprinted by permission of *The Wall Street Journal*

SUMMARY

Statistical descriptive measures include measures of central tendency, measures of variability, and measures of shape. Measures of central tendency and measures of variability are computed differently for ungrouped and grouped data. Measures of central tendency are useful in describing data because they communicate information about the more central portions of the data. The most common measures of central tendency are the three Ms': mode, median, and mean. In addition, percentiles and quartiles are measures of central tendency.

The mode is the most frequently occurring value in a set of data. Among other things, the mode is used in business for determining sizes.

The median is the middle term in an ordered array of numbers containing an odd number of terms. For an array with an even number of terms, the median is the average of the two middle terms. A median is unaffected by the magnitude of extreme values. This characteristic makes the median a most useful and appropriate measure of location in reporting such things as income, age, and prices of houses.

The arithmetic mean is widely used and is usually what researchers are referring to when they use the word *mean*. The arithmetic mean is the average. The population mean and the sample mean are computed in the same way but are denoted by different symbols. The arithmetic mean is affected by every value and can be inordinately influenced by extreme values.

Percentiles divide a set of data into 100 groups, which means 99 percentiles are needed. Quartiles divide data into four groups. The three quartiles are Q_1, which is the lower quartile; Q_2, which is the middle quartile and equals the median; and Q_3, which is the upper quartile.

Measures of variability are statistical tools used in combination with measures of central tendency to describe data. Measures of variability provide information about the spread of the data values. These measures include the range, mean absolute deviation, variance, standard deviation, interquartile range, z scores, and coefficient of variation for ungrouped data.

One of the most elementary measures of variability is the range. It is the difference between the largest and smallest values. Although the range is easy to compute, it has limited usefulness. The interquartile range is the difference between the third and first quartile. It equals the range of the middle 50% of the data.

The mean absolute deviation (MAD) is computed by averaging the absolute values of the deviations from the mean. The mean absolute deviation provides the magnitude of the average deviation but without specifying its direction. The mean absolute deviation has limited usage in statistics, but interest is growing for the use of MAD in the field of forecasting.

Variance is widely used as a tool in statistics but is used little as a stand-alone measure of variability. The variance is the average of the squared deviations about the mean.

The square root of the variance is the standard deviation. It also is a widely used tool in statistics, but it is used more often than the variance as a stand-alone measure. The standard deviation is best understood by examining its applications in determining where data are in relation to the mean. The empirical rule and Chebyshev's theorem are statements about the proportions of data values that are within various numbers of standard deviations from the mean.

The empirical rule reveals the percentage of values that are within one, two, or three standard deviations of the mean for a set of data. The empirical rule applies only if the data are in a bell-shaped distribution.

Chebyshev's theorem also delineates the proportion of values that are within a given number of standard deviations from the mean. However, it applies to any distribution. The z score represents the number of standard deviations a value is from the mean for normally distributed data.

The coefficient of variation is a ratio of a standard deviation to its mean, given as a percentage. It is especially useful in comparing standard deviations or variances that represent data with different means.

Some measures of central tendency and some measures of variability are presented for grouped data. These measures include mean, median, mode, variance, and standard deviation. Generally, these measures are only approximate for grouped data because the values of the actual raw data are unknown.

Two measures of shape are skewness and kurtosis. Skewness is the lack of symmetry in a distribution. If a distribution is skewed, it is stretched in one direction or the other. The skewed part of a graph is its long, thin portion. One measure of skewness is the Pearsonian coefficient of skewness.

Kurtosis is the degree of peakedness of a distribution. A tall, thin distribution is referred to as leptokurtic. A flat distribution is platykurtic, and a distribution with a more normal peakedness is said to be mesokurtic.

A box-and-whisker plot is a graphical depiction of a distribution. The plot is constructed by using the median, the lower quartile, and the upper quartile. It can yield information about skewness and outliers.

