CHAPTER 3

Descriptive Statistics: Numerical Measures

CONTENTS

STATISTICS *in* PRACTICE

SMALL FRY DESIGN*

SANTA ANA, CALIFORNIA

Founded in 1997, Small Fry Design is a toy and accessory company that designs and imports products for infants. The company's product line includes teddy bears, mobiles, musical toys, rattles, and security blankets and features high-quality soft toy designs with an emphasis on color, texture, and sound. The products are designed in the United States and manufactured in China.

Small Fry Design uses independent representatives to sell the products to infant furnishing retailers, children's accessory and apparel stores, gift shops, upscale department stores, and major catalog companies. Currently, Small Fry Design products are distributed in more than 1000 retail outlets throughout the United States.

Cash flow management is one of the most critical activities in the day-to-day operation of this company. Ensuring sufficient incoming cash to meet both current and ongoing debt obligations can mean the difference between business success and failure. A critical factor in cash flow management is the analysis and control of accounts receivable. By measuring the average age and dollar value of outstanding invoices, management can predict cash availability and monitor changes in the status of accounts receivable. The company set the following goals: the average age for outstanding invoices should not exceed 45 days, and the dollar value of invoices more than 60 days old should not exceed 5% of the dollar value of all accounts receivable.

In a recent summary of accounts receivable status, the following descriptive statistics were provided for the age of outstanding invoices:

Mean	40 days
Median	35 days
Mode	31 days

*The authors are indebted to John A. McCarthy, President of Small Fry Design, for providing this Statistics in Practice.

Small Fry Design's "King of the Jungle" mobile.
© Joe-Higgins/South-Western.

Interpretation of these statistics shows that the mean or average age of an invoice is 40 days. The median shows that half of the invoices remain outstanding 35 days or more. The mode of 31 days, the most frequent invoice age, indicates that the most common length of time an invoice is outstanding is 31 days. The statistical summary also showed that only 3% of the dollar value of all accounts receivable was more than 60 days old. Based on the statistical information, management was satisfied that accounts receivable and incoming cash flow were under control.

In this chapter, you will learn how to compute and interpret some of the statistical measures used by Small Fry Design. In addition to the mean, median, and mode, you will learn about other descriptive statistics such as the range, variance, standard deviation, percentiles, and correlation. These numerical measures will assist in the understanding and interpretation of data.

In Chapter 2 we discussed tabular and graphical presentations used to summarize data. In this chapter, we present several numerical measures that provide additional alternatives for summarizing data.

We start by developing numerical summary measures for data sets consisting of a single variable. When a data set contains more than one variable, the same numerical measures can be computed separately for each variable. However, in the two-variable case, we will also develop measures of the relationship between the variables.

Numerical measures of location, dispersion, shape, and association are introduced. If the measures are computed for data from a sample, they are called **sample statistics**. If the measures are computed for data from a population, they are called **population parameters**. In statistical inference, a sample statistic is referred to as the **point estimator** of the corresponding population parameter. In Chapter 7 we will discuss in more detail the process of point estimation.

In the three chapter appendixes we show how Minitab, Excel, and StatTools can be used to compute the numerical measures described in the chapter.

3.1 Measures of Location

Mean

Perhaps the most important measure of location is the **mean**, or average value, for a variable. The mean provides a measure of central location for the data. If the data are for a sample, the mean is denoted by $\bar{x}$; if the data are for a population, the mean is denoted by the Greek letter μ.

In statistical formulas, it is customary to denote the value of variable x for the first observation by x_1, the value of variable x for the second observation by x_2, and so on. In general, the value of variable x for the ith observation is denoted by x_i. For a sample with n observations, the formula for the sample mean is as follows.

The sample mean $\bar{x}$ is a sample statistic.

SAMPLE MEAN

$$\bar{x} = \frac{\Sigma x_i}{n} \tag{3.1}$$

In the preceding formula, the numerator is the sum of the values of the n observations. That is,

$$\Sigma x_i = x_1 + x_2 + \cdots + x_n$$

The Greek letter Σ is the summation sign.

To illustrate the computation of a sample mean, let us consider the following class size data for a sample of five college classes.

46 54 42 46 32

We use the notation x_1, x_2, x_3, x_4, x_5 to represent the number of students in each of the five classes.

$$x_1 = 46 \qquad x_2 = 54 \qquad x_3 = 42 \qquad x_4 = 46 \qquad x_5 = 32$$

Hence, to compute the sample mean, we can write

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{46 + 54 + 42 + 46 + 32}{5} = 44$$

The sample mean class size is 44 students.

Another illustration of the computation of a sample mean is given in the following situation. Suppose that a college placement office sent a questionnaire to a sample of business school graduates requesting information on monthly starting salaries. Table 3.1 shows the

TABLE 3.1 MONTHLY STARTING SALARIES FOR A SAMPLE OF 12 BUSINESS SCHOOL GRADUATES

WEB file
StartSalary

Graduate	Monthly Starting Salary ($)	Graduate	Monthly Starting Salary ($)
1	3450	7	3490
2	3550	8	3730
3	3650	9	3540
4	3480	10	3925
5	3355	11	3520
6	3310	12	3480

collected data. The mean monthly starting salary for the sample of 12 business college graduates is computed as

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_{12}}{12}$$
$$= \frac{3450 + 3550 + \cdots + 3480}{12}$$
$$= \frac{42{,}480}{12} = 3540$$

Equation (3.1) shows how the mean is computed for a sample with n observations. The formula for computing the mean of a population remains the same, but we use different notation to indicate that we are working with the entire population. The number of observations in a population is denoted by N and the symbol for a population mean is μ.

The sample mean $\bar{x}$ is a point estimator of the population mean μ.

POPULATION MEAN

$$\mu = \frac{\Sigma x_i}{N} \qquad (3.2)$$

Median

The **median** is another measure of central location. The median is the value in the middle when the data are arranged in ascending order (smallest value to largest value). With an odd number of observations, the median is the middle value. An even number of observations has no single middle value. In this case, we follow convention and define the median as the average of the values for the middle two observations. For convenience the definition of the median is restated as follows.

MEDIAN

Arrange the data in ascending order (smallest value to largest value).

(a) For an odd number of observations, the median is the middle value.
(b) For an even number of observations, the median is the average of the two middle values.

Let us apply this definition to compute the median class size for the sample of five college classes. Arranging the data in ascending order provides the following list.

32 42 46 46 54

Because $n = 5$ is odd, the median is the middle value. Thus the median class size is 46 students. Even though this data set contains two observations with values of 46, each observation is treated separately when we arrange the data in ascending order.

Suppose we also compute the median starting salary for the 12 business college graduates in Table 3.1. We first arrange the data in ascending order.

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Middle Two Values

Because $n = 12$ is even, we identify the middle two values: 3490 and 3520. The median is the average of these values.

$$\text{Median} = \frac{3490 + 3520}{2} = 3505$$

The median is the measure of location most often reported for annual income and property value data because a few extremely large incomes or property values can inflate the mean. In such cases, the median is the preferred measure of central location.

Although the mean is the more commonly used measure of central location, in some situations the median is preferred. The mean is influenced by extremely small and large data values. For instance, suppose that one of the graduates (see Table 3.1) had a starting salary of $10,000 per month (maybe the individual's family owns the company). If we change the highest monthly starting salary in Table 3.1 from $3925 to $10,000 and recompute the mean, the sample mean changes from $3540 to $4046. The median of $3505, however, is unchanged, because $3490 and $3520 are still the middle two values. With the extremely high starting salary included, the median provides a better measure of central location than the mean. We can generalize to say that whenever a data set contains extreme values, the median is often the preferred measure of central location.

Mode

A third measure of location is the **mode**. The mode is defined as follows.

MODE

The mode is the value that occurs with greatest frequency.

To illustrate the identification of the mode, consider the sample of five class sizes. The only value that occurs more than once is 46. Because this value, occurring with a frequency of 2, has the greatest frequency, it is the mode. As another illustration, consider the sample of starting salaries for the business school graduates. The only monthly starting salary that occurs more than once is $3480. Because this value has the greatest frequency, it is the mode.

Situations can arise for which the greatest frequency occurs at two or more different values. In these instances more than one mode exists. If the data contain exactly two modes, we say that the data are *bimodal*. If data contain more than two modes, we say that the data are *multimodal*. In multimodal cases the mode is almost never reported because listing three or more modes would not be particularly helpful in describing a location for the data.

Percentiles

A **percentile** provides information about how the data are spread over the interval from the smallest value to the largest value. For data that do not contain numerous repeated values, the pth percentile divides the data into two parts. Approximately p percent of the observations have values less than the pth percentile; approximately $(100 - p)$ percent of the observations have values greater than the pth percentile. The pth percentile is formally defined as follows.

PERCENTILE

The pth percentile is a value such that *at least* p percent of the observations are less than or equal to this value and *at least* $(100 - p)$ percent of the observations are greater than or equal to this value.

Colleges and universities frequently report admission test scores in terms of percentiles. For instance, suppose an applicant obtains a raw score of 54 on the verbal portion of an admission test. How this student performed in relation to other students taking the same test may not be readily apparent. However, if the raw score of 54 corresponds to the 70th percentile, we know that approximately 70% of the students scored lower than this individual and approximately 30% of the students scored higher than this individual.

The following procedure can be used to compute the pth percentile.

CALCULATING THE pTH PERCENTILE

Following these steps makes it easy to calculate percentiles.

Step 1. Arrange the data in ascending order (smallest value to largest value).

Step 2. Compute an index i

$$i = \left(\frac{p}{100}\right)n$$

where p is the percentile of interest and n is the number of observations.

Step 3. (a) If *i is not an integer, round up.* The next integer *greater* than i denotes the position of the pth percentile.
(b) If *i is an integer,* the pth percentile is the average of the values in positions i and $i + 1$.

As an illustration of this procedure, let us determine the 85th percentile for the starting salary data in Table 3.1.

Step 1. Arrange the data in ascending order.

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Step 2.

$$i = \left(\frac{p}{100}\right)n = \left(\frac{85}{100}\right)12 = 10.2$$

Step 3. Because i is not an integer, *round up.* The position of the 85th percentile is the next integer greater than 10.2, the 11th position.

Returning to the data, we see that the 85th percentile is the data value in the 11th position, or 3730.

As another illustration of this procedure, let us consider the calculation of the 50th percentile for the starting salary data. Applying step 2, we obtain

$$i = \left(\frac{50}{100}\right)12 = 6$$

Because i is an integer, step 3(b) states that the 50th percentile is the average of the sixth and seventh data values; thus the 50th percentile is (3490 + 3520)/2 = 3505. Note that the *50th percentile is also the median.*

Quartiles

Quartiles are just specific percentiles; thus, the steps for computing percentiles can be applied directly in the computation of quartiles.

It is often desirable to divide data into four parts, with each part containing approximately one-fourth, or 25% of the observations. Figure 3.1 shows a data distribution divided into four parts. The division points are referred to as the **quartiles** and are defined as

Q_1 = first quartile, or 25th percentile
Q_2 = second quartile, or 50th percentile (also the median)
Q_3 = third quartile, or 75th percentile.

The starting salary data are again arranged in ascending order. We already identified Q_2, the second quartile (median), as 3505.

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

The computations of quartiles Q_1 and Q_3 require the use of the rule for finding the 25th and 75th percentiles. These calculations follow.

For Q_1,

$$i = \left(\frac{p}{100}\right)n = \left(\frac{25}{100}\right)12 = 3$$

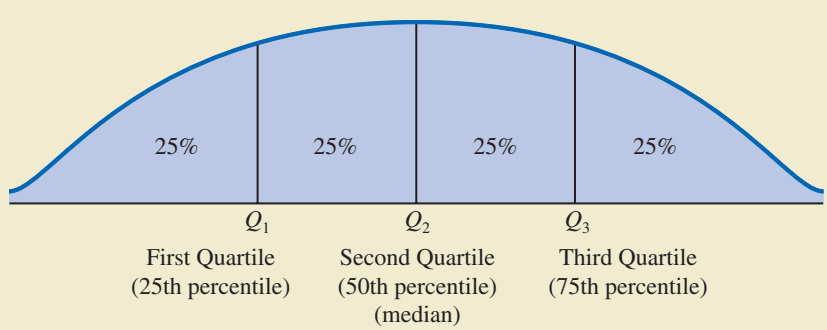
Because i is an integer, step 3(b) indicates that the first quartile, or 25th percentile, is the average of the third and fourth data values; thus, Q_1 = (3450 + 3480)/2 = 3465.

For Q_3,

$$i = \left(\frac{p}{100}\right)n = \left(\frac{75}{100}\right)12 = 9$$

Again, because i is an integer, step 3(b) indicates that the third quartile, or 75th percentile, is the average of the ninth and tenth data values; thus, Q_3 = (3550 + 3650)/2 = 3600.

FIGURE 3.1 LOCATION OF THE QUARTILES



The quartiles divide the starting salary data into four parts, with each part containing 25% of the observations.

3310	3355	3450	3480	3480	3490	3520	3540	3550	3650	3730	3925
		$Q_1 = 3465$			$Q_2 = 3505$ (Median)			$Q_3 = 3600$			

We defined the quartiles as the 25th, 50th, and 75th percentiles. Thus, we computed the quartiles in the same way as percentiles. However, other conventions are sometimes used to compute quartiles, and the actual values reported for quartiles may vary slightly depending on the convention used. Nevertheless, the objective of all procedures for computing quartiles is to divide the data into four equal parts.

NOTES AND COMMENTS

It is better to use the median than the mean as a measure of central location when a data set contains extreme values. Another measure, sometimes used when extreme values are present, is the *trimmed mean.* It is obtained by deleting a percentage of the smallest and largest values from a data set and then computing the mean of the remaining values. For example, the 5% trimmed mean is obtained by removing the smallest 5% and the largest 5% of the data values and then computing the mean of the remaining values. Using the sample with $n = 12$ starting salaries, $0.05(12) = 0.6$. Rounding this value to 1 indicates that the 5% trimmed mean would remove the 1 smallest data value and the 1 largest data value. The 5% trimmed mean using the 10 remaining observations is 3524.50.

Exercises

Methods

1. Consider a sample with data values of 10, 20, 12, 17, and 16. Compute the mean and median.
2. Consider a sample with data values of 10, 20, 21, 17, 16, and 12. Compute the mean and median.
3. Consider a sample with data values of 27, 25, 20, 15, 30, 34, 28, and 25. Compute the 20th, 25th, 65th, and 75th percentiles.
4. Consider a sample with data values of 53, 55, 70, 58, 64, 57, 53, 69, 57, 68, and 53. Compute the mean, median, and mode.

SELF test

Applications

5. The Dow Jones Travel Index reported what business travelers pay for hotel rooms per night in major U.S. cities (*The Wall Street Journal,* January 16, 2004). The average hotel room rates for 20 cities are as follows:

WEB file
Hotels

Atlanta	$163	Minneapolis	$125
Boston	177	New Orleans	167
Chicago	166	New York	245
Cleveland	126	Orlando	146
Dallas	123	Phoenix	139
Denver	120	Pittsburgh	134
Detroit	144	San Francisco	167
Houston	173	Seattle	162
Los Angeles	160	St. Louis	145
Miami	192	Washington, D.C.	207

a. What is the mean hotel room rate?
b. What is the median hotel room rate?
c. What is the mode?
d. What is the first quartile?
e. What is the third quartile?

6. During the 2007–2008 NCAA college basketball season, men's basketball teams attempted an all-time high number of 3-point shots, averaging 19.07 shots per game (Associated Press Sports, January 24, 2009). In an attempt to discourage so many 3-point shots and encourage more inside play, the NCAA rules committee moved the 3-point line back from 19 feet, 9 inches to 20 feet, 9 inches at the beginning of the 2008–2009 basketball season. Shown in the following table are the 3-point shots taken and the 3-point shots made for a sample of 19 NCAA basketball games during the 2008–2009 season.

WEB file
3Points

3-Point Shots	Shots Made	3-Point Shots	Shots Made
23	4	17	7
20	6	19	10
17	5	22	7
18	8	25	11
13	4	15	6
16	4	10	5
8	5	11	3
19	8	25	8
28	5	23	7
21	7		

a. What is the mean number of 3-point shots taken per game?
b. What is the mean number of 3-point shots made per game?
c. Using the closer 3-point line, players were making 35.2% of their shots. What percentage of shots were players making from the new 3-point line?
d. What was the impact of the NCAA rules change that moved the 3-point line back to 20 feet, 9 inches for the 2008–2009 season? Would you agree with the Associated Press Sports article that stated, "Moving back the 3-point line hasn't changed the game dramatically"? Explain.

7. Endowment income is a critical part of the annual budgets at colleges and universities. A study by the National Association of College and University Business Officers reported that the 435 colleges and universities surveyed held a total of $413 billion in endowments. The 10 wealthiest universities are shown below (*The Wall Street Journal,* January 27, 2009). Amounts are in billion of dollars.

University	Endowment ($billion)	University	Endowment ($billion)
Columbia	7.2	Princeton	16.4
Harvard	36.6	Stanford	17.2
M.I.T.	10.1	Texas	16.1
Michigan	7.6	Texas A&M	6.7
Northwestern	7.2	Yale	22.9

a. What is the mean endowment for these universities?
b. What is the median endowment?
c. What is the mode endowment?
d. Compute the first and third quartiles?

e. What is the total endowment at these 10 universities? These universities represent 2.3% of the 435 colleges and universities surveyed. What percentage of the total $413 billion in endowments is held by these 10 universities?

f. *The Wall Street Journal* reported that over a recent five-month period, a downturn in the economy has caused endowments to decline 23%. What is the estimate of the dollar amount of the decline in the total endowments held by these 10 universities? Given this situation, what are some of the steps you would expect university administrators to be considering?

SELF test

8. The cost of consumer purchases such as single-family housing, gasoline, Internet services, tax preparation, and hospitalization were provided in *The Wall-Street Journal* (January 2, 2007). Sample data typical of the cost of tax-return preparation by services such as H&R Block are shown below.

WEB file
TaxCost

120	230	110	115	160
130	150	105	195	155
105	360	120	120	140
100	115	180	235	255

a. Compute the mean, median, and mode.

b. Compute the first and third quartiles.

c. Compute and interpret the 90th percentile.

9. The National Association of Realtors provided data showing that home sales were the slowest in 10 years (Associated Press, December 24, 2008). Sample data with representative sales prices for existing homes and new homes follow. Data are in thousands of dollars:

Existing Homes	315.5	202.5	140.2	181.3	470.2	169.9	112.8	230.0	177.5
New Homes	275.9	350.2	195.8	525.0	225.3	215.5	175.0	149.5	

a. What is the median sales price for existing homes?

b. What is the median sales price for new homes?

c. Do existing homes or new homes have the higher median sales price? What is the difference between the median sales prices?

d. A year earlier the median sales price for existing homes was $208.4 thousand and the median sales price for new homes was $249 thousand. Compute the percentage change in the median sales price of existing and new homes over the one-year period. Did existing homes or new homes have the larger percentage change in median sales price?

10. A panel of economists provided forecasts of the U.S. economy for the first six months of 2007 (*The Wall Street Journal,* January 2, 2007). The percent changes in the gross domestic product (GDP) forecasted by 30 economists are as follows.

WEB file
Economy

2.6	3.1	2.3	2.7	3.4	0.9	2.6	2.8	2.0	2.4
2.7	2.7	2.7	2.9	3.1	2.8	1.7	2.3	2.8	3.5
0.4	2.5	2.2	1.9	1.8	1.1	2.0	2.1	2.5	0.5

a. What is the minimum forecast for the percent change in the GDP? What is the maximum?

b. Compute the mean, median, and mode.

c. Compute the first and third quartiles.

d. Did the economists provide an optimistic or pessimistic outlook for the U.S. economy? Discuss.

11. In automobile mileage and gasoline-consumption testing, 13 automobiles were road tested for 300 miles in both city and highway driving conditions. The following data were recorded for miles-per-gallon performance.

City:	16.2	16.7	15.9	14.4	13.2	15.3	16.8	16.0	16.1	15.3	15.2	15.3	16.2
Highway:	19.4	20.6	18.3	18.6	19.2	17.4	17.2	18.6	19.0	21.1	19.4	18.5	18.7

Use the mean, median, and mode to make a statement about the difference in performance for city and highway driving.

12. Walt Disney Company bought Pixar Animation Studios, Inc., in a deal worth $7.4 billion (CNN Money website, January 24, 2006). The animated movies produced by Disney and Pixar during the previous 10 years are listed in the following table. The box office revenues are in millions of dollars. Compute the total revenue, the mean, the median, and the quartiles to compare the box office success of the movies produced by both companies. Do the statistics suggest at least one of the reasons Disney was interested in buying Pixar? Discuss.

WEB file
Disney

Disney Movies	Revenue ($millions)	Pixar Movies	Revenue ($millions)
Pocahontas	346	*Toy Story*	362
Hunchback of Notre Dame	325	*A Bug's Life*	363
Hercules	253	*Toy Story 2*	485
Mulan	304	*Monsters, Inc.*	525
Tarzan	448	*Finding Nemo*	865
Dinosaur	354	*The Incredibles*	631
The Emperor's New Groove	169		
Lilo & Stitch	273		
Treasure Planet	110		
The Jungle Book 2	136		
Brother Bear	250		
Home on the Range	104		
Chicken Little	249		

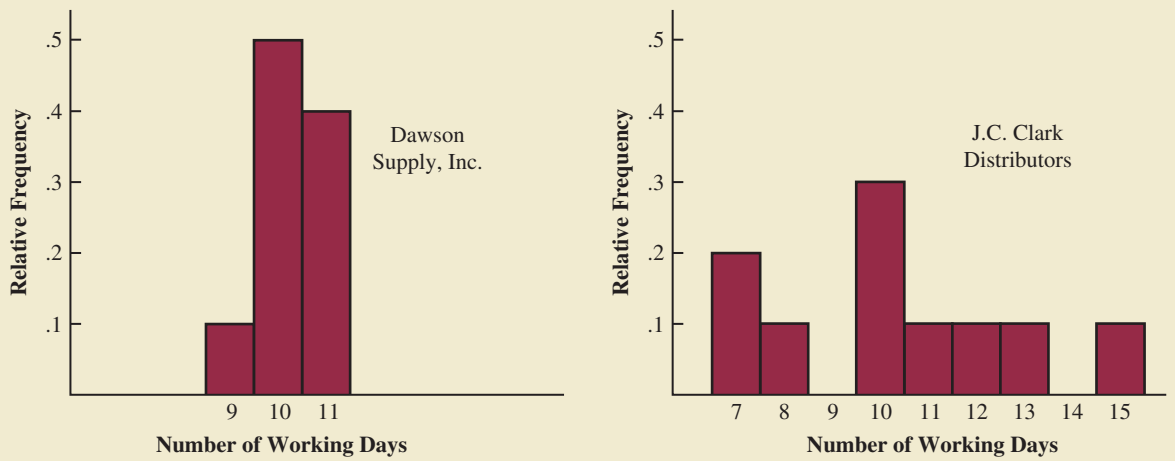
3.2 Measures of Variability

The variability in the delivery time creates uncertainty for production scheduling. Methods in this section help measure and understand variability.

In addition to measures of location, it is often desirable to consider measures of variability, or dispersion. For example, suppose that you are a purchasing agent for a large manufacturing firm and that you regularly place orders with two different suppliers. After several months of operation, you find that the mean number of days required to fill orders is 10 days for both of the suppliers. The histograms summarizing the number of working days required to fill orders from the suppliers are shown in Figure 3.2. Although the mean number of days is 10 for both suppliers, do the two suppliers demonstrate the same degree of reliability in terms of making deliveries on schedule? Note the dispersion, or variability, in delivery times indicated by the histograms. Which supplier would you prefer?

For most firms, receiving materials and supplies on schedule is important. The 7- or 8-day deliveries shown for J.C. Clark Distributors might be viewed favorably; however, a few of the slow 13- to 15-day deliveries could be disastrous in terms of keeping a workforce busy

FIGURE 3.2 HISTORICAL DATA SHOWING THE NUMBER OF DAYS REQUIRED TO FILL ORDERS



and production on schedule. This example illustrates a situation in which the variability in the delivery times may be an overriding consideration in selecting a supplier. For most purchasing agents, the lower variability shown for Dawson Supply, Inc., would make Dawson the preferred supplier.

We turn now to a discussion of some commonly used measures of variability.

Range

The simplest measure of variability is the **range**.

RANGE

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

Let us refer to the data on starting salaries for business school graduates in Table 3.1. The largest starting salary is 3925 and the smallest is 3310. The range is $3925 - 3310 = 615$.

Although the range is the easiest of the measures of variability to compute, it is seldom used as the only measure. The reason is that the range is based on only two of the observations and thus is highly influenced by extreme values. Suppose one of the graduates received a starting salary of $10,000 per month. In this case, the range would be $10{,}000 - 3310 = 6690$ rather than 615. This large value for the range would not be especially descriptive of the variability in the data because 11 of the 12 starting salaries are closely grouped between 3310 and 3730.

Interquartile Range

A measure of variability that overcomes the dependency on extreme values is the **interquartile range (IQR)**. This measure of variability is the difference between the third quartile, Q_3, and the first quartile, Q_1. In other words, the interquartile range is the range for the middle 50% of the data.

INTERQUARTILE RANGE

$$\text{IQR} = Q_3 - Q_1 \tag{3.3}$$

For the data on monthly starting salaries, the quartiles are $Q_3 = 3600$ and $Q_1 = 3465$. Thus, the interquartile range is $3600 - 3465 = 135$.

Variance

The **variance** is a measure of variability that utilizes all the data. The variance is based on the difference between the value of each observation (x_i) and the mean. The difference between each x_i and the mean ($\bar{x}$ for a sample, μ for a population) is called a *deviation about the mean.* For a sample, a deviation about the mean is written ($x_i - \bar{x}$); for a population, it is written ($x_i - \mu$). In the computation of the variance, the deviations about the mean are *squared.*

If the data are for a population, the average of the squared deviations is called the *population variance.* The population variance is denoted by the Greek symbol σ^2. For a population of N observations and with μ denoting the population mean, the definition of the population variance is as follows.

POPULATION VARIANCE

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \tag{3.4}$$

In most statistical applications, the data being analyzed are for a sample. When we compute a sample variance, we are often interested in using it to estimate the population variance σ^2. Although a detailed explanation is beyond the scope of this text, it can be shown that if the sum of the squared deviations about the sample mean is divided by $n - 1$, and not n, the resulting sample variance provides an unbiased estimate of the population variance. For this reason, the *sample variance,* denoted by s^2, is defined as follows.

The sample variance s^2 is the estimator of the population variance σ^2.

SAMPLE VARIANCE

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \tag{3.5}$$

To illustrate the computation of the sample variance, we will use the data on class size for the sample of five college classes as presented in Section 3.1. A summary of the data, including the computation of the deviations about the mean and the squared deviations about the mean, is shown in Table 3.2. The sum of squared deviations about the mean is $\Sigma(x_i - \bar{x})^2 = 256$. Hence, with $n - 1 = 4$, the sample variance is

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{256}{4} = 64$$

Before moving on, let us note that the units associated with the sample variance often cause confusion. Because the values being summed in the variance calculation, $(x_i - \bar{x})^2$, are squared, the units associated with the sample variance are also *squared.* For instance, the

TABLE 3.2 COMPUTATION OF DEVIATIONS AND SQUARED DEVIATIONS ABOUT THE MEAN FOR THE CLASS SIZE DATA

Number of Students in Class (x_i)	Mean Class Size ($\bar{x}$)	Deviation About the Mean ($x_i - \bar{x}$)	Squared Deviation About the Mean $(x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	−12	144
		0	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

The variance is useful in comparing the variability of two or more variables.

sample variance for the class size data is $s^2 = 64$ (students)^2. The squared units associated with variance make it difficult to obtain an intuitive understanding and interpretation of the numerical value of the variance. We recommend that you think of the variance as a measure useful in comparing the amount of variability for two or more variables. In a comparison of the variables, the one with the largest variance shows the most variability. Further interpretation of the value of the variance may not be necessary.

As another illustration of computing a sample variance, consider the starting salaries listed in Table 3.1 for the 12 business school graduates. In Section 3.1, we showed that the sample mean starting salary was 3540. The computation of the sample variance ($s^2 = 27{,}440.91$) is shown in Table 3.3.

TABLE 3.3 COMPUTATION OF THE SAMPLE VARIANCE FOR THE STARTING SALARY DATA

Monthly Salary (x_i)	Sample Mean ($\bar{x}$)	Deviation About the Mean ($x_i - \bar{x}$)	Squared Deviation About the Mean $(x_i - \bar{x})^2$
3450	3540	−90	8,100
3550	3540	10	100
3650	3540	110	12,100
3480	3540	−60	3,600
3355	3540	−185	34,225
3310	3540	−230	52,900
3490	3540	−50	2,500
3730	3540	190	36,100
3540	3540	0	0
3925	3540	385	148,225
3520	3540	−20	400
3480	3540	−60	3,600
		0	301,850
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

Using equation (3.5),

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{301{,}850}{11} = 27{,}440.91$$

In Tables 3.2 and 3.3 we show both the sum of the deviations about the mean and the sum of the squared deviations about the mean. For any data set, the sum of the deviations about the mean will *always equal zero.* Note that in Tables 3.2 and 3.3, $\Sigma(x_i - \bar{x}) = 0$. The positive deviations and negative deviations cancel each other, causing the sum of the deviations about the mean to equal zero.

Standard Deviation

The **standard deviation** is defined to be the positive square root of the variance. Following the notation we adopted for a sample variance and a population variance, we use s to denote the sample standard deviation and σ to denote the population standard deviation. The standard deviation is derived from the variance in the following way.

The sample standard deviation s is the estimator of the population standard deviation σ.

STANDARD DEVIATION

$$\text{Sample standard deviation} = s = \sqrt{s^2} \tag{3.6}$$

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} \tag{3.7}$$

Recall that the sample variance for the sample of class sizes in five college classes is $s^2 = 64$. Thus, the sample standard deviation is $s = \sqrt{64} = 8$. For the data on starting salaries, the sample standard deviation is $s = \sqrt{27{,}440.91} = 165.65$.

The standard deviation is easier to interpret than the variance because the standard deviation is measured in the same units as the data.

What is gained by converting the variance to its corresponding standard deviation? Recall that the units associated with the variance are squared. For example, the sample variance for the starting salary data of business school graduates is $s^2 = 27{,}440.91$ (dollars)^2. Because the standard deviation is the square root of the variance, the units of the variance, dollars squared, are converted to dollars in the standard deviation. Thus, the standard deviation of the starting salary data is \$165.65. In other words, the standard deviation is measured in the same units as the original data. For this reason the standard deviation is more easily compared to the mean and other statistics that are measured in the same units as the original data.

Coefficient of Variation

The coefficient of variation is a relative measure of variability; it measures the standard deviation relative to the mean.

In some situations we may be interested in a descriptive statistic that indicates how large the standard deviation is relative to the mean. This measure is called the **coefficient of variation** and is usually expressed as a percentage.

COEFFICIENT OF VARIATION

$$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\% \tag{3.8}$$

For the class size data, we found a sample mean of 44 and a sample standard deviation of 8. The coefficient of variation is $[(8/44) \times 100]\% = 18.2\%$. In words, the coefficient of variation tells us that the sample standard deviation is 18.2% of the value of the sample mean. For the starting salary data with a sample mean of 3540 and a sample standard deviation of 165.65, the coefficient of variation, $[(165.65/3540) \times 100]\% = 4.7\%$, tells us the sample standard deviation is only 4.7% of the value of the sample mean. In general, the coefficient of variation is a useful statistic for comparing the variability of variables that have different standard deviations and different means.