KEY TERMS

arithmetic mean
bimodal
box-and-whisker plot
Chebyshev's theorem
coefficient of skewness
coefficient of variation (CV)
deviation from the mean
empirical rule
interquartile range
kurtosis
leptokurtic
mean absolute deviation (MAD)
measures of central tendency
measures of shape
measures of variability
median
mesokurtic
mode
multimodal
percentiles
platykurtic
quartiles
range
skewness
standard deviation
sum of squares of x
variance
z score

FORMULAS

Population mean (ungrouped)

$$\mu = \frac{\Sigma x}{N}$$

Sample mean (ungrouped)

$$\bar{x} = \frac{\Sigma x}{n}$$

Mean absolute deviation

$$\text{MAD} = \frac{\Sigma |x - \mu|}{N}$$

Population variance (ungrouped)

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N}$$

$$\sigma^2 = \frac{\Sigma x^2 - N\mu^2}{N}$$

Population standard deviation (ungrouped)

$$\sigma = \sqrt{\sigma^2}$$

$$\sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{N}}$$

$$\sigma = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N}}$$

$$\sigma = \sqrt{\frac{\Sigma x^2 - N\mu^2}{N}}$$

Grouped mean

$$\mu_{\text{grouped}} = \frac{\Sigma fM}{N}$$

Grouped Median

$$Median = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W)$$

Population variance (grouped)

$$\sigma^2 = \frac{\Sigma f(M - \mu)^2}{N} = \frac{\Sigma fM^2 - \frac{(\Sigma fM)^2}{N}}{N}$$

Population standard deviation (grouped)

$$\sigma = \sqrt{\frac{\Sigma f(M-\mu)^2}{N}} = \sqrt{\frac{\Sigma fM^2 - \frac{(\Sigma fM)^2}{N}}{N}}$$

Sample variance

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

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

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

Sample standard deviation

$$s = \sqrt{s^2}$$

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

$$s = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$$

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

Chebyshev's theorem

$$1 - \frac{1}{k^2}$$

z score

$$z = \frac{x-\mu}{\sigma}$$

Coefficient of variation

$$\text{CV} = \frac{\sigma}{\mu}(100)$$

Interquartile range

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

Sample variance (grouped)

$$s^2 = \frac{\Sigma f(M-\bar{x})^2}{n-1} = \frac{\Sigma fM^2 - \frac{(\Sigma fM)^2}{n}}{n-1}$$

Sample standard deviation (grouped)

$$s = \sqrt{\frac{\Sigma f(M-\bar{x})^2}{n-1}} = \sqrt{\frac{\Sigma fM^2 - \frac{(\Sigma fM)^2}{n}}{n-1}}$$

Pearsonian coefficient of skewness

$$S_k = \frac{3(\mu - M_d)}{\sigma}$$

SUPPLEMENTARY PROBLEMS

CALCULATING THE STATISTICS

3.41 The 2000 U.S. Census asked every household to report information on each person living there. Suppose for a sample of 30 households selected, the number of persons living in each was reported as follows.

2	3	1	2	6	4	2	1	5	3	2	3	1	2	2
1	3	1	2	2	4	2	1	2	8	3	2	1	1	3

Compute the mean, median, mode, range, lower and upper quartiles, and interquartile range for these data.

3.42 The 2000 U.S. Census also asked for each person's age. Suppose that a sample of 40 households taken from the census data showed the age of the first person recorded on the census form to be as follows.

42	29	31	38	55	27	28
33	49	70	25	21	38	47
63	22	38	52	50	41	19
22	29	81	52	26	35	38
29	31	48	26	33	42	58
40	32	24	34	25		

Compute P_{10}, P_{80}, Q_1, Q_3, the interquartile range, and the range for these data.

3.43 Shown below are the top 20 companies in the computer industry by sales according to netvalley.com in a recent year. Compute the mean, median, P_{30}, P_{60}, P_{90}, Q_1, Q_3, range, and interquartile range on these data.

Company	Sales ($ millions)
IBM	91,134
Hewlett Packard	86,696
Verizon Communications	75,112
Dell	49,205
Microsoft	39,788
Intel	38,826
Motorola	36,843
Sprint	34,680
Canon	34,222
Ingram Micro	28,808
Cisco Systems	24,801
EDS	19,757
Xerox	15,701
Computer Sciences	14,059
Apple	13,931
Texas Instruments	13,392
Oracle	11,799
Sanmina-SCI	11,735
Arrow Electronics	11,164
Sun Microsystems	11,070

3.44 Shown in right column are the top 10 companies receiving the largest dollar volume of contract awards from the U.S. Department of Defense in a recent year. Use this population data to compute a mean and a standard deviation for these top 10 companies.