NOTES AND COMMENTS

1. Statistical software packages and spreadsheets can be used to develop the descriptive statistics presented in this chapter. After the data are entered into a worksheet, a few simple commands can be used to generate the desired output. In three chapter-ending appendixes we show how Minitab, Excel, and StatTools can be used to develop descriptive statistics.
2. The standard deviation is a commonly used measure of the risk associated with investing in stock and stock funds (*BusinessWeek,* January 17, 2000). It provides a measure of how monthly returns fluctuate around the long-run average return.
3. Rounding the value of the sample mean $\bar{x}$ and the values of the squared deviations $(x_i - \bar{x})^2$ may introduce errors when a calculator is used in the computation of the variance and standard deviation. To reduce rounding errors, we recommend carrying at least six significant digits during intermediate calculations. The resulting variance or standard deviation can then be rounded to fewer digits.
4. An alternative formula for the computation of the sample variance is

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

where $\Sigma x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2$.

Exercises

Methods

13. Consider a sample with data values of 10, 20, 12, 17, and 16. Compute the range and interquartile range.
14. Consider a sample with data values of 10, 20, 12, 17, and 16. Compute the variance and standard deviation.
15. Consider a sample with data values of 27, 25, 20, 15, 30, 34, 28, and 25. Compute the range, interquartile range, variance, and standard deviation.

SELF test

Applications

SELF test

16. A bowler's scores for six games were 182, 168, 184, 190, 170, and 174. Using these data as a sample, compute the following descriptive statistics:
    a. Range
    b. Variance
    c. Standard deviation
    d. Coefficient of variation
17. A home theater in a box is the easiest and cheapest way to provide surround sound for a home entertainment center. A sample of prices is shown here (*Consumer Reports Buying Guide,* 2004). The prices are for models with a DVD player and for models without a DVD player.

Models with DVD Player	Price	Models without DVD Player	Price
Sony HT-1800DP	$450	Pioneer HTP-230	$300
Pioneer HTD-330DV	300	Sony HT-DDW750	300
Sony HT-C800DP	400	Kenwood HTB-306	360
Panasonic SC-HT900	500	RCA RT-2600	290
Panasonic SC-MTI	400	Kenwood HTB-206	300

    a. Compute the mean price for models with a DVD player and the mean price for models without a DVD player. What is the additional price paid to have a DVD player included in a home theater unit?
    b. Compute the range, variance, and standard deviation for the two samples. What does this information tell you about the prices for models with and without a DVD player?

18. Car rental rates per day for a sample of seven Eastern U.S. cities are as follows (*The Wall Street Journal,* January 16, 2004).

City	Daily Rate
Boston	$43
Atlanta	35
Miami	34
New York	58
Orlando	30
Pittsburgh	30
Washington, D.C.	36

a. Compute the mean, variance, and standard deviation for the car rental rates.
b. A similar sample of seven Western U.S. cities showed a sample mean car rental rate of $38 per day. The variance and standard deviation were 12.3 and 3.5, respectively. Discuss any difference between the car rental rates in Eastern and Western U.S. cities.

19. The *Los Angeles Times* regularly reports the air quality index for various areas of Southern California. A sample of air quality index values for Pomona provided the following data: 28, 42, 58, 48, 45, 55, 60, 49, and 50.
a. Compute the range and interquartile range.
b. Compute the sample variance and sample standard deviation.
c. A sample of air quality index readings for Anaheim provided a sample mean of 48.5, a sample variance of 136, and a sample standard deviation of 11.66. What comparisons can you make between the air quality in Pomona and that in Anaheim on the basis of these descriptive statistics?

20. The following data were used to construct the histograms of the number of days required to fill orders for Dawson Supply, Inc., and J.C. Clark Distributors (see Figure 3.2).

Dawson Supply Days for Delivery:	11	10	9	10	11	11	10	11	10	10
Clark Distributors Days for Delivery:	8	10	13	7	10	11	10	7	15	12

Use the range and standard deviation to support the previous observation that Dawson Supply provides the more consistent and reliable delivery times.

21. How do grocery costs compare across the country? Using a market basket of 10 items including meat, milk, bread, eggs, coffee, potatoes, cereal, and orange juice, *Where to Retire* magazine calculated the cost of the market basket in six cities and in six retirement areas across the country (*Where to Retire,* November/December 2003). The data with market basket cost to the nearest dollar are as follows:

City	Cost	Retirement Area	Cost
Buffalo, NY	$33	Biloxi-Gulfport, MS	$29
Des Moines, IA	27	Asheville, NC	32
Hartford, CT	32	Flagstaff, AZ	32
Los Angeles, CA	38	Hilton Head, SC	34
Miami, FL	36	Fort Myers, FL	34
Pittsburgh, PA	32	Santa Fe, NM	31

a. Compute the mean, variance, and standard deviation for the sample of cities and the sample of retirement areas.
b. What observations can be made based on the two samples?

WEB file
BackToSchool

22. The National Retail Federation reported that college freshman spend more on back-to-school items than any other college group (*USA Today,* August 4, 2006). Sample data comparing the back-to-school expenditures for 25 freshmen and 20 seniors are shown in the data file BackToSchool.
    a. What is the mean back-to-school expenditure for each group? Are the data consistent with the National Retail Federation's report?
    b. What is the range for the expenditures in each group?
    c. What is the interquartile range for the expenditures in each group?
    d. What is the standard deviation for expenditures in each group?
    e. Do freshmen or seniors have more variation in back-to-school expenditures?

23. Scores turned in by an amateur golfer at the Bonita Fairways Golf Course in Bonita Springs, Florida, during 2005 and 2006 are as follows:

2005 Season:	74	78	79	77	75	73	75	77
2006 Season:	71	70	75	77	85	80	71	79

    a. Use the mean and standard deviation to evaluate the golfer's performance over the two-year period.
    b. What is the primary difference in performance between 2005 and 2006? What improvement, if any, can be seen in the 2006 scores?

24. The following times were recorded by the quarter-mile and mile runners of a university track team (times are in minutes).

Quarter-Mile Times:	.92	.98	1.04	.90	.99
Mile Times:	4.52	4.35	4.60	4.70	4.50

After viewing this sample of running times, one of the coaches commented that the quarter-milers turned in the more consistent times. Use the standard deviation and the coefficient of variation to summarize the variability in the data. Does the use of the coefficient of variation indicate that the coach's statement should be qualified?

3.3 Measures of Distribution Shape, Relative Location, and Detecting Outliers

We have described several measures of location and variability for data. In addition, it is often important to have a measure of the shape of a distribution. In Chapter 2 we noted that a histogram provides a graphical display showing the shape of a distribution. An important numerical measure of the shape of a distribution is called **skewness**.

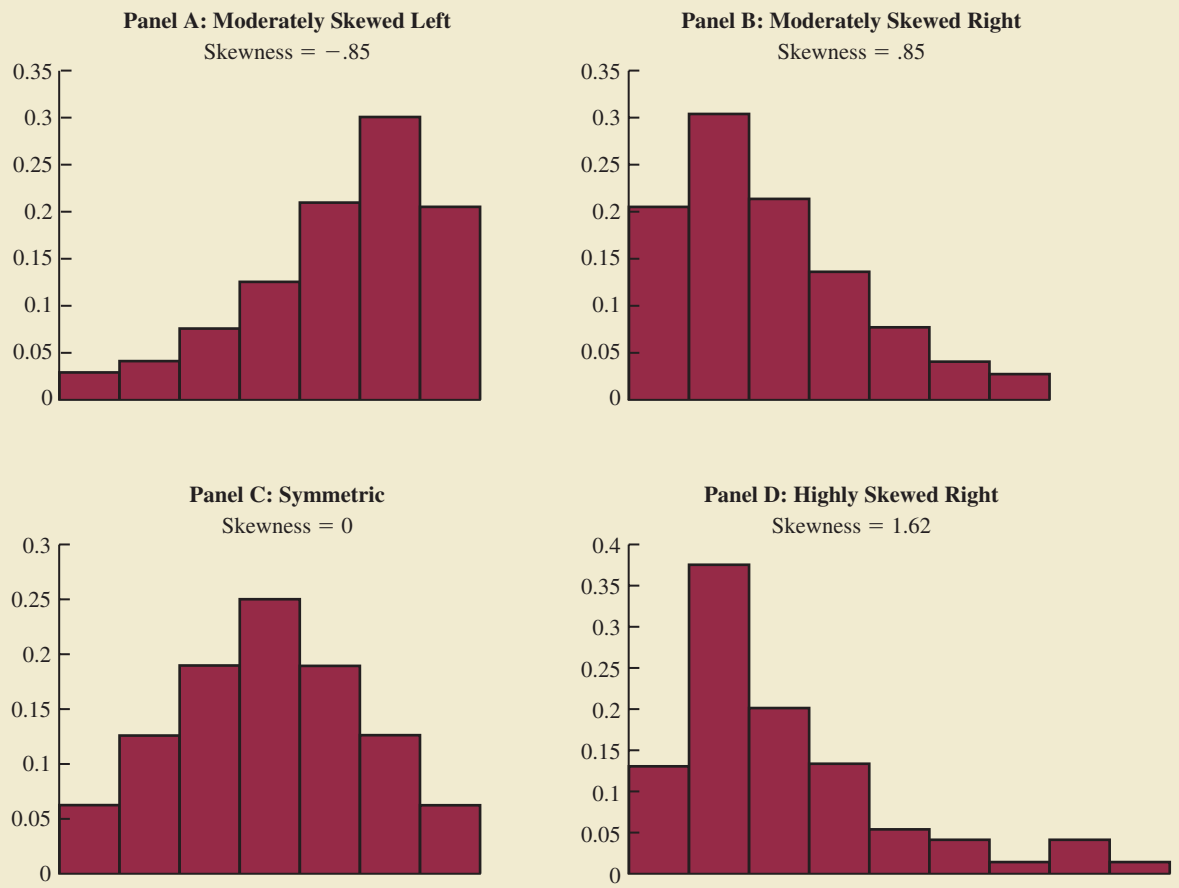
Distribution Shape

Shown in Figure 3.3 are four histograms constructed from relative frequency distributions. The histograms in Panels A and B are moderately skewed. The one in Panel A is skewed to the left; its skewness is −.85. The histogram in Panel B is skewed to the right; its skewness is +.85. The histogram in Panel C is symmetric; its skewness is zero. The histogram in Panel D is highly skewed to the right; its skewness is 1.62. The formula used to compute skewness is somewhat complex.[1] However, the skewness can be easily

[1]The formula for the skewness of sample data:

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

FIGURE 3.3 HISTOGRAMS SHOWING THE SKEWNESS FOR FOUR DISTRIBUTIONS



computed using statistical software. For data skewed to the left, the skewness is negative; for data skewed to the right, the skewness is positive. If the data are symmetric, the skewness is zero.

For a symmetric distribution, the mean and the median are equal. When the data are positively skewed, the mean will usually be greater than the median; when the data are negatively skewed, the mean will usually be less than the median. The data used to construct the histogram in Panel D are customer purchases at a women's apparel store. The mean purchase amount is $77.60 and the median purchase amount is $59.70. The relatively few large purchase amounts tend to increase the mean, while the median remains unaffected by the large purchase amounts. The median provides the preferred measure of location when the data are highly skewed.

z-Scores

In addition to measures of location, variability, and shape, we are also interested in the relative location of values within a data set. Measures of relative location help us determine how far a particular value is from the mean.

By using both the mean and standard deviation, we can determine the relative location of any observation. Suppose we have a sample of n observations, with the values denoted

by $x_1, x_2, \ldots, x_n$. In addition, assume that the sample mean, $\bar{x}$, and the sample standard deviation, s, are already computed. Associated with each value, x_i, is another value called its **z-score**. Equation (3.9) shows how the z-score is computed for each x_i.

z-SCORE

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.9}$$

where

z_i = the z-score for x_i
$\bar{x}$ = the sample mean
s = the sample standard deviation

The z-score is often called the *standardized value*. The z-score, z_i, can be interpreted as the *number of standard deviations x_i is from the mean $\bar{x}$*. For example, $z_1 = 1.2$ would indicate that x_1 is 1.2 standard deviations greater than the sample mean. Similarly, $z_2 = -.5$ would indicate that x_2 is .5, or 1/2, standard deviation less than the sample mean. A z-score greater than zero occurs for observations with a value greater than the mean, and a z-score less than zero occurs for observations with a value less than the mean. A z-score of zero indicates that the value of the observation is equal to the mean.

The z-score for any observation can be interpreted as a measure of the relative location of the observation in a data set. Thus, observations in two different data sets with the same z-score can be said to have the same relative location in terms of being the same number of standard deviations from the mean.

The z-scores for the class size data are computed in Table 3.4. Recall the previously computed sample mean, $\bar{x} = 44$, and sample standard deviation, $s = 8$. The z-score of -1.50 for the fifth observation shows it is farthest from the mean; it is 1.50 standard deviations below the mean.

Chebyshev's Theorem

Chebyshev's theorem enables us to make statements about the proportion of data values that must be within a specified number of standard deviations of the mean.

TABLE 3.4 z-SCORES FOR THE CLASS SIZE DATA

Number of Students in Class (x_i)	Deviation About the Mean ($x_i - \bar{x}$)	z-Score $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = .25
54	10	10/8 = 1.25
42	−2	−2/8 = −.25
46	2	2/8 = .25
32	−12	−12/8 = −1.50

CHEBYSHEV'S THEOREM

At least $(1 - 1/z^2)$ of the data values must be within z standard deviations of the mean, where z is any value greater than 1.

Some of the implications of this theorem, with $z = 2, 3$, and 4 standard deviations, follow.

- At least .75, or 75%, of the data values must be within $z = 2$ standard deviations of the mean.
- At least .89, or 89%, of the data values must be within $z = 3$ standard deviations of the mean.
- At least .94, or 94%, of the data values must be within $z = 4$ standard deviations of the mean.

For an example using Chebyshev's theorem, suppose that the midterm test scores for 100 students in a college business statistics course had a mean of 70 and a standard deviation of 5. How many students had test scores between 60 and 80? How many students had test scores between 58 and 82?

For the test scores between 60 and 80, we note that 60 is two standard deviations below the mean and 80 is two standard deviations above the mean. Using Chebyshev's theorem, we see that at least .75, or at least 75%, of the observations must have values within two standard deviations of the mean. Thus, at least 75% of the students must have scored between 60 and 80.

Chebyshev's theorem requires $z > 1$; but z need not be an integer.

For the test scores between 58 and 82, we see that $(58 - 70)/5 = -2.4$ indicates 58 is 2.4 standard deviations below the mean and that $(82 - 70)/5 = +2.4$ indicates 82 is 2.4 standard deviations above the mean. Applying Chebyshev's theorem with $z = 2.4$, we have

$$\left(1 - \frac{1}{z^2}\right) = \left(1 - \frac{1}{(2.4)^2}\right) = .826$$

At least 82.6% of the students must have test scores between 58 and 82.

Empirical Rule

The empirical rule is based on the normal probability distribution, which will be discussed in Chapter 6. The normal distribution is used extensively throughout the text.

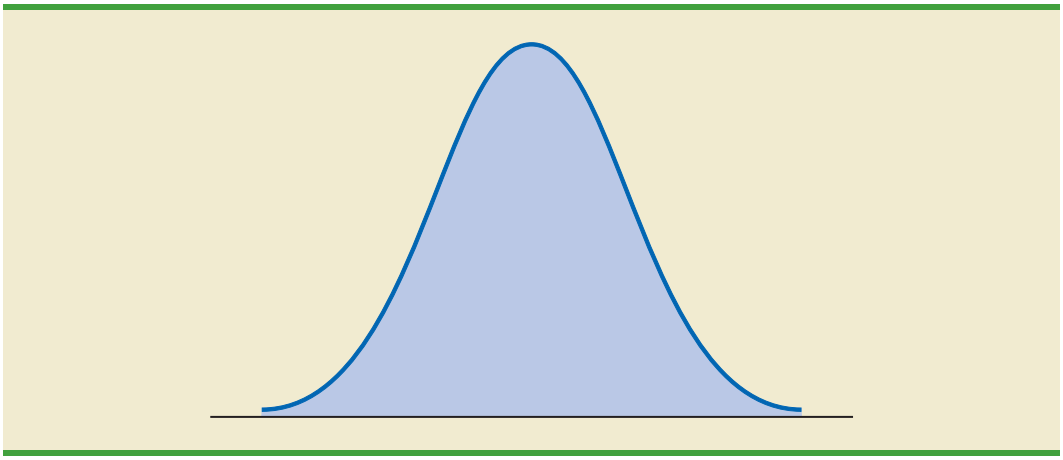
One of the advantages of Chebyshev's theorem is that it applies to any data set regardless of the shape of the distribution of the data. Indeed, it could be used with any of the distributions in Figure 3.3. In many practical applications, however, data sets exhibit a symmetric mound-shaped or bell-shaped distribution like the one shown in Figure 3.4. When the data are believed to approximate this distribution, the **empirical rule** can be used to determine the percentage of data values that must be within a specified number of standard deviations of the mean.

EMPIRICAL RULE

For data having a bell-shaped distribution:

- Approximately 68% of the data values will be within one standard deviation of the mean.
- Approximately 95% of the data values will be within two standard deviations of the mean.
- Almost all of the data values will be within three standard deviations of the mean.

FIGURE 3.4 A SYMMETRIC MOUND-SHAPED OR BELL-SHAPED DISTRIBUTION



For example, liquid detergent cartons are filled automatically on a production line. Filling weights frequently have a bell-shaped distribution. If the mean filling weight is 16 ounces and the standard deviation is .25 ounces, we can use the empirical rule to draw the following conclusions.

- Approximately 68% of the filled cartons will have weights between 15.75 and 16.25 ounces (within one standard deviation of the mean).
- Approximately 95% of the filled cartons will have weights between 15.50 and 16.50 ounces (within two standard deviations of the mean).
- Almost all filled cartons will have weights between 15.25 and 16.75 ounces (within three standard deviations of the mean).

Detecting Outliers

Sometimes a data set will have one or more observations with unusually large or unusually small values. These extreme values are called **outliers**. Experienced statisticians take steps to identify outliers and then review each one carefully. An outlier may be a data value that has been incorrectly recorded. If so, it can be corrected before further analysis. An outlier may also be from an observation that was incorrectly included in the data set; if so, it can be removed. Finally, an outlier may be an unusual data value that has been recorded correctly and belongs in the data set. In such cases it should remain.

It is a good idea to check for outliers before making decisions based on data analysis. Errors are often made in recording data and entering data into the computer. Outliers should not necessarily be deleted, but their accuracy and appropriateness should be verified.

Standardized values (z-scores) can be used to identify outliers. Recall that the empirical rule allows us to conclude that for data with a bell-shaped distribution, almost all the data values will be within three standard deviations of the mean. Hence, in using z-scores to identify outliers, we recommend treating any data value with a z-score less than -3 or greater than $+3$ as an outlier. Such data values can then be reviewed for accuracy and to determine whether they belong in the data set.

Refer to the z-scores for the class size data in Table 3.4. The z-score of -1.50 shows the fifth class size is farthest from the mean. However, this standardized value is well within the -3 to $+3$ guideline for outliers. Thus, the z-scores do not indicate that outliers are present in the class size data.

NOTES AND COMMENTS

1. Chebyshev's theorem is applicable for any data set and can be used to state the minimum number of data values that will be within a certain number of standard deviations of the mean. If the data are known to be approximately bell-shaped, more can be said. For instance, the

empirical rule allows us to say that *approximately* 95% of the data values will be within two standard deviations of the mean; Chebyshev's theorem allows us to conclude only that at least 75% of the data values will be in that interval.

2. Before analyzing a data set, statisticians usually make a variety of checks to ensure the validity of data. In a large study it is not uncommon for errors to be made in recording data values or in entering the values into a computer. Identifying outliers is one tool used to check the validity of the data.

Exercises

Methods

25. Consider a sample with data values of 10, 20, 12, 17, and 16. Compute the z-score for each of the five observations.

26. Consider a sample with a mean of 500 and a standard deviation of 100. What are the z-scores for the following data values: 520, 650, 500, 450, and 280?

SELF test

27. Consider a sample with a mean of 30 and a standard deviation of 5. Use Chebyshev's theorem to determine the percentage of the data within each of the following ranges:
    a. 20 to 40
    b. 15 to 45
    c. 22 to 38
    d. 18 to 42
    e. 12 to 48

28. Suppose the data have a bell-shaped distribution with a mean of 30 and a standard deviation of 5. Use the empirical rule to determine the percentage of data within each of the following ranges:
    a. 20 to 40
    b. 15 to 45
    c. 25 to 35

Applications

SELF test

29. The results of a national survey showed that on average, adults sleep 6.9 hours per night. Suppose that the standard deviation is 1.2 hours.
    a. Use Chebyshev's theorem to calculate the percentage of individuals who sleep between 4.5 and 9.3 hours.
    b. Use Chebyshev's theorem to calculate the percentage of individuals who sleep between 3.9 and 9.9 hours.
    c. Assume that the number of hours of sleep follows a bell-shaped distribution. Use the empirical rule to calculate the percentage of individuals who sleep between 4.5 and 9.3 hours per day. How does this result compare to the value that you obtained using Chebyshev's theorem in part (a)?

30. The Energy Information Administration reported that the mean retail price per gallon of regular grade gasoline was $2.05 (Energy Information Administration, May 2009). Suppose that the standard deviation was $.10 and that the retail price per gallon has a bell-shaped distribution.
    a. What percentage of regular grade gasoline sold between $1.95 and $2.15 per gallon?
    b. What percentage of regular grade gasoline sold between $1.95 and $2.25 per gallon?
    c. What percentage of regular grade gasoline sold for more than $2.25 per gallon?

31. The national average for the math portion of the College Board's Scholastic Aptitude Test (SAT) is 515 (*The World Almanac,* 2009). The College Board periodically rescales the test scores such that the standard deviation is approximately 100. Answer the following questions using a bell-shaped distribution and the empirical rule for the verbal test scores.

a. What percentage of students have an SAT verbal score greater than 615?
b. What percentage of students have an SAT verbal score greater than 715?
c. What percentage of students have an SAT verbal score between 415 and 515?
d. What percentage of students have an SAT verbal score between 315 and 615?

32. The high costs in the California real estate market have caused families who cannot afford to buy bigger homes to consider backyard sheds as an alternative form of housing expansion. Many are using the backyard structures for home offices, art studios, and hobby areas as well as for additional storage. The mean price of a customized wooden, shingled backyard structure is $3100 (*Newsweek,* September 29, 2003). Assume that the standard deviation is $1200.
a. What is the z-score for a backyard structure costing $2300?
b. What is the z-score for a backyard structure costing $4900?
c. Interpret the z-scores in parts (a) and (b). Comment on whether either should be considered an outlier.
d. The *Newsweek* article described a backyard shed-office combination built in Albany, California, for $13,000. Should this structure be considered an outlier? Explain.

33. Florida Power & Light (FP&L) Company has enjoyed a reputation for quickly fixing its electric system after storms. However, during the hurricane seasons of 2004 and 2005, a new reality was that the company's historical approach to emergency electric system repairs was no longer good enough (*The Wall Street Journal,* January 16, 2006). Data showing the days required to restore electric service after seven hurricanes during 2004 and 2005 follow.

Hurricane	Days to Restore Service
Charley	13
Frances	12
Jeanne	8
Dennis	3
Katrina	8
Rita	2
Wilma	18

Based on this sample of seven, compute the following descriptive statistics:
a. Mean, median, and mode
b. Range and standard deviation
c. Should Wilma be considered an outlier in terms of the days required to restore electric service?
d. The seven hurricanes resulted in 10 million service interruptions to customers. Do the statistics show that FP&L should consider updating its approach to emergency electric system repairs? Discuss.

34. A sample of 10 NCAA college basketball game scores provided the following data (*USA Today,* January 26, 2004).

WEB file
NCAA

Winning Team	Points	Losing Team	Points	Winning Margin
Arizona	90	Oregon	66	24
Duke	85	Georgetown	66	19
Florida State	75	Wake Forest	70	5
Kansas	78	Colorado	57	21
Kentucky	71	Notre Dame	63	8
Louisville	65	Tennessee	62	3
Oklahoma State	72	Texas	66	6

Winning Team	Points	Losing Team	Points	Winning Margin
Purdue	76	Michigan State	70	6
Stanford	77	Southern Cal	67	10
Wisconsin	76	Illinois	56	20

a. Compute the mean and standard deviation for the points scored by the winning team.
b. Assume that the points scored by the winning teams for all NCAA games follow a bell-shaped distribution. Using the mean and standard deviation found in part (a), estimate the percentage of all NCAA games in which the winning team scores 84 or more points. Estimate the percentage of NCAA games in which the winning team scores more than 90 points.
c. Compute the mean and standard deviation for the winning margin. Do the data contain outliers? Explain.

35. *Consumer Reports* posts reviews and ratings of a variety of products on its website. The following is a sample of 20 speaker systems and their ratings. The ratings are on a scale of 1 to 5, with 5 being best.

WEB file
Speakers

Speaker	Rating	Speaker	Rating
Infinity Kappa 6.1	4.00	ACI Sapphire III	4.67
Allison One	4.12	Bose 501 Series	2.14
Cambridge Ensemble II	3.82	DCM KX-212	4.09
Dynaudio Contour 1.3	4.00	Eosone RSF1000	4.17
Hsu Rsch. HRSW12V	4.56	Joseph Audio RM7si	4.88
Legacy Audio Focus	4.32	Martin Logan Aerius	4.26
Mission 73li	4.33	Omni Audio SA 12.3	2.32
PSB 400i	4.50	Polk Audio RT12	4.50
Snell Acoustics D IV	4.64	Sunfire True Subwoofer	4.17
Thiel CS1.5	4.20	Yamaha NS-A636	2.17

a. Compute the mean and the median.
b. Compute the first and third quartiles.
c. Compute the standard deviation.
d. The skewness of this data is −1.67. Comment on the shape of the distribution.
e. What are the z-scores associated with Allison One and Omni Audio?
f. Do the data contain any outliers? Explain.

3.4 Exploratory Data Analysis

In Chapter 2 we introduced the stem-and-leaf display as a technique of exploratory data analysis. Recall that exploratory data analysis enables us to use simple arithmetic and easy-to-draw pictures to summarize data. In this section we continue exploratory data analysis by considering five-number summaries and box plots.

Five-Number Summary

In a **five-number summary**, the following five numbers are used to summarize the data:

1. Smallest value
2. First quartile (Q_1)
3. Median (Q_2)
4. Third quartile (Q_3)
5. Largest value

The easiest way to develop a five-number summary is to first place the data in ascending order. Then it is easy to identify the smallest value, the three quartiles, and the largest value. The monthly starting salaries shown in Table 3.1 for a sample of 12 business school graduates are repeated here in ascending order.

3310	3355	3450	3480	3480	3490	3520	3540	3550	3650	3730	3925
		$Q_1 = 3465$			$Q_2 = 3505$ (Median)			$Q_3 = 3600$			

The median of 3505 and the quartiles $Q_1 = 3465$ and $Q_3 = 3600$ were computed in Section 3.1. Reviewing the data shows a smallest value of 3310 and a largest value of 3925. Thus the five-number summary for the salary data is 3310, 3465, 3505, 3600, 3925. Approximately one-fourth, or 25%, of the observations are between adjacent numbers in a five-number summary.

Box Plot

A **box plot** is a graphical summary of data that is based on a five-number summary. A key to the development of a box plot is the computation of the median and the quartiles, Q_1 and Q_3. The interquartile range, IQR $= Q_3 - Q_1$, is also used. Figure 3.5 is the box plot for the monthly starting salary data. The steps used to construct the box plot follow.

1. A box is drawn with the ends of the box located at the first and third quartiles. For the salary data, $Q_1 = 3465$ and $Q_3 = 3600$. This box contains the middle 50% of the data.
2. A vertical line is drawn in the box at the location of the median (3505 for the salary data).
3. By using the interquartile range, IQR $= Q_3 - Q_1$, *limits* are located. The limits for the box plot are 1.5(IQR) below Q_1 and 1.5(IQR) above Q_3. For the salary data, IQR $= Q_3 - Q_1 = 3600 - 3465 = 135$. Thus, the limits are $3465 - 1.5(135) = 3262.5$ and $3600 + 1.5(135) = 3802.5$. Data outside these limits are considered *outliers.*
4. The dashed lines in Figure 3.5 are called *whiskers.* The whiskers are drawn from the ends of the box to the smallest and largest values *inside the limits* computed in step 3. Thus, the whiskers end at salary values of 3310 and 3730.
5. Finally, the location of each outlier is shown with the symbol *. In Figure 3.5 we see one outlier, 3925.

Box plots provide another way to identify outliers. But they do not necessarily identify the same values as those with a z-score less than −3 or greater than +3. Either or both procedures may be used.

In Figure 3.5 we included lines showing the location of the upper and lower limits. These lines were drawn to show how the limits are computed and where they are located.

FIGURE 3.5 BOX PLOT OF THE STARTING SALARY DATA WITH LINES SHOWING THE LOWER AND UPPER LIMITS

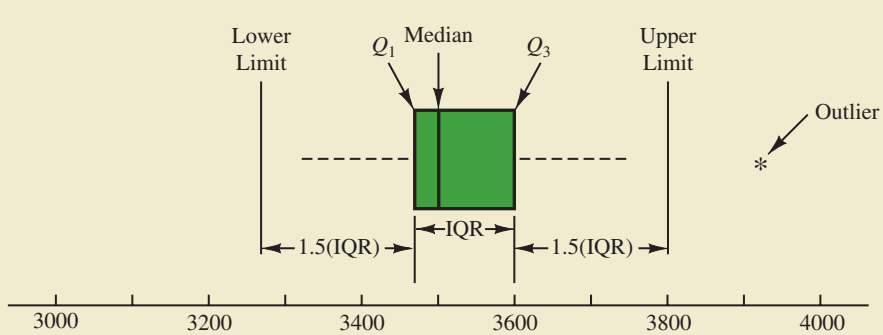
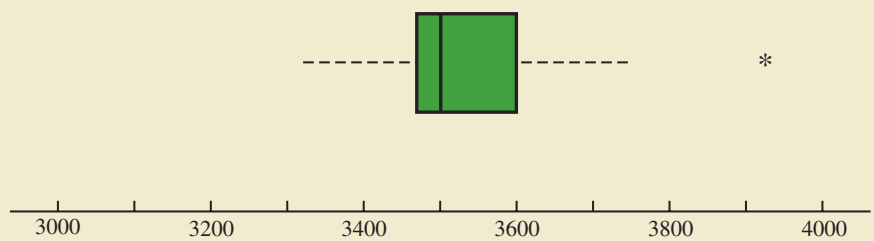


FIGURE 3.6 BOX PLOT OF MONTHLY STARTING SALARY DATA



Although the limits are always computed, generally they are not drawn on the box plots. Figure 3.6 shows the usual appearance of a box plot for the salary data.

WEB file
MajorSalary

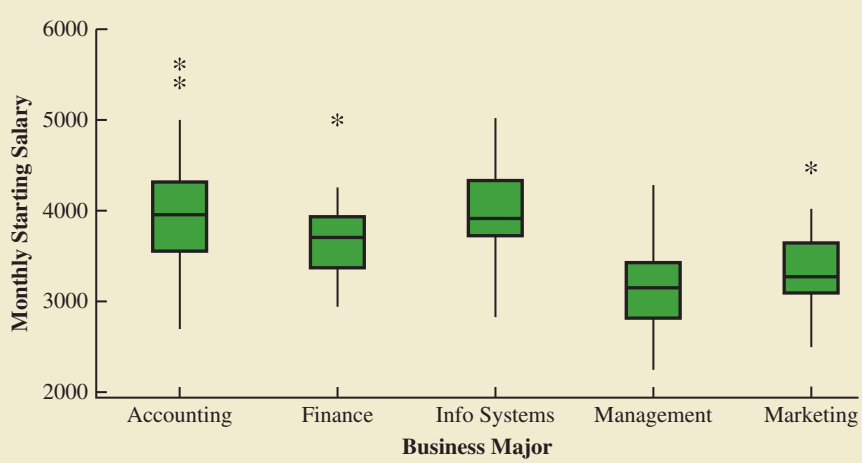
In order to compare monthly starting salaries for business school graduates by major, a sample of 111 recent graduates was selected. The major and the monthly starting salary were recorded for each graduate. Figure 3.7 shows the Minitab box plots for accounting, finance, information systems, management, and marketing majors. Note that the major is shown on the horizontal axis and each box plot is shown vertically above the corresponding major. Displaying box plots in this manner is an excellent graphical technique for making comparisons among two or more groups.