Company	Amount of Contracts ($ billions)
Lockheed Martin	27.32
Boeing	20.86
Northrop Grumman	16.77
General Dynamics	11.47
Raytheon	10.41
KBR	5.97
L-3 Communications	5.04
United Technologies	4.57
BAE Systems	4.50
SAIC	3.40

3.45 Shown here are the U.S. oil refineries with the largest capacity in terms of barrels per day according to the U.S. Energy Information Administration. Use these as population data and answer the questions.

Refinery Location	Company	Capacity
Baytown, Texas	ExxonMobil	567,000
Baton Rouge, Louisiana	ExxonMobil	503,000
Texas City, Texas	BP	467,720
Lake Charles, Louisiana	Citgo	429,500
Whiting, Indiana	BP	410,000
Beaumont, Texas	ExxonMobil	348,500
Philadelphia, Pennsylvania	Sunoco	335,000
Pascagoula, Mississippi	Chevron	330,000
Deer Park, Texas	partnership	329,800
Wood River, Illinois	WRB	306,000
Port Arthur, Texas	Premcor	289,000

a. What are the values of the mean and the median? Compare the answers and state which you prefer as a measure of location for these data and why.
b. What are the values of the range and interquartile range? How do they differ?
c. What are the values of variance and standard deviation for these data?
d. What is the z score for Pascagoula, Mississippi? What is the z score for Texas City, Texas? Interpret these z scores.
e. Calculate the Pearsonian coefficient of skewness and comment on the skewness of this distribution.

3.46 The U.S. Department of the Interior releases figures on mineral production. Following are the 14 leading states in nonfuel mineral production in the United States.

State	Value ($ billions)
Arizona	4.35
California	4.24
Nevada	3.88
Florida	2.89
Utah	2.79
Texas	2.72
Minnesota	2.19
Missouri	1.94
Georgia	1.81
Colorado	1.75
Michigan	1.75
Pennsylvania	1.55
Alaska	1.47
Wyoming	1.30

a. Calculate the mean, median, and mode.
b. Calculate the range, interquartile range, mean absolute deviation, sample variance, and sample standard deviation.
c. Compute the Pearsonian coefficient of skewness for these data.
d. Sketch a box-and-whisker plot.

3.47 The radio music listener market is diverse. Listener formats might include adult contemporary, album rock, top 40, oldies, rap, country and western, classical, and jazz. In targeting audiences, market researchers need to be concerned about the ages of the listeners attracted to particular formats. Suppose a market researcher surveyed a sample of 170 listeners of country music radio stations and obtained the following age distribution.

Age	Frequency
15–under 20	9
20–under 25	16
25–under 30	27
30–under 35	44
35–under 40	42
40–under 45	23
45–under 50	7
50–under 55	2

a. What are the mean and modal ages of country music listeners?
b. What are the variance and standard deviation of the ages of country music listeners?

3.48 A research agency administers a demographic survey to 90 telemarketing companies to determine the size of their operations. When asked to report how many employees now work in their telemarketing operation, the companies gave responses ranging from 1 to 100. The agency's analyst organizes the figures into a frequency distribution.

Number of Employees Working in Telemarketing	Number of Companies
0–under 20	32
20–under 40	16
40–under 60	13
60–under 80	10
80–under 100	19

a. Compute the mean, median, and mode for this distribution.
b. Compute the sample standard deviation for these data.

TESTING YOUR UNDERSTANDING

3.49 Financial analysts like to use the standard deviation as a measure of risk for a stock. The greater the deviation in a stock price over time, the more risky it is to invest in the stock. However, the average prices of some stocks are considerably higher than the average price of others, allowing for the potential of a greater standard deviation of price. For example, a standard deviation of $5.00 on a $10.00 stock is considerably different from a $5.00 standard deviation on a $40.00 stock. In this situation, a coefficient of variation might provide insight into risk. Suppose stock X costs an average of $32.00 per share and showed a standard deviation of $3.45 for the past 60 days. Suppose stock Y costs an average of $84.00 per share and showed a standard deviation of $5.40 for the past 60 days. Use the coefficient of variation to determine the variability for each stock.