What observations can you make about monthly starting salaries by major using the box plots in Figured 3.7? Specifically, we note the following:

- The higher salaries are in accounting; the lower salaries are in management and marketing.
- Based on the medians, accounting and information systems have similar and higher median salaries. Finance is next with management and marketing showing lower median salaries.
- High salary outliers exist for accounting, finance, and marketing majors.
- Finance salaries appear to have the least variation, while accounting salaries appear to have the most variation.

Perhaps you can see additional interpretations based on these box plots.

FIGURE 3.7 MINITAB BOX PLOTS OF MONTLY STARTING SALARY BY MAJOR



NOTES AND COMMENTS

1. An advantage of the exploratory data analysis procedures is that they are easy to use; few numerical calculations are necessary. We simply sort the data values into ascending order and identify the five-number summary. The box plot can then be constructed. It is not necessary to compute the mean and the standard deviation for the data.
2. In Appendix 3.1, we show how to construct a box plot for the starting salary data using Minitab. The box plot obtained looks just like the one in Figure 3.6, but turned on its side.

Exercises

Methods

36. Consider a sample with data values of 27, 25, 20, 15, 30, 34, 28, and 25. Provide the five-number summary for the data.
37. Show the box plot for the data in exercise 36.

SELF test

38. Show the five-number summary and the box plot for the following data: 5, 15, 18, 10, 8, 12, 16, 10, 6.
39. A data set has a first quartile of 42 and a third quartile of 50. Compute the lower and upper limits for the corresponding box plot. Should a data value of 65 be considered an outlier?

Applications

40. Naples, Florida, hosts a half-marathon (13.1-mile race) in January each year. The event attracts top runners from throughout the United States as well as from around the world. In January 2009, 22 men and 31 women entered the 19–24 age class. Finish times in minutes are as follows (*Naples Daily News,* January 19, 2009). Times are shown in order of finish.

WEB file
Runners

Finish	Men	Women	Finish	Men	Women	Finish	Men	Women
1	65.30	109.03	11	109.05	123.88	21	143.83	136.75
2	66.27	111.22	12	110.23	125.78	22	148.70	138.20
3	66.52	111.65	13	112.90	129.52	23		139.00
4	66.85	111.93	14	113.52	129.87	24		147.18
5	70.87	114.38	15	120.95	130.72	25		147.35
6	87.18	118.33	16	127.98	131.67	26		147.50
7	96.45	121.25	17	128.40	132.03	27		147.75
8	98.52	122.08	18	130.90	133.20	28		153.88
9	100.52	122.48	19	131.80	133.50	29		154.83
10	108.18	122.62	20	138.63	136.57	30		189.27
						31		189.28

a. George Towett of Marietta, Georgia, finished in first place for the men and Lauren Wald of Gainesville, Florida, finished in first place for the women. Compare the first-place finish times for men and women. If the 53 men and women runners had competed as one group, in what place would Lauren have finished?
b. What is the median time for men and women runners? Compare men and women runners based on their median times.
c. Provide a five-number summary for both the men and the women.
d. Are there outliers in either group?

e. Show the box plots for the two groups. Did men or women have the most variation in finish times? Explain.

SELF test

41. Annual sales, in millions of dollars, for 21 pharmaceutical companies follow.

8408	1374	1872	8879	2459	11413
608	14138	6452	1850	2818	1356
10498	7478	4019	4341	739	2127
3653	5794	8305			

a. Provide a five-number summary.
b. Compute the lower and upper limits.
c. Do the data contain any outliers?
d. Johnson & Johnson's sales are the largest on the list at $14,138 million. Suppose a data entry error (a transposition) had been made and the sales had been entered as $41,138 million. Would the method of detecting outliers in part (c) identify this problem and allow for correction of the data entry error?
e. Show a box plot.

42. *Consumer Reports* provided overall customer satisfaction scores for AT&T, Sprint, T-Mobile, and Verizon cell-phone services in major metropolitan areas throughout the United States. The rating for each service reflects the overall customer satisfaction considering a variety of factors such as cost, connectivity problems, dropped calls, static interference, and customer support. A satisfaction scale from 0 to 100 was used with 0 indicating completely dissatisfied and 100 indicating completely satisfied. The ratings for the four cell-phone services in 20 metropolitan areas are as shown (*Consumer Reports,* January 2009).

WEB file
CellService

Metropolitan Area	AT&T	Sprint	T-Mobile	Verizon
Atlanta	70	66	71	79
Boston	69	64	74	76
Chicago	71	65	70	77
Dallas	75	65	74	78
Denver	71	67	73	77
Detroit	73	65	77	79
Jacksonville	73	64	75	81
Las Vegas	72	68	74	81
Los Angeles	66	65	68	78
Miami	68	69	73	80
Minneapolis	68	66	75	77
Philadelphia	72	66	71	78
Phoenix	68	66	76	81
San Antonio	75	65	75	80
San Diego	69	68	72	79
San Francisco	66	69	73	75
Seattle	68	67	74	77
St. Louis	74	66	74	79
Tampa	73	63	73	79
Washington	72	68	71	76

a. Consider T-Mobile first. What is the median rating?
b. Develop a five-number summary for the T-Mobile service.
c. Are there outliers for T-Mobile? Explain.
d. Repeat parts (b) and (c) for the other three cell-phone services.

e. Show the box plots for the four cell-phone services on one graph. Discuss what a comparison of the box plots tells about the four services. Which service did *Consumer Reports* recommend as being best in terms of overall customer satisfaction?

43. The Philadelphia Phillies defeated the Tampa Bay Rays 4 to 3 to win the 2008 major league baseball World Series (*The Philadelphia Inquirer,* October 29, 2008). Earlier in the major league baseball playoffs, the Philadelphia Phillies defeated the Los Angeles Dodgers to win the National League Championship, while the Tampa Bay Rays defeated the Boston Red Sox to win the American League Championship. The file MLBSalaries contains the salaries for the 28 players on each of these four teams (USA Today Salary Database, October 2008). The data, shown in thousands of dollars, have been ordered from the highest salary to the lowest salary for each team.

WEB file
MLBSalaries

a. Analyze the salaries for the World Champion Philadelphia Phillies. What is the total payroll for the team? What is the median salary? What is the five-number summary?

b. Were there salary outliers for the Philadelphia Phillies? If so, how many and what were the salary amounts?

c. What is the total payroll for each of the other three teams? Develop the five-number summary for each team and identify any outliers.

d. Show the box plots of the salaries for all four teams. What are your interpretations? Of these four teams, does it appear that the team with the higher salaries won the league championships and the World Series?

WEB file
Mutual

44. A listing of 46 mutual funds and their 12-month total return percentage is shown in Table 3.5 (*Smart Money,* February 2004).
a. What are the mean and median return percentages for these mutual funds?
b. What are the first and third quartiles?
c. Provide a five-number summary.
d. Do the data contain any outliers? Show a box plot.

TABLE 3.5 TWELVE-MONTH RETURN FOR MUTUAL FUNDS

Mutual Fund	Return (%)	Mutual Fund	Return (%)
Alger Capital Appreciation	23.5	Nations Small Company	21.4
Alger LargeCap Growth	22.8	Nations SmallCap Index	24.5
Alger MidCap Growth	38.3	Nations Strategic Growth	10.4
Alger SmallCap	41.3	Nations Value Inv	10.8
AllianceBernstein Technology	40.6	One Group Diversified Equity	10.0
Federated American Leaders	15.6	One Group Diversified Int'l	10.9
Federated Capital Appreciation	12.4	One Group Diversified Mid Cap	15.1
Federated Equity-Income	11.5	One Group Equity Income	6.6
Federated Kaufmann	33.3	One Group Int'l Equity Index	13.2
Federated Max-Cap Index	16.0	One Group Large Cap Growth	13.6
Federated Stock	16.9	One Group Large Cap Value	12.8
Janus Adviser Int'l Growth	10.3	One Group Mid Cap Growth	18.7
Janus Adviser Worldwide	3.4	One Group Mid Cap Value	11.4
Janus Enterprise	24.2	One Group Small Cap Growth	23.6
Janus High-Yield	12.1	PBHG Growth	27.3
Janus Mercury	20.6	Putnam Europe Equity	20.4
Janus Overseas	11.9	Putnam Int'l Capital Opportunity	36.6
Janus Worldwide	4.1	Putnam International Equity	21.5
Nations Convertible Securities	13.6	Putnam Int'l New Opportunity	26.3
Nations Int'l Equity	10.7	Strong Advisor Mid Cap Growth	23.7
Nations LargeCap Enhd. Core	13.2	Strong Growth 20	11.7
Nations LargeCap Index	13.5	Strong Growth Inv	23.2
Nation MidCap Index	19.5	Strong Large Cap Growth	14.5

3.5 Measures of Association Between Two Variables

Thus far we have examined numerical methods used to summarize the data for *one variable at a time.* Often a manager or decision maker is interested in the *relationship between two variables.* In this section we present covariance and correlation as descriptive measures of the relationship between two variables.

We begin by reconsidering the application concerning a stereo and sound equipment store in San Francisco as presented in Section 2.4. The store's manager wants to determine the relationship between the number of weekend television commercials shown and the sales at the store during the following week. Sample data with sales expressed in hundreds of dollars are provided in Table 3.6. It shows 10 observations ($n = 10$), one for each week. The scatter diagram in Figure 3.8 shows a positive relationship, with higher sales (y) associated with a greater number of commercials (x). In fact, the scatter diagram suggests that a straight line could be used as an approximation of the relationship. In the following discussion, we introduce **covariance** as a descriptive measure of the linear association between two variables.

Covariance

For a sample of size n with the observations (x_1, y_1), (x_2, y_2), and so on, the sample covariance is defined as follows:

SAMPLE COVARIANCE

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \qquad (3.10)$$

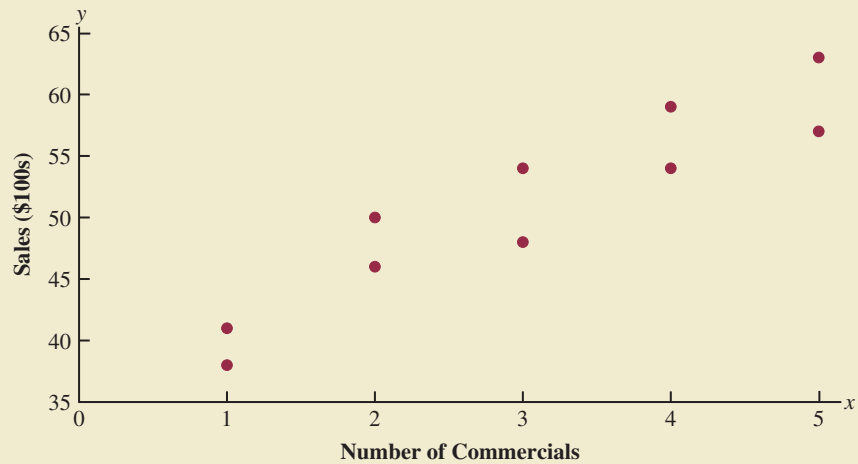
This formula pairs each x_i with a y_i. We then sum the products obtained by multiplying the deviation of each x_i from its sample mean $\bar{x}$ by the deviation of the corresponding y_i from its sample mean $\bar{y}$; this sum is then divided by $n - 1$.

TABLE 3.6 SAMPLE DATA FOR THE STEREO AND SOUND EQUIPMENT STORE

Week	Number of Commercials x	Sales Volume ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

WEB file
Stereo

FIGURE 3.8 SCATTER DIAGRAM FOR THE STEREO AND SOUND EQUIPMENT STORE



To measure the strength of the linear relationship between the number of commercials x and the sales volume y in the stereo and sound equipment store problem, we use equation (3.10) to compute the sample covariance. The calculations in Table 3.7 show the computation of $\Sigma(x_i - \bar{x})(y_i - \bar{y})$. Note that $\bar{x} = 30/10 = 3$ and $\bar{y} = 510/10 = 51$. Using equation (3.10), we obtain a sample covariance of

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{9} = 11$$

TABLE 3.7 CALCULATIONS FOR THE SAMPLE COVARIANCE

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
Totals	30	510	0	0	99

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{10 - 1} = 11$$

The formula for computing the covariance of a population of size N is similar to equation (3.10), but we use different notation to indicate that we are working with the entire population.

POPULATION COVARIANCE

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{N} \tag{3.11}$$

In equation (3.11) we use the notation μ_x for the population mean of the variable x and μ_y for the population mean of the variable y. The population covariance σ_{xy} is defined for a population of size N.

Interpretation of the Covariance

To aid in the interpretation of the sample covariance, consider Figure 3.9. It is the same as the scatter diagram of Figure 3.7 with a vertical dashed line at $\bar{x} = 3$ and a horizontal dashed line at $\bar{y} = 51$. The lines divide the graph into four quadrants. Points in quadrant I correspond to x_i greater than $\bar{x}$ and y_i greater than $\bar{y}$, points in quadrant II correspond to x_i less than $\bar{x}$ and y_i greater than $\bar{y}$, and so on. Thus, the value of $(x_i - \bar{x})(y_i - \bar{y})$ must be positive for points in quadrant I, negative for points in quadrant II, positive for points in quadrant III, and negative for points in quadrant IV.

The covariance is a measure of the linear association between two variables.

If the value of s_{xy} is positive, the points with the greatest influence on s_{xy} must be in quadrants I and III. Hence, a positive value for s_{xy} indicates a positive linear association between x and y; that is, as the value of x increases, the value of y increases. If the value of s_{xy} is negative, however, the points with the greatest influence on s_{xy} are in quadrants II and IV. Hence, a negative value for s_{xy} indicates a negative linear association between x and y; that is, as the value of x increases, the value of y decreases. Finally, if the points are evenly distributed across all four quadrants, the value of s_{xy} will be close to zero, indicating no linear association between x and y. Figure 3.10 shows the values of s_{xy} that can be expected with three different types of scatter diagrams.

FIGURE 3.9 PARTITIONED SCATTER DIAGRAM FOR THE STEREO AND SOUND EQUIPMENT STORE

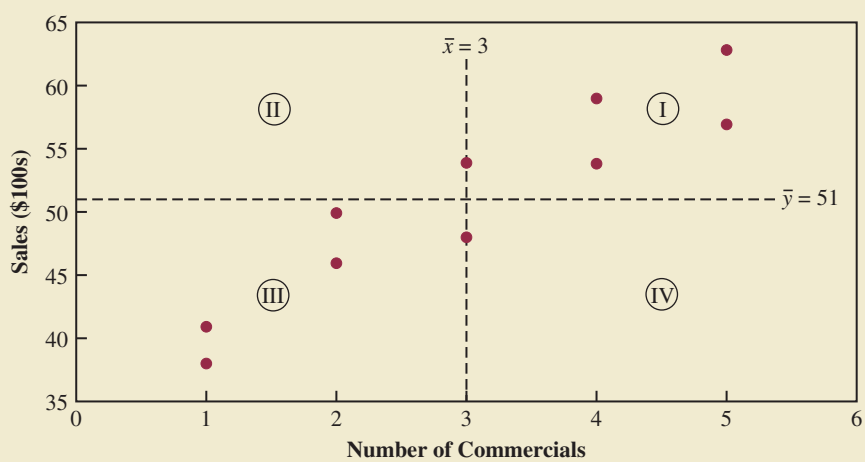


FIGURE 3.10 INTERPRETATION OF SAMPLE COVARIANCE

Referring again to Figure 3.9, we see that the scatter diagram for the stereo and sound equipment store follows the pattern in the top panel of Figure 3.10. As we should expect, the value of the sample covariance indicates a positive linear relationship with $s_{xy} = 11$.

From the preceding discussion, it might appear that a large positive value for the covariance indicates a strong positive linear relationship and that a large negative value indicates a strong negative linear relationship. However, one problem with using covariance as a measure of the strength of the linear relationship is that the value of the covariance depends on the units of measurement for x and y. For example, suppose we are interested in the relationship between height x and weight y for individuals. Clearly the strength of the relationship should be the same whether we measure height in feet or inches. Measuring the height in inches, however, gives us much larger numerical values for $(x_i - \bar{x})$ than when we measure height in feet. Thus, with height measured in inches, we would obtain a larger value for the numerator $\Sigma(x_i - \bar{x})(y_i - \bar{y})$ in equation (3.10)—and hence a larger covariance—when in fact the relationship does not change. A measure of the relationship between two variables that is not affected by the units of measurement for x and y is the **correlation coefficient**.

Correlation Coefficient

For sample data, the Pearson product moment correlation coefficient is defined as follows.

PEARSON PRODUCT MOMENT CORRELATION COEFFICIENT: SAMPLE DATA

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3.12}$$

where

r_{xy} = sample correlation coefficient
s_{xy} = sample covariance
s_x = sample standard deviation of x
s_y = sample standard deviation of y

Equation (3.12) shows that the Pearson product moment correlation coefficient for sample data (commonly referred to more simply as the *sample correlation coefficient*) is computed by dividing the sample covariance by the product of the sample standard deviation of x and the sample standard deviation of y.

Let us now compute the sample correlation coefficient for the stereo and sound equipment store. Using the data in Table 3.7, we can compute the sample standard deviations for the two variables:

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{20}{9}} = 1.49$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{566}{9}} = 7.93$$

Now, because $s_{xy} = 11$, the sample correlation coefficient equals

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{(1.49)(7.93)} = .93$$

The formula for computing the correlation coefficient for a population, denoted by the Greek letter ρ_{xy} (rho, pronounced "row"), follows.

PEARSON PRODUCT MOMENT CORRELATION COEFFICIENT: POPULATION DATA

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.13}$$

where

ρ_{xy} = population correlation coefficient
σ_{xy} = population covariance
σ_x = population standard deviation for x
σ_y = population standard deviation for y

The sample correlation coefficient r_{xy} is the estimator of the population correlation coefficient ρ_{xy}.

The sample correlation coefficient r_{xy} provides an estimate of the population correlation coefficient ρ_{xy}.

Interpretation of the Correlation Coefficient

First let us consider a simple example that illustrates the concept of a perfect positive linear relationship. The scatter diagram in Figure 3.11 depicts the relationship between x and y based on the following sample data.

x_i	y_i
5	10
10	30
15	50

FIGURE 3.11 SCATTER DIAGRAM DEPICTING A PERFECT POSITIVE LINEAR RELATIONSHIP

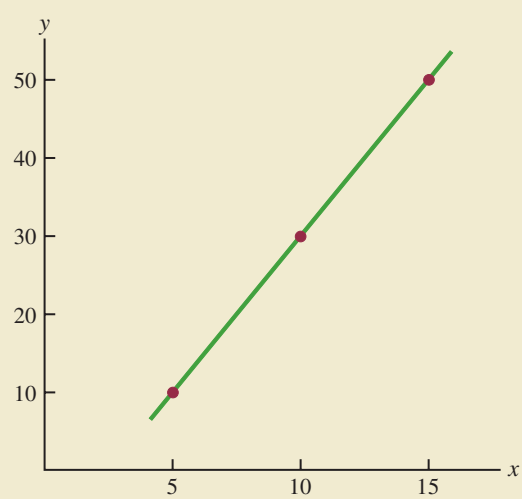


TABLE 3.8 COMPUTATIONS USED IN CALCULATING THE SAMPLE CORRELATION COEFFICIENT

	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	5	10	−5	25	−20	400	100
	10	30	0	0	0	0	0
	15	50	5	25	20	400	100
Totals	30	90	0	50	0	800	200
	$\bar{x} = 10$	$\bar{y} = 30$					

The straight line drawn through each of the three points shows a perfect linear relationship between x and y. In order to apply equation (3.12) to compute the sample correlation we must first compute s_{xy}, s_x, and s_y. Some of the computations are shown in Table 3.8. Using the results in this table, we find

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{200}{2} = 100$$

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{50}{2}} = 5$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{800}{2}} = 20$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{100}{5(20)} = 1$$

The correlation coefficient ranges from −1 to +1. Values close to −1 or +1 indicate a strong linear relationship. The closer the correlation is to zero, the weaker the relationship.

Thus, we see that the value of the sample correlation coefficient is 1.

In general, it can be shown that if all the points in a data set fall on a positively sloped straight line, the value of the sample correlation coefficient is +1; that is, a sample correlation coefficient of +1 corresponds to a perfect positive linear relationship between x and y. Moreover, if the points in the data set fall on a straight line having negative slope, the value of the sample correlation coefficient is −1; that is, a sample correlation coefficient of −1 corresponds to a perfect negative linear relationship between x and y.

Let us now suppose that a certain data set indicates a positive linear relationship between x and y but that the relationship is not perfect. The value of r_{xy} will be less than 1, indicating that the points in the scatter diagram are not all on a straight line. As the points deviate more and more from a perfect positive linear relationship, the value of r_{xy} becomes smaller and smaller. A value of r_{xy} equal to zero indicates no linear relationship between x and y, and values of r_{xy} near zero indicate a weak linear relationship.

For the data involving the stereo and sound equipment store, $r_{xy} = .93$. Therefore, we conclude that a strong positive linear relationship occurs between the number of commercials and sales. More specifically, an increase in the number of commercials is associated with an increase in sales.

In closing, we note that correlation provides a measure of linear association and not necessarily causation. A high correlation between two variables does not mean that changes in one variable will cause changes in the other variable. For example, we may find that the quality rating and the typical meal price of restaurants are positively correlated. However, simply increasing the meal price at a restaurant will not cause the quality rating to increase.

Exercises

Methods

SELF test

45. Five observations taken for two variables follow.

x_i	4	6	11	3	16
y_i	50	50	40	60	30

a. Develop a scatter diagram with x on the horizontal axis.
b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?
c. Compute and interpret the sample covariance.
d. Compute and interpret the sample correlation coefficient.

46. Five observations taken for two variables follow.

x_i	6	11	15	21	27
y_i	6	9	6	17	12

a. Develop a scatter diagram for these data.
b. What does the scatter diagram indicate about a relationship between x and y?
c. Compute and interpret the sample covariance.
d. Compute and interpret the sample correlation coefficient.

Applications

47. Nielsen Media Research provides two measures of the television viewing audience: a television program *rating,* which is the percentage of households with televisions watching a program, and a television program *share,* which is the percentage of households watching a program among those with televisions in use. The following data show the Nielsen television ratings and share data for the Major League Baseball World Series over a nine-year period (Associated Press, October 27, 2003).

Rating	19	17	17	14	16	12	15	12	13
Share	32	28	29	24	26	20	24	20	22

a. Develop a scatter diagram with rating on the horizontal axis.
b. What is the relationship between rating and share? Explain.
c. Compute and interpret the sample covariance.
d. Compute the sample correlation coefficient. What does this value tell us about the relationship between rating and share?

48. A department of transportation's study on driving speed and miles per gallon for midsize automobiles resulted in the following data:

Speed (Miles per Hour)	30	50	40	55	30	25	60	25	50	55
Miles per Gallon	28	25	25	23	30	32	21	35	26	25

Compute and interpret the sample correlation coefficient.

49. At the beginning of 2009, the economic downturn resulted in the loss of jobs and an increase in delinquent loans for housing. The national unemployment rate was 6.5% and the percentage of delinquent loans was 6.12% (*The Wall Street Journal,* January 27, 2009). In projecting where the real estate market was headed in the coming year, economists studied the relationship between the jobless rate and the percentage of delinquent loans. The expectation was that if the jobless rate continued to increase, there would also be an

increase in the percentage of delinquent loans. The data below show the jobless rate and the delinquent loan percentage for 27 major real estate markets.

WEB file
Housing

Metro Area	Jobless Rate (%)	Delinquent Loan (%)	Metro Area	Jobless Rate (%)	Delinquent Loan (%)
Atlanta	7.1	7.02	New York	6.2	5.78
Boston	5.2	5.31	Orange County	6.3	6.08
Charlotte	7.8	5.38	Orlando	7.0	10.05
Chicago	7.8	5.40	Philadelphia	6.2	4.75
Dallas	5.8	5.00	Phoenix	5.5	7.22
Denver	5.8	4.07	Portland	6.5	3.79
Detroit	9.3	6.53	Raleigh	6.0	3.62
Houston	5.7	5.57	Sacramento	8.3	9.24
Jacksonville	7.3	6.99	St. Louis	7.5	4.40
Las Vegas	7.6	11.12	San Diego	7.1	6.91
Los Angeles	8.2	7.56	San Francisco	6.8	5.57
Miami	7.1	12.11	Seattle	5.5	3.87
Minneapolis	6.3	4.39	Tampa	7.5	8.42
Nashville	6.6	4.78			

a. Compute the correlation coefficient. Is there a positive correlation between the jobless rate and the percentage of delinquent housing loans? What is your interpretation?
b. Show a scatter diagram of the relationship between jobless rate and the percentage of delinquent housing loans.

50. The Dow Jones Industrial Average (DJIA) and the Standard & Poor's 500 Index (S&P 500) are both used to measure the performance of the stock market. The DJIA is based on the price of stocks for 30 large companies; the S&P 500 is based on the price of stocks for 500 companies. If both the DJIA and S&P 500 measure the performance of the stock market, how are they correlated? The following data show the daily percent increase or daily percent decrease in the DJIA and S&P 500 for a sample of nine days over a three-month period (*The Wall Street Journal,* January 15 to March 10, 2006).

WEB file
StockMarket

DJIA	.20	.82	−.99	.04	−.24	1.01	.30	.55	−.25
S&P 500	.24	.19	−.91	.08	−.33	.87	.36	.83	−.16

a. Show a scatter diagram.
b. Compute the sample correlation coefficient for these data.
c. Discuss the association between the DJIA and S&P 500. Do you need to check both before having a general idea about the daily stock market performance?

51. The daily high and low temperatures for 14 cities around the world are shown (The Weather Channel, April 22, 2009).

WEB file
WorldTemp

City	High	Low	City	High	Low
Athens	68	50	London	67	45
Beijing	70	49	Moscow	44	29
Berlin	65	44	Paris	69	44
Cairo	96	64	Rio de Janeiro	76	69
Dublin	57	46	Rome	69	51
Geneva	70	45	Tokyo	70	58
Hong Kong	80	73	Toronto	44	39

a. What is the sample mean high temperature?
b. What is the sample mean low temperature?
c. What is the correlation between the high and low temperatures? Discuss.

3.6 The Weighted Mean and Working with Grouped Data

In Section 3.1, we presented the mean as one of the most important measures of central location. The formula for the mean of a sample with n observations is restated as follows.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{3.14}$$

In this formula, each x_i is given equal importance or weight. Although this practice is most common, in some instances, the mean is computed by giving each observation a weight that reflects its importance. A mean computed in this manner is referred to as a **weighted mean**.

Weighted Mean

The weighted mean is computed as follows:

WEIGHTED MEAN

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} \tag{3.15}$$

where

x_i = value of observation i
w_i = weight for observation i

When the data are from a sample, equation (3.15) provides the weighted sample mean. When the data are from a population, μ replaces $\bar{x}$ and equation (3.15) provides the weighted population mean.

As an example of the need for a weighted mean, consider the following sample of five purchases of a raw material over the past three months.

Purchase	Cost per Pound ($)	Number of Pounds
1	3.00	1200
2	3.40	500
3	2.80	2750
4	2.90	1000
5	3.25	800

Note that the cost per pound varies from $2.80 to $3.40, and the quantity purchased varies from 500 to 2750 pounds. Suppose that a manager asked for information about the mean cost per pound of the raw material. Because the quantities ordered vary, we must use the formula for a weighted mean. The five cost-per-pound data values are $x_1 = 3.00$, $x_2 = 3.40$, $x_3 = 2.80$, $x_4 = 2.90$, and $x_5 = 3.25$. The weighted mean cost per pound is found by weighting each cost

by its corresponding quantity. For this example, the weights are $w_1 = 1200$, $w_2 = 500$, $w_3 = 2750$, $w_4 = 1000$, and $w_5 = 800$. Based on equation (3.15), the weighted mean is calculated as follows:

$$\bar{x} = \frac{1200(3.00) + 500(3.40) + 2750(2.80) + 1000(2.90) + 800(3.25)}{1200 + 500 + 2750 + 1000 + 800}$$

$$= \frac{18{,}500}{6250} = 2.96$$

Thus, the weighted mean computation shows that the mean cost per pound for the raw material is \$2.96. Note that using equation (3.14) rather than the weighted mean formula would have provided misleading results. In this case, the mean of the five cost-per-pound values is (3.00 + 3.40 + 2.80 + 2.90 + 3.25)/5 = 15.35/5 = \$3.07, which overstates the actual mean cost per pound purchased.

Computing a grade point average is a good example of the use of a weighted mean.

The choice of weights for a particular weighted mean computation depends upon the application. An example that is well known to college students is the computation of a grade point average (GPA). In this computation, the data values generally used are 4 for an A grade, 3 for a B grade, 2 for a C grade, 1 for a D grade, and 0 for an F grade. The weights are the number of credits hours earned for each grade. Exercise 54 at the end of this section provides an example of this weighted mean computation. In other weighted mean computations, quantities such as pounds, dollars, or volume are frequently used as weights. In any case, when observations vary in importance, the analyst must choose the weight that best reflects the importance of each observation in the determination of the mean.

Grouped Data

In most cases, measures of location and variability are computed by using the individual data values. Sometimes, however, data are available only in a grouped or frequency distribution form. In the following discussion, we show how the weighted mean formula can be used to obtain approximations of the mean, variance, and standard deviation for **grouped data**.

In Section 2.2 we provided a frequency distribution of the time in days required to complete year-end audits for the public accounting firm of Sanderson and Clifford. The frequency distribution of audit times is shown in Table 3.9. Based on this frequency distribution, what is the sample mean audit time?

To compute the mean using only the grouped data, we treat the midpoint of each class as being representative of the items in the class. Let M_i denote the midpoint for class i and let f_i denote the frequency of class i. The weighted mean formula (3.15) is then used with the data values denoted as M_i and the weights given by the frequencies f_i. In this case, the denominator of equation (3.15) is the sum of the frequencies, which is the

TABLE 3.9 FREQUENCY DISTRIBUTION OF AUDIT TIMES

Audit Time (days)	Frequency
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

sample size n. That is, $\Sigma f_i = n$. Thus, the equation for the sample mean for grouped data is as follows.

SAMPLE MEAN FOR GROUPED DATA

$$\bar{x} = \frac{\Sigma f_i M_i}{n} \tag{3.16}$$

where

M_i = the midpoint for class i
f_i = the frequency for class i
n = the sample size

With the class midpoints, M_i, halfway between the class limits, the first class of 10–14 in Table 3.9 has a midpoint at $(10 + 14)/2 = 12$. The five class midpoints and the weighted mean computation for the audit time data are summarized in Table 3.10. As can be seen, the sample mean audit time is 19 days.

To compute the variance for grouped data, we use a slightly altered version of the formula for the variance provided in equation (3.5). In equation (3.5), the squared deviations of the data about the sample mean $\bar{x}$ were written $(x_i - \bar{x})^2$. However, with grouped data, the values are not known. In this case, we treat the class midpoint, M_i, as being representative of the x_i values in the corresponding class. Thus, the squared deviations about the sample mean, $(x_i - \bar{x})^2$, are replaced by $(M_i - \bar{x})^2$. Then, just as we did with the sample mean calculations for grouped data, we weight each value by the frequency of the class, f_i. The sum of the squared deviations about the mean for all the data is approximated by $\Sigma f_i(M_i - \bar{x})^2$. The term $n - 1$ rather than n appears in the denominator in order to make the sample variance the estimate of the population variance. Thus, the following formula is used to obtain the sample variance for grouped data.