3.50 The Polk Company reported that the average age of a car on U.S. roads in a recent year was 7.5 years. Suppose the distribution of ages of cars on U.S. roads is approximately bellshaped. If 99.7% of the ages are between 1 year and 14 years, what is the standard deviation of car age? Suppose the standard deviation is 1.7 years and the mean is 7.5 years. Between what two values would 95% of the car ages fall?

3.51 According to a *Human Resources* report, a worker in the industrial countries spends on average 419 minutes a day on the job. Suppose the standard deviation of time spent on the job is 27 minutes.

a. If the distribution of time spent on the job is approximately bell shaped, between what two times would 68% of the figures be? 95%? 99.7%?

b. If the shape of the distribution of times is unknown, approximately what percentage of the times would be between 359 and 479 minutes?

c. Suppose a worker spent 400 minutes on the job. What would that worker's z score be, and what would it tell the researcher?

3.52 During the 1990s, businesses were expected to show a lot of interest in Central and Eastern European countries. As new markets began to open, American businesspeople needed a better understanding of the market potential there. The following are the per capita GDP figures for eight of these European countries published by the *World Almanac*. **Note:** The per capita GDP for the U.S. is $44,000.

Country	Per Capita GDP (U.S. $)
Albania	5,700
Bulgaria	10,700
Croatia	13,400
Czech Republic	21,900
Hungary	17,600
Poland	14,300
Romania	9,100
Bosnia/Herzegovina	5,600

a. Compute the mean and standard deviation for Albania, Bulgaria, Croatia, and Czech Republic.

b. Compute the mean and standard deviation for Hungary, Poland, Romania, and Bosnia/Herzegovina.

c. Use a coefficient of variation to compare the two standard deviations. Treat the data as population data.

3.53 According to the Bureau of Labor Statistics, the average annual salary of a worker in Detroit, Michigan, is $35,748. Suppose the median annual salary for a worker in this group is $31,369 and the mode is $29,500. Is the distribution of salaries for this group skewed? If so, how and why? Which of these measures of central tendency would you use to describe these data? Why?

3.54 According to the U.S. Army Corps of Engineers, the top 20 U.S. ports, ranked by total tonnage (in million tons), were as follows.

Port	Total Tonnage
South Louisiana, LA	212.7
Houston, TX	211.7
New York, NY and NJ	152.1
Huntington, WV, KY, and OH	83.9
Long Beach, CA	79.9
Beaumont, TX	78.9
Corpus Christi, TX	77.6
New Orleans, LA	65.9
Baton Rouge, LA	59.3
Texas City, TX	57.8
Mobile, AL	57.7
Los Angeles, CA	54.9
Lake Charles, LA	52.7
Tampa, FL	49.2
Plaquemines, LA	47.9
Duluth-Superior MN and WI	44.7
Valdez, AK	44.4
Baltimore, MD	44.1
Pittsburgh, PA	43.6
Philadelphia, PA	39.4

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

b. Discuss the shape of the distribution from the plot.

c. Are there outliers?

d. What are they and why do you think they are outliers?

3.55 *Runzheimer International* publishes data on overseas business travel costs. They report that the average per diem total for a business traveler in Paris, France, is $349. Suppose the shape of the distribution of the per diem costs of a business traveler to Paris is unknown, but that 53% of the per diem figures are between $317 and $381. What is the value of the standard deviation? The average per diem total for a business traveler in Moscow is $415. If the shape of the distribution of per diem costs of a business traveler in Moscow is unknown and if 83% of the per diem costs in Moscow lie between $371 and $459, what is the standard deviation?