SAMPLE VARIANCE FOR GROUPED DATA

$$s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} \tag{3.17}$$

TABLE 3.10 COMPUTATION OF THE SAMPLE MEAN AUDIT TIME FOR GROUPED DATA

Audit Time (days)	Class Midpoint (M_i)	Frequency (f_i)	f_iM_i
10–14	12	4	48
15–19	17	8	136
20–24	22	5	110
25–29	27	2	54
30–34	32	1	32
		20	380

Sample mean $\bar{x} = \frac{\Sigma f_i M_i}{n} = \frac{380}{20} = 19$ days

TABLE 3.11 COMPUTATION OF THE SAMPLE VARIANCE OF AUDIT TIMES FOR GROUPED DATA (SAMPLE MEAN $\bar{x} = 19$)

Audit Time (days)	Class Midpoint (M_i)	Frequency (f_i)	Deviation ($M_i - \bar{x}$)	Squared Deviation $(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
10–14	12	4	−7	49	196
15–19	17	8	−2	4	32
20–24	22	5	3	9	45
25–29	27	2	8	64	128
30–34	32	1	13	169	169
		20			570
					$\Sigma f_i(M_i - \bar{x})^2$

$$\text{Sample variance } s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} = \frac{570}{19} = 30$$

The calculation of the sample variance for audit times based on the grouped data is shown in Table 3.11. The sample variance is 30.

The standard deviation for grouped data is simply the square root of the variance for grouped data. For the audit time data, the sample standard deviation is $s = \sqrt{30} = 5.48$.

Before closing this section on computing measures of location and dispersion for grouped data, we note that formulas (3.16) and (3.17) are for a sample. Population summary measures are computed similarly. The grouped data formulas for a population mean and variance follow.

POPULATION MEAN FOR GROUPED DATA

$$\mu = \frac{\Sigma f_i M_i}{N} \tag{3.18}$$

POPULATION VARIANCE FOR GROUPED DATA

$$\sigma^2 = \frac{\Sigma f_i(M_i - \mu)^2}{N} \tag{3.19}$$

NOTES AND COMMENTS

In computing descriptive statistics for grouped data, the class midpoints are used to approximate the data values in each class. As a result, the descriptive statistics for grouped data approximate the descriptive statistics that would result from using the original data directly. We therefore recommend computing descriptive statistics from the original data rather than from grouped data whenever possible.

Exercises

Methods

52. Consider the following data and corresponding weights.

x_i	Weight (w_i)
3.2	6
2.0	3
2.5	2
5.0	8

a. Compute the weighted mean.
b. Compute the sample mean of the four data values without weighting. Note the difference in the results provided by the two computations.

SELF test

53. Consider the sample data in the following frequency distribution.

Class	Midpoint	Frequency
3–7	5	4
8–12	10	7
13–17	15	9
18–22	20	5

a. Compute the sample mean.
b. Compute the sample variance and sample standard deviation.

Applications

SELF test

54. The grade point average for college students is based on a weighted mean computation. For most colleges, the grades are given the following data values: A (4), B (3), C (2), D (1), and F (0). After 60 credit hours of course work, a student at State University earned 9 credit hours of A, 15 credit hours of B, 33 credit hours of C, and 3 credit hours of D.
a. Compute the student's grade point average.
b. Students at State University must maintain a 2.5 grade point average for their first 60 credit hours of course work in order to be admitted to the business college. Will this student be admitted?

55. Morningstar tracks the total return for a large number of mutual funds. The following table shows the total return and the number of funds for four categories of mutual funds (*Morningstar Funds500*, 2008).

Type of Fund	Number of Funds	Total Return (%)
Domestic Equity	9191	4.65
International Equity	2621	18.15
Specialty Stock	1419	11.36
Hybrid	2900	6.75

a. Using the number of funds as weights, compute the weighted average total return for the mutual funds covered by Morningstar.
b. Is there any difficulty associated with using the "number of funds" as the weights in computing the weighted average total return for Morningstar in part (a)? Discuss. What else might be used for weights?
c. Suppose you had invested $10,000 in mutual funds at the beginning of 2007 and diversified the investment by placing $2000 in Domestic Equity funds, $4000 in

International Equity funds, $3000 in Specialty Stock funds, and $1000 in Hybrid funds. What is the expected return on the portfolio?

56. Based on a survey of 425 master's programs in business administration, the *U. S. News & World Report* ranked the Indiana University Kelley Business School as the 20th best business program in the country (*America's Best Graduate Schools,* 2009). The ranking was based in part on surveys of business school deans and corporate recruiters. Each survey respondent was asked to rate the overall academic quality of the master's program on a scale from 1 "marginal" to 5 "outstanding." Use the sample of responses shown below to compute the weighted mean score for the business school deans and the corporate recruiters. Discuss.

Quality Assessment	Business School Deans	Corporate Recruiters
5	44	31
4	66	34
3	60	43
2	10	12
1	0	0

57. The following frequency distribution shows the price per share of the 30 companies in the Dow Jones Industrial Average (*Barron's,* February 2, 2009).

Price per Share	Number of Companies
$0–9	4
$10–19	5
$20–29	7
$30–39	3
$40–49	4
$50–59	4
$60–69	0
$70–79	2
$80–89	0
$90–99	1

a. Compute the mean price per share and the standard deviation of the price per share for the Dow Jones Industrial Average companies.
b. On January 16, 2006, the mean price per share was $45.83 and the standard deviation was $18.14. Comment on the changes in the price per share over the three-year period.

Summary

In this chapter we introduced several descriptive statistics that can be used to summarize the location, variability, and shape of a data distribution. Unlike the tabular and graphical procedures introduced in Chapter 2, the measures introduced in this chapter summarize the data in terms of numerical values. When the numerical values obtained are for a sample, they are called sample statistics. When the numerical values obtained are for a population, they are called population parameters. Some of the notation used for sample statistics and population parameters follow.

In statistical inference, the sample statistic is referred to as the point estimator of the population parameter.

	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Variance	s^2	σ^2
Standard deviation	s	σ
Covariance	s_{xy}	σ_{xy}
Correlation	r_{xy}	ρ_{xy}

As measures of central location, we defined the mean, median, and mode. Then the concept of percentiles was used to describe other locations in the data set. Next, we presented the range, interquartile range, variance, standard deviation, and coefficient of variation as measures of variability or dispersion. Our primary measure of the shape of a data distribution was the skewness. Negative values indicate a data distribution skewed to the left. Positive values indicate a data distribution skewed to the right. We then described how the mean and standard deviation could be used, applying Chebyshev's theorem and the empirical rule, to provide more information about the distribution of data and to identify outliers.

In Section 3.4 we showed how to develop a five-number summary and a box plot to provide simultaneous information about the location, variability, and shape of the distribution. In Section 3.5 we introduced covariance and the correlation coefficient as measures of association between two variables. In the final section, we showed how to compute a weighted mean and how to calculate a mean, variance, and standard deviation for grouped data.

The descriptive statistics we discussed can be developed using statistical software packages and spreadsheets. In the chapter-ending appendixes we show how to use Minitab, Excel, and StatTools to develop the descriptive statistics introduced in this chapter.

Glossary

Sample statistic A numerical value used as a summary measure for a sample (e.g., the sample mean, $\bar{x}$, the sample variance, s^2, and the sample standard deviation, s).

Population parameter A numerical value used as a summary measure for a population (e.g., the population mean, μ, the population variance, σ^2, and the population standard deviation, σ).

Point estimator The sample statistic, such as $\bar{x}$, s^2, and s, when used to estimate the corresponding population parameter.

Mean A measure of central location computed by summing the data values and dividing by the number of observations.

Median A measure of central location provided by the value in the middle when the data are arranged in ascending order.

Mode A measure of location, defined as the value that occurs with greatest frequency.

Percentile A value such that at least p percent of the observations are less than or equal to this value and at least $(100 - p)$ percent of the observations are greater than or equal to this value. The 50th percentile is the median.

Quartiles The 25th, 50th, and 75th percentiles, referred to as the first quartile, the second quartile (median), and third quartile, respectively. The quartiles can be used to divide a data set into four parts, with each part containing approximately 25% of the data.

Range A measure of variability, defined to be the largest value minus the smallest value.

Interquartile range (IQR) A measure of variability, defined to be the difference between the third and first quartiles.

Variance A measure of variability based on the squared deviations of the data values about the mean.

Standard deviation A measure of variability computed by taking the positive square root of the variance.

Coefficient of variation A measure of relative variability computed by dividing the standard deviation by the mean and multiplying by 100.

Skewness A measure of the shape of a data distribution. Data skewed to the left result in negative skewness; a symmetric data distribution results in zero skewness; and data skewed to the right result in positive skewness.

z-score A value computed by dividing the deviation about the mean $(x_i - \bar{x})$ by the standard deviation s. A z-score is referred to as a standardized value and denotes the number of standard deviations x_i is from the mean.
Chebyshev's theorem A theorem that can be used to make statements about the proportion of data values that must be within a specified number of standard deviations of the mean.
Empirical rule A rule that can be used to compute the percentage of data values that must be within one, two, and three standard deviations of the mean for data that exhibit a bell-shaped distribution.
Outlier An unusually small or unusually large data value.
Five-number summary An exploratory data analysis technique that uses five numbers to summarize the data: smallest value, first quartile, median, third quartile, and largest value.
Box plot A graphical summary of data based on a five-number summary.
Covariance A measure of linear association between two variables. Positive values indicate a positive relationship; negative values indicate a negative relationship.
Correlation coefficient A measure of linear association between two variables that takes on values between -1 and $+1$. Values near $+1$ indicate a strong positive linear relationship; values near -1 indicate a strong negative linear relationship; and values near zero indicate the lack of a linear relationship.
Weighted mean The mean obtained by assigning each observation a weight that reflects its importance.
Grouped data Data available in class intervals as summarized by a frequency distribution. Individual values of the original data are not available.

Key Formulas

Sample Mean

$$\bar{x} = \frac{\Sigma x_i}{n} \tag{3.1}$$

Population Mean

$$\mu = \frac{\Sigma x_i}{N} \tag{3.2}$$

Interquartile Range

$$\text{IQR} = Q_3 - Q_1 \tag{3.3}$$

Population Variance

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \tag{3.4}$$

Sample Variance

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \tag{3.5}$$

Standard Deviation

$$\text{Sample standard deviation} = s = \sqrt{s^2} \tag{3.6}$$

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} \tag{3.7}$$

Coefficient of Variation

$$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\% \tag{3.8}$$

z-Score

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.9}$$

Sample Covariance

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{3.10}$$

Population Covariance

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{N} \tag{3.11}$$

Pearson Product Moment Correlation Coefficient: Sample Data

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3.12}$$

Pearson Product Moment Correlation Coefficient: Population Data

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.13}$$

Weighted Mean

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} \tag{3.15}$$

Sample Mean for Grouped Data

$$\bar{x} = \frac{\Sigma f_i M_i}{n} \tag{3.16}$$

Sample Variance for Grouped Data

$$s^2 = \frac{\Sigma f_i (M_i - \bar{x})^2}{n - 1} \tag{3.17}$$

Population Mean for Grouped Data

$$\mu = \frac{\Sigma f_i M_i}{N} \tag{3.18}$$

Population Variance for Grouped Data

$$\sigma^2 = \frac{\Sigma f_i (M_i - \mu)^2}{N} \tag{3.19}$$

Supplementary Exercises

58. According to an annual consumer spending survey, the average monthly Bank of America Visa credit card charge was $1838 (*U.S. Airways Attaché Magazine,* December 2003). A sample of monthly credit card charges provides the following data.

WEB file
Visa

236	1710	1351	825	7450
316	4135	1333	1584	387
991	3396	170	1428	1688

a. Compute the mean and median.
b. Compute the first and third quartiles.
c. Compute the range and interquartile range.
d. Compute the variance and standard deviation.
e. The skewness measure for these data is 2.12. Comment on the shape of this distribution. Is it the shape you would expect? Why or why not?
f. Do the data contain outliers?

59. The U.S. Census Bureau provides statistics on family life in the United States, including the age at the time of first marriage, current marital status, and size of household (U.S. Census Bureau website, March 20, 2006). The following data show the age at the time of first marriage for a sample of men and a sample of women.

WEB file
Ages

Men	26	23	28	25	27	30	26	35	28
	21	24	27	29	30	27	32	27	25
Women	20	28	23	30	24	29	26	25	
	22	22	25	23	27	26	19		

a. Determine the median age at the time of first marriage for men and women.
b. Compute the first and third quartiles for both men and women.
c. Twenty-five years ago the median age at the time of first marriage was 25 for men and 22 for women. What insight does this information provide about the decision of when to marry among young people today?

60. Dividend yield is the annual dividend per share a company pays divided by the current market price per share expressed as a percentage. A sample of 10 large companies provided the following dividend yield data (*The Wall Street Journal,* January 16, 2004).

Company	Yield %	Company	Yield %
Altria Group	5.0	General Motors	3.7
American Express	0.8	JPMorgan Chase	3.5
Caterpillar	1.8	McDonald's	1.6
Eastman Kodak	1.9	United Technology	1.5
ExxonMobil	2.5	Wal-Mart Stores	0.7

a. What are the mean and median dividend yields?
b. What are the variance and standard deviation?
c. Which company provides the highest dividend yield?
d. What is the z-score for McDonald's? Interpret this z-score.
e. What is the z-score for General Motors? Interpret this z-score.
f. Based on z-scores, do the data contain any outliers?

61. The U.S. Department of Education reports that about 50% of all college students use a student loan to help cover college expenses (National Center for Educational Studies, January 2006). A sample of students who graduated with student loan debt is shown here. The data, in thousands of dollars, show typical amounts of debt upon graduation.

10.1	14.8	5.0	10.2	12.4	12.2	2.0	11.5	17.8	4.0

a. For those students who use a student loan, what is the mean loan debt upon graduation?
b. What is the variance? Standard deviation?

62. Small business owners often look to payroll service companies to handle their employee payroll. Reasons are that small business owners face complicated tax regulations and penalties for employment tax errors are costly. According to the Internal Revenue Service, 26% of all small business employment tax returns contained errors that resulted in a tax penalty to the owner (*The Wall Street Journal,* January 30, 2006). The tax penalty for a sample of 20 small business owners follows:

WEB file
Penalty

820	270	450	1010	890	700	1350	350	300	1200
390	730	2040	230	640	350	420	270	370	620

a. What is the mean tax penalty for improperly filed employment tax returns?
b. What is the standard deviation?
c. Is the highest penalty, $2040, an outlier?
d. What are some of the advantages of a small business owner hiring a payroll service company to handle employee payroll services, including the employment tax returns?

63. Public transportation and the automobile are two methods an employee can use to get to work each day. Samples of times recorded for each method are shown. Times are in minutes.

Public Transportation:	28	29	32	37	33	25	29	32	41	34
Automobile:	29	31	33	32	34	30	31	32	35	33

a. Compute the sample mean time to get to work for each method.
b. Compute the sample standard deviation for each method.
c. On the basis of your results from parts (a) and (b), which method of transportation should be preferred? Explain.
d. Develop a box plot for each method. Does a comparison of the box plots support your conclusion in part (c)?

64. The National Association of Realtors reported the median home price in the United States and the increase in median home price over a five-year period (*The Wall Street Journal,* January 16, 2006). Use the sample home prices shown here to answer the following questions.

WEB file
Homes

995.9	48.8	175.0	263.5	298.0	218.9	209.0
628.3	111.0	212.9	92.6	2325.0	958.0	212.5

a. What is the sample median home price?
b. In January 2001, the National Association of Realtors reported a median home price of $139,300 in the United States. What was the percentage increase in the median home price over the five-year period?
c. What are the first quartile and the third quartile for the sample data?
d. Provide a five-number summary for the home prices.
e. Do the data contain any outliers?
f. What is the mean home price for the sample? Why does the National Association of Realtors prefer to use the median home price in its reports?