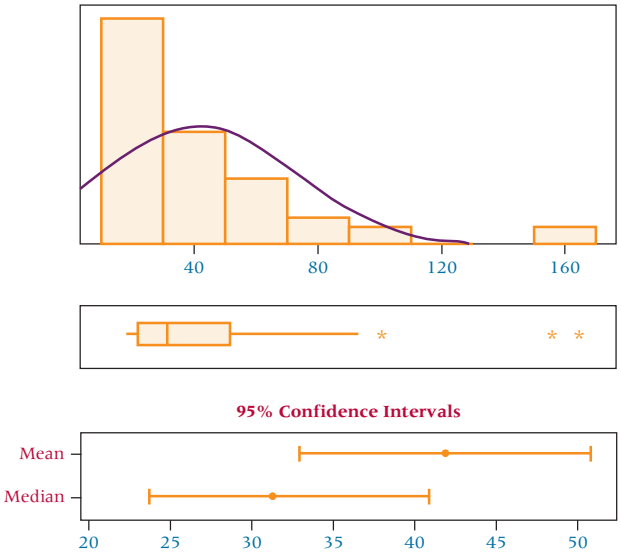
INTERPRETING THE OUTPUT

3.56 Netvalley.com compiled a list of the top 100 banks in the United States according to total assets. Leading the list was Bank of America, followed by JPMorgan Chase and Citibank. Following is an Excel analysis of total assets ($ billions) of these banks using the descriptive statistics feature. Study the output and describe in your own words what you can learn about the assets of these top 100 banks.

Top 100 Banks in U.S.	
Mean	76.5411
Standard error	17.93374
Median	21.97
Mode	13.01
Standard deviation	179.3374
Sample variance	32161.9
Kurtosis	22.2632
Skewness	4.586275
Range	1096.01
Minimum	8.99
Maximum	1105
Sum	7654.11
Count	100

3.57 *Hispanic Business* magazine publishes a list of the top 50 advertisers in the Hispanic market. The advertising spending for each of these 50 advertisers (in $ millions) was entered into a Minitab spreadsheet and the data were analyzed using Minitab's Graphical Summary. Study the output from this analysis and describe the advertising expenditures of these top Hispanic market advertisers.

Top 50 Advertisers in the Hispanic Market



Anderson-Darling Normality Test	
A-Squared	4.12
P-Value <	0.005
Mean	41.853
St Dev	31.415
Variance	986.892
Skewness	2.34450
Kurtosis	6.25496
N	50
Minimum	18.079
1st Quartile	21.617
Median	31.167
3rd Quartile	51.154
Maximum	162.695
95% Confidence Interval for Mean	
32.925	50.781
95% Confidence Interval for Median	
23.672	40.810
95% Confidence Interval for St Dev	
26.242	39.147

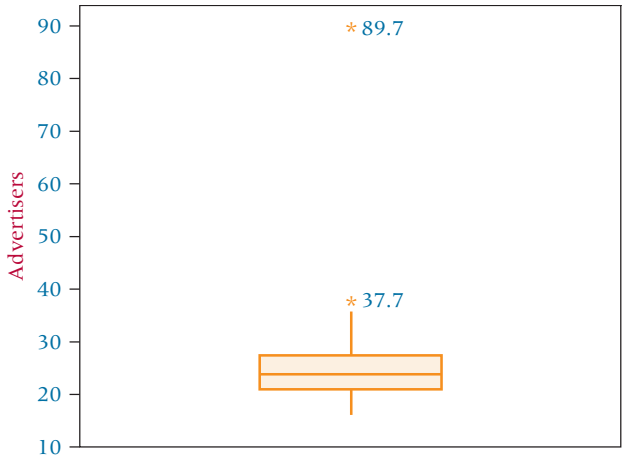
3.58 Excel was used to analyze the number of employees for the top 60 employers with headquarters around the world outside of the United States. The data was compiled by myglobalcareer.com, extracted from *My Global Career 500*, and was analyzed using Excel's descriptive statistics feature. Summarize what you have learned about the number of employees for these companies by studying the output.

Top 60 Employers Outside of the U.S.	
Mean	211942.9
Standard error	12415.31
Median	175660
Mode	150000
Standard deviation	96168.57
Sample variance	9.25E+09
Kurtosis	1.090976
Skewness	1.356847
Range	387245
Minimum	115300
Maximum	502545
Sum	12716575
Count	60

3.59 The Nielsen Company compiled a list of the top 25 advertisers in African American media. Shown below are a Minitab descriptive statistics analysis of the annual advertising spending in $ million by these companies in African American media and a box plot of these data. Study this output and summarize the expenditures of these top 25 advertisers in your own words.

```
Variable      N N*   Mean SE Mean StDev Minimum     Q1
Advertisers  25  0  27.24    2.84 14.19   16.20  21.25

Variable      Median      Q3   Maximum
Advertisers    24.00   27.50     89.70
```



ANALYZING THE DATABASES

see www.wiley.com/college/black

1. What are the mean and the median amounts of new capital expenditures for industries in the Manufacturing database? Comparing the mean and the median for these data, what do these statistics tell you about the data?
2. Analyze U.S. finished motor gasoline production in the 12-Year Gasoline database using a descriptive statistics feature of either Excel or Minitab. By examining such statistics as the mean, median, mode, range, standard deviation, and a measure of skewness, describe U.S. motor gasoline production over this period of time.
3. Using measures of central tendency including the mean, median, and mode, describe annual food spending and annual household income for the 200 households in the Consumer Food database. Compare the two results by determining approximately what percent of annual household income is spent on food.
4. Using the Financial database, study earnings per share for Type 2 and Type 7 companies (chemical and petrochemical) using statistics. Compute a coefficient of variation for both Type 2 and Type 7. Compare the two coefficients and comment on them.
5. Using the Hospital database, construct a box-and-whisker plot for births. Thinking about hospitals and birthing facilities, comment on why the box-and-whisker plot looks like it does.

CASE

COCA-COLA GOES SMALL IN RUSSIA

The Coca-Cola Company is the number-one seller of soft drinks in the world. Every day an average of more than 1.5 billion servings of Coca-Cola, Diet Coke, Sprite, Fanta, and other products of Coca-Cola are enjoyed around the world. The company has the world's largest production and distribution system for soft drinks and sells more than twice as many soft drinks as its nearest competitor. Coca-Cola products are sold in more than 200 countries around the globe.

For several reasons, the company believes it will continue to grow internationally. One reason is that disposable income is rising. Another is that outside the United States and Europe, the world is getting younger. In addition, reaching world markets is becoming easier as political barriers fall and transportation difficulties are overcome. Still another reason is that the sharing of ideas, cultures, and news around the world creates market opportunities. Part of the company mission is for Coca-Cola to maintain the world's most powerful trademark and effectively utilize the world's most effective and pervasive distribution system.

In June 1999, Coca-Cola Russia introduced a 200-milliliter (about 6.8 oz.) Coke bottle in Volgograd, Russia, in a campaign to market Coke to its poorest customers. This strategy was successful for Coca-Cola in other countries, such as India. The bottle sells for 12 cents, making it affordable to almost everyone. In 2001, Coca-Cola enjoyed a 25% volume growth in Russia, including an 18% increase in unit case sales of Coca-Cola.

Today, Coca-Cola beverages are produced and sold in Russia by the company's authorized local bottling partner, Coca-Cola HBC Russia, based in Moscow. The Coca-Cola business system directly employs approximately 4000 people in Russia, and more than 70% of all supplies required by the company are sourced locally.

Discussion

1. Because of the variability of bottling machinery, it is likely that every 200-milliliter bottle of Coca-Cola does not contain exactly 200 milliliters of fluid. Some bottles may contain more fluid and others less. Because 200-milliliter bottle

fills are somewhat unusual, a production engineer wants to test some of the bottles from the first production runs to determine how close they are to the 200-milliliter specification. Suppose the following data are the fill measurements from a random sample of 50 bottles. Use the techniques presented in this chapter to describe the sample. Consider measures of central tendency, variability, and skewness. Based on this analysis, how is the bottling process working?

200.1	199.9	200.2	200.2	200.0
200.1	200.9	200.1	200.3	200.5
199.7	200.4	200.3	199.8	199.3
200.1	199.4	199.6	199.2	200.2
200.4	199.8	199.9	200.2	199.6
199.6	200.4	200.4	200.6	200.6
200.1	200.8	199.9	200.0	199.9
200.3	200.5	199.9	201.1	199.7
200.2	200.5	200.2	199.7	200.9
200.2	199.5	200.6	200.3	199.8

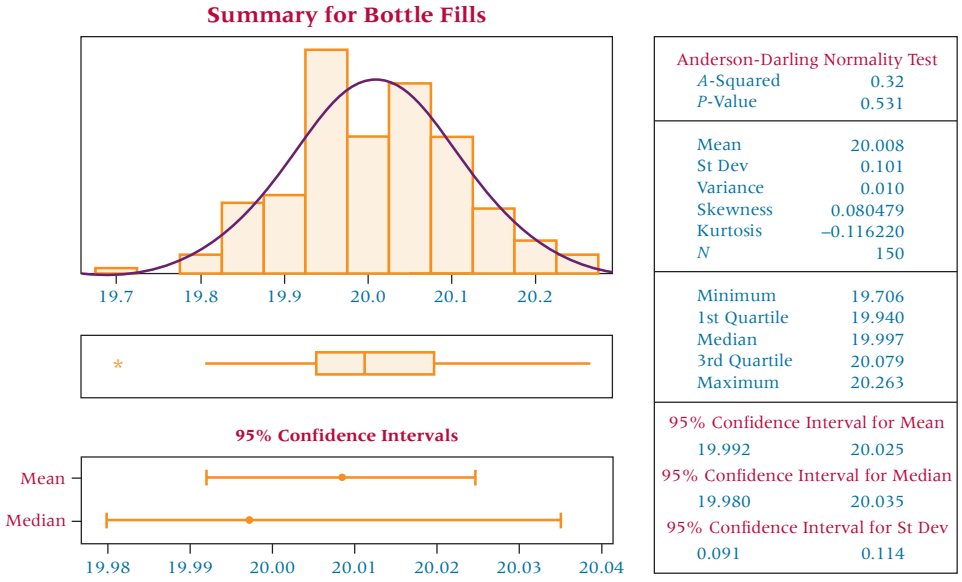
2. Suppose that at another plant Coca-Cola is filling bottles with the more traditional 20 ounces of fluid. A lab randomly samples 150 bottles and tests the bottles for fill volume. The descriptive statistics are given in both Minitab and Excel computer output. Write a brief report to supervisors summarizing what this output is saying about the process.

Minitab Output

Descriptive Statistics: Bottle Fills

Variable	Total Count	Mean	SE Mean	StDev	Variance	CoefVar	Minimum	Q_1
Bottle Fills	150	20.008	0.00828	0.101	0.0103	0.51	19.706	19.940

Variable	Median	Q_3	Maximum	Range	IQR
Bottle Fills	19.997	20.079	20.263	0.557	0.139



Excel Output

Bottle Fills	
Mean	20.00817
Standard error	0.008278
Median	19.99697
Mode	#N/A
Standard deviation	0.101388
Sample variance	0.010279
Kurtosis	–0.11422
Skewness	0.080425
Range	0.557666
Minimum	19.70555
Maximum	20.26322
Sum	3001.225
Count	150

Source: Adapted from "Coke, Avis Adjust in Russia," *Advertising Age*, July 5, 1999, p. 25; Coca-Cola Web site at http://www.coca-cola.com/home.html. The Coca-Cola company's 2001 annual report is found at: http://www2.coca-cola.com/investors/annualreport/2001/index.html. The Coca-Cola's Web site for information on Russia is located at: http://www2.cocacola.com/ourcompany/cfs/cfs_include/cfs_russia_include.html.
View Coca-Cola Company's 2007 annual report at: http://www.thecoca-colacompany.com/investors/pdfs/2007_annual_review.pdf

USING THE COMPUTER

EXCEL

- While Excel has the capability of producing many of the statistics in this chapter piecemeal, there is one Excel feature, **Descriptive Statistics**, that produces many of these statistics in one output.
- To use the **Descriptive Statistics** feature, begin by selecting the **Data** tab on the Excel worksheet. From the **Analysis** panel at the right top of the **Data** tab worksheet, click on **Data Analysis**. If your Excel worksheet does not show the **Data Analysis** option, then you can load it as an add-in following directions given in Chapter 2. From the **Data Analysis** pulldown menu, select **Descriptive Statistics**. In the **Descriptive Statistics** dialog box, enter the location of the data to be analyzed in **Input Range**. Check **Labels in the First Row** if your data contains a label in the first row (cell). Check the box beside **Summary statistics.** The **Summary statistics** feature computes a wide variety of descriptive statistics. The output includes the mean, the median, the mode, the standard deviation, the sample variance, a measure of kurtosis, a measure of skewness, the range, the minimum, the maximum, the sum and the count.
- The **Rank and Percentile** feature of the **Data Analysis** tool of Excel has the capability of ordering the data, assigning ranks to the data, and yielding the percentiles of the data. To access this command, click on **Data Analysis** (see above) and select **Rank and Percentile** from the menu. In the **Rank and Percentile** dialog box, enter the location of the data to be analyzed in **Input Range.** Check **Labels in the First Row** if your data contains a label in the first row (cell).
- Many of the individual statistics presented in this chapter can be computed using the **Insert Function** (f_x) of Excel. To access the **Insert Function**, go to the **Formulas** tab on an Excel worksheet (top center tab). The **Insert Function** is on the far left of the menu bar. In the **Insert Function** dialog box at the top, there is a pulldown menu where it says **Or select a category.** From the pulldown menu associated with this command, select **Statistical.** There are 83 different statistics that can be computed using one of these commands. Select the one that you want to compute and enter the location of the data. Some of the more useful commands in this menu are **AVERAGE, MEDIAN, MODE, SKEW, STDEV,** and **VAR.**

MINITAB

- Minitab is capable of performing many of the tasks presented in this chapter, including descriptive statistics and box plots. To begin Descriptive Statistics, select **Stat** on the menu bar, and then from the pulldown menu select **Basic Statistics.** From the **Basic Statistics** menu, select either **Display Descriptive Statistics** or **Graphical Summary.** If you select **DisplayDescriptiveStatistics** in the dialog box that appears, input the column(s) to be analyzed in the box labeled **Variables.** If you click **OK**, then your output will include the sample size, mean, median, standard deviation, minimum, the first quartile, and the third quartile. If in the **Display Descriptive Statistics** dialog box you select the option **Graphs**, you will have several other output options that are relatively self-explanatory. The options include **Histogram of data; Histogram of data, with normal curve; Individual value plot;** and **Boxplot of data.** If in the **Display Descriptive Statistics** dialog box you select the option **Statistics**, you have the option of selecting any of 24 statistics offered to appear in your output.
- On the other hand, you may opt to use the **Graphical Summary** option under **Basic Statistics.** If you use this option, in the dialog box that appears input the column(s) to be analyzed in the box labeled **Variables.** If you click **OK**, then your output will include a histogram graph of the data with the normal curve superimposed, a box plot of the data, the same descriptive statistics listed above for the output from **Display Descriptive Statistics**, along with skewness and kurtosis statistics and other output that pertain to Chapter 8 topics.
- A variety of descriptive statistics can be obtained through the use of **Column Statistics** or **Row Statistics**. To begin, select **Calc** from the menu bar. From the pulldown menu, select either **Column Statistics** or **Row Statistics**, depending on where the data are located. For **Column Statistics**, in the space below **Input variable**, enter the column to be analyzed. For **Row Statistics**, in the space below **Input variables**, enter the rows to be analyzed. Check which **Statistic** you want to compute from **Sum, Mean, Standard deviation, Minimum, Maximum, Range, Median**, and **N total.** Minitab will only allow you to select one statistic at a time.
- Minitab can produce box-and-whisker plots. To begin, select **Graph** from the menu bar, and then select **Boxplot** from the pulldown menu.
- In the Boxplot dialog box, there are four options: **One Y Simple, One Y With Groups, Multiple Y's Simple**, and **Multiple Y's With Groups;** and you must select one.
- After you select one of these types of box plots, another dialog box appears with several options to select, including: **Scale..., Labels..., Data View..., Multiple Graphs...**, and **Data Options...**
- Enter the location of the variables to be graphed in the top box labeled **Graph variables**, and click **OK.**