65. The U.S. Census Bureau's American Community Survey reported the percentage of children under 18 years of age who had lived below the poverty level during the previous 12 months (U.S. Census Bureau website, August 2008). The region of the country, Northeast (NE), Southeast (SE), Midwest (MW), Southwest (SW), and West (W) and the percentage of children under 18 who had lived below the poverty level are shown for each state.

WEB file
PovertyLevel

State	Region	Poverty %	State	Region	Poverty %
Alabama	SE	23.0	Montana	W	17.3
Alaska	W	15.1	Nebraska	MW	14.4
Arizona	SW	19.5	Nevada	W	13.9
Arkansas	SE	24.3	New Hampshire	NE	9.6
California	W	18.1	New Jersey	NE	11.8
Colorado	W	15.7	New Mexico	SW	25.6
Connecticut	NE	11.0	New York	NE	20.0
Delaware	NE	15.8	North Carolina	SE	20.2
Florida	SE	17.5	North Dakota	MW	13.0
Georgia	SE	20.2	Ohio	MW	18.7
Hawaii	W	11.4	Oklahoma	SW	24.3
Idaho	W	15.1	Oregon	W	16.8
Illinois	MW	17.1	Pennsylvania	NE	16.9
Indiana	MW	17.9	Rhode Island	NE	15.1
Iowa	MW	13.7	South Carolina	SE	22.1
Kansas	MW	15.6	South Dakota	MW	16.8
Kentucky	SE	22.8	Tennessee	SE	22.7
Louisiana	SE	27.8	Texas	SW	23.9
Maine	NE	17.6	Utah	W	11.9
Maryland	NE	9.7	Vermont	NE	13.2
Massachusetts	NE	12.4	Virginia	SE	12.2
Michigan	MW	18.3	Washington	W	15.4
Minnesota	MW	12.2	West Virginia	SE	25.2
Mississippi	SE	29.5	Wisconsin	MW	14.9
Missouri	MW	18.6	Wyoming	W	12.0

a. What is the median poverty level percentage for the 50 states?
b. What are the first and third quartiles? What is your interpretation of the quartiles?
c. Show a box plot for the data. Interpret the box plot in terms of what it tells you about the level of poverty for children in the United States. Are any states considered outliers? Discuss.
d. Identify the states in the lower quartile. What is your interpretation of this group and what region or regions are represented most in the lower quartile?

66. *Travel + Leisure* magazine presented its annual list of the 500 best hotels in the world (*Travel + Leisure,* January 2009). The magazine provides a rating for each hotel along with a brief description that includes the size of the hotel, amenities, and the cost per night for a double room. A sample of 12 of the top-rated hotels in the United States follows.

WEB file
Travel

Hotel	Location	Rooms	Cost/Night
Boulders Resort & Spa	Phoenix, AZ	220	499
Disney's Wilderness Lodge	Orlando, FL	727	340
Four Seasons Hotel Beverly Hills	Los Angeles, CA	285	585
Four Seasons Hotel	Boston, MA	273	495
Hay-Adams	Washington, DC	145	495
Inn on Biltmore Estate	Asheville, NC	213	279
Loews Ventana Canyon Resort	Phoenix, AZ	398	279
Mauna Lani Bay Hotel	Island of Hawaii	343	455
Montage Laguna Beach	Laguna Beach, CA	250	595
Sofitel Water Tower	Chicago, IL	414	367
St. Regis Monarch Beach	Dana Point, CA	400	675
The Broadmoor	Colorado Springs, CO	700	420

a. What is the mean number of rooms?
b. What is the mean cost per night for a double room?

c. Develop a scatter diagram with the number of rooms on the horizontal axis and the cost per night on the vertical axis. Does there appear to be a relationship between the number of rooms and the cost per night? Discuss.
d. What is the sample correlation coefficient? What does it tell you about the relationship between the number of rooms and the cost per night for a double room? Does this appear reasonable? Discuss.

67. Morningstar tracks the performance of a large number of companies and publishes an evaluation of each. Along with a variety of financial data, Morningstar includes a Fair Value estimate for the price that should be paid for a share of the company's common stock. Data for 30 companies are available in the file named FairValue. The data include the Fair Value estimate per share of common stock, the most recent price per share, and the earning per share for the company (*Morningstar Stocks500,* 2008).

WEB file
FairValue

a. Develop a scatter diagram for the Fair Value and Share Price data with Share Price on the horizontal axis. What is the sample correlation coefficient, and what can you say about the relationship between the variables?
b. Develop a scatter diagram for the Fair Value and Earnings per Share data with Earnings per Share on the horizontal axis. What is the sample correlation coefficient, and what can you say about the relationship between the variables?

68. Does a major league baseball team's record during spring training indicate how the team will play during the regular season? Over the last six years, the correlation coefficient between a team's winning percentage in spring training and its winning percentage in the regular season is .18 (*The Wall Street Journal,* March 30, 2009). Shown are the winning percentages for the 14 American League teams during the 2008 season.

WEB file
SpringTraining

Team	Spring Training	Regular Season	Team	Spring Training	Regular Season
Baltimore Orioles	.407	.422	Minnesota Twins	.500	.540
Boston Red Sox	.429	.586	New York Yankees	.577	.549
Chicago White Sox	.417	.546	Oakland A's	.692	.466
Cleveland Indians	.569	.500	Seattle Mariners	.500	.377
Detroit Tigers	.569	.457	Tampa Bay Rays	.731	.599
Kansas City Royals	.533	.463	Texas Rangers	.643	.488
Los Angeles Angels	.724	.617	Toronto Blue Jays	.448	.531

a. What is the correlation coefficient between the spring training and the regular season winning percentages?
b. What is your conclusion about a team's record during spring training indicating how the team will play during the regular season? What are some of the reasons why this occurs? Discuss.

69. The days to maturity for a sample of five money market funds are shown here. The dollar amounts invested in the funds are provided. Use the weighted mean to determine the mean number of days to maturity for dollars invested in these five money market funds.

Days to Maturity	Dollar Value ($millions)
20	20
12	30
7	10
5	15
6	10

70. Automobiles traveling on a road with a posted speed limit of 55 miles per hour are checked for speed by a state police radar system. Following is a frequency distribution of speeds.

Speed (miles per hour)	Frequency
45–49	10
50–54	40
55–59	150
60–64	175
65–69	75
70–74	15
75–79	10
Total	475

a. What is the mean speed of the automobiles traveling on this road?
b. Compute the variance and the standard deviation.

Case Problem 1 Pelican Stores

Pelican Stores, a division of National Clothing, is a chain of women's apparel stores operating throughout the country. The chain recently ran a promotion in which discount coupons were sent to customers of other National Clothing stores. Data collected for a sample of 100 in-store credit card transactions at Pelican Stores during one day while the promotion was running are contained in the file named PelicanStores. Table 3.12 shows a portion of the data set. The proprietary card method of payment refers to charges made using a National Clothing charge card. Customers who made a purchase using a discount

WEB file
PelicanStores

TABLE 3.12 SAMPLE OF 100 CREDIT CARD PURCHASES AT PELICAN STORES

Customer	Type of Customer	Items	Net Sales	Method of Payment	Gender	Marital Status	Age
1	Regular	1	39.50	Discover	Male	Married	32
2	Promotional	1	102.40	Proprietary Card	Female	Married	36
3	Regular	1	22.50	Proprietary Card	Female	Married	32
4	Promotional	5	100.40	Proprietary Card	Female	Married	28
5	Regular	2	54.00	MasterCard	Female	Married	34
6	Regular	1	44.50	MasterCard	Female	Married	44
7	Promotional	2	78.00	Proprietary Card	Female	Married	30
8	Regular	1	22.50	Visa	Female	Married	40
9	Promotional	2	56.52	Proprietary Card	Female	Married	46
10	Regular	1	44.50	Proprietary Card	Female	Married	36
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
96	Regular	1	39.50	MasterCard	Female	Married	44
97	Promotional	9	253.00	Proprietary Card	Female	Married	30
98	Promotional	10	287.59	Proprietary Card	Female	Married	52
99	Promotional	2	47.60	Proprietary Card	Female	Married	30
100	Promotional	1	28.44	Proprietary Card	Female	Married	44

coupon are referred to as promotional customers and customers who made a purchase but did not use a discount coupon are referred to as regular customers. Because the promotional coupons were not sent to regular Pelican Stores customers, management considers the sales made to people presenting the promotional coupons as sales it would not otherwise make. Of course, Pelican also hopes that the promotional customers will continue to shop at its stores.

Most of the variables shown in Table 3.12 are self-explanatory, but two of the variables require some clarification.

Items	The total number of items purchased
Net Sales	The total amount ($) charged to the credit card

Pelican's management would like to use this sample data to learn about its customer base and to evaluate the promotion involving discount coupons.

Managerial Report

Use the methods of descriptive statistics presented in this chapter to summarize the data and comment on your findings. At a minimum, your report should include the following:

1. Descriptive statistics on net sales and descriptive statistics on net sales by various classifications of customers.
2. Descriptive statistics concerning the relationship between age and net sales.

Case Problem 2 Motion Picture Industry

The motion picture industry is a competitive business. More than 50 studios produce a total of 300 to 400 new motion pictures each year, and the financial success of each motion picture varies considerably. The opening weekend gross sales, the total gross sales, the number of theaters the movie was shown in, and the number of weeks the motion picture was in the top 60 for gross sales are common variables used to measure the success of a motion picture. Data collected for a sample of 100 motion pictures produced in 2005 are contained in the file named Movies. Table 3.13 shows the data for the first 10 motion pictures in the file.

WEB file
Movies

TABLE 3.13 PERFORMANCE DATA FOR 10 MOTION PICTURES

Motion Picture	Opening Gross Sales ($millions)	Total Gross Sales ($millions)	Number of Theaters	Weeks in Top 60
Coach Carter	29.17	67.25	2574	16
Ladies in Lavender	0.15	6.65	119	22
Batman Begins	48.75	205.28	3858	18
Unleashed	10.90	24.47	1962	8
Pretty Persuasion	0.06	0.23	24	4
Fever Pitch	12.40	42.01	3275	14
Harry Potter and the Goblet of Fire	102.69	287.18	3858	13
Monster-in-Law	23.11	82.89	3424	16
White Noise	24.11	55.85	2279	7
Mr. and Mrs. Smith	50.34	186.22	3451	21

Managerial Report

Use the numerical methods of descriptive statistics presented in this chapter to learn how these variables contribute to the success of a motion picture. Include the following in your report.

1. Descriptive statistics for each of the four variables along with a discussion of what the descriptive statistics tell us about the motion picture industry.
2. What motion pictures, if any, should be considered high-performance outliers? Explain.
3. Descriptive statistics showing the relationship between total gross sales and each of the other variables. Discuss.

Case Problem 3 Business Schools of Asia-Pacific

WEB file
Asian

The pursuit of a higher education degree in business is now international. A survey shows that more and more Asians choose the master of business administration (MBA) degree route to corporate success. As a result, the number of applicants for MBA courses at Asia-Pacific schools continues to increase.

Across the region, thousands of Asians show an increasing willingness to temporarily shelve their careers and spend two years in pursuit of a theoretical business qualification. Courses in these schools are notoriously tough and include economics, banking, marketing, behavioral sciences, labor relations, decision making, strategic thinking, business law, and more. The data set in Table 3.14 shows some of the characteristics of the leading Asia-Pacific business schools.

Managerial Report

Use the methods of descriptive statistics to summarize the data in Table 3.14. Discuss your findings.

1. Include a summary for each variable in the data set. Make comments and interpretations based on maximums and minimums, as well as the appropriate means and proportions. What new insights do these descriptive statistics provide concerning Asia-Pacific business schools?
2. Summarize the data to compare the following:
   a. Any difference between local and foreign tuition costs.
   b. Any difference between mean starting salaries for schools requiring and not requiring work experience.
   c. Any difference between starting salaries for schools requiring and not requiring English tests.
3. Do starting salaries appear to be related to tuition?
4. Present any additional graphical and numerical summaries that will be beneficial in communicating the data in Table 3.14 to others.

Case Problem 4 Heavenly Chocolates Website Transactions

Heavenly Chocolates manufactures and sells quality chocolate products at its plant and retail store located in Saratoga Springs, New York. Two years ago the company developed a website and began selling its products over the Internet. Website sales have exceeded the company's expectations, and mangement is now considering stragegies to increase sales even further. To learn more about the website customers, a sample of 50 Heavenly Chocolate transactions was selected from the previous month's sales. Data showing the day

TABLE 3.14 DATA FOR 25 ASIA-PACIFIC BUSINESS SCHOOLS

Business School	Full-Time Enrollment	Students per Faculty	Local Tuition ($)	Foreign Tuition ($)	Age	%Foreign	GMAT	English Test	Work Experience	Starting Salary ($)
Melbourne Business School	200	5	24,420	29,600	28	47	Yes	No	Yes	71,400
University of New South Wales (Sydney)	228	4	19,993	32,582	29	28	Yes	No	Yes	65,200
Indian Institute of Management (Ahmedabad)	392	5	4,300	4,300	22	0	No	No	No	7,100
Chinese University of Hong Kong	90	5	11,140	11,140	29	10	Yes	No	No	31,000
International University of Japan (Niigata)	126	4	33,060	33,060	28	60	Yes	Yes	No	87,000
Asian Institute of Management (Manila)	389	5	7,562	9,000	25	50	Yes	No	Yes	22,800
Indian Institute of Management (Bangalore)	380	5	3,935	16,000	23	1	Yes	No	No	7,500
National University of Singapore	147	6	6,146	7,170	29	51	Yes	Yes	Yes	43,300
Indian Institute of Management (Calcutta)	463	8	2,880	16,000	23	0	No	No	No	7,400
Australian National University (Canberra)	42	2	20,300	20,300	30	80	Yes	Yes	Yes	46,600
Nanyang Technological University (Singapore)	50	5	8,500	8,500	32	20	Yes	No	Yes	49,300
University of Queensland (Brisbane)	138	17	16,000	22,800	32	26	No	No	Yes	49,600
Hong Kong University of Science and Technology	60	2	11,513	11,513	26	37	Yes	No	Yes	34,000
Macquarie Graduate School of Management (Sydney)	12	8	17,172	19,778	34	27	No	No	Yes	60,100
Chulalongkorn University (Bangkok)	200	7	17,355	17,355	25	6	Yes	No	Yes	17,600
Monash Mt. Eliza Business School (Melbourne)	350	13	16,200	22,500	30	30	Yes	Yes	Yes	52,500
Asian Institute of Management (Bangkok)	300	10	18,200	18,200	29	90	No	Yes	Yes	25,000
University of Adelaide	20	19	16,426	23,100	30	10	No	No	Yes	66,000
Massey University (Palmerston North, New Zealand)	30	15	13,106	21,625	37	35	No	Yes	Yes	41,400
Royal Melbourne Institute of Technology Business Graduate School	30	7	13,880	17,765	32	30	No	Yes	Yes	48,900
Jamnalal Bajaj Institute of Management Studies (Mumbai)	240	9	1,000	1,000	24	0	No	No	Yes	7,000
Curtin Institute of Technology (Perth)	98	15	9,475	19,097	29	43	Yes	No	Yes	55,000
Lahore University of Management Sciences	70	14	11,250	26,300	23	2.5	No	No	No	7,500
Universiti Sains Malaysia (Penang)	30	5	2,260	2,260	32	15	No	Yes	Yes	16,000
De La Salle University (Manila)	44	17	3,300	3,600	28	3.5	Yes	No	Yes	13,100

WEB file
Shoppers

TABLE 3.15 A SAMPLE OF 50 HEAVENLY CHOCOLATES WEBSITE TRANSACTIONS

Customer	Day	Browser	Time (min)	Pages Viewed	Amount Spent ($)
1	Mon	Internet Explorer	12.0	4	54.52
2	Wed	Other	19.5	6	94.90
3	Mon	Internet Explorer	8.5	4	26.68
4	Tue	Firefox	11.4	2	44.73
5	Wed	Internet Explorer	11.3	4	66.27
6	Sat	Firefox	10.5	6	67.80
7	Sun	Internet Explorer	11.4	2	36.04
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
48	Fri	Internet Explorer	9.7	5	103.15
49	Mon	Other	7.3	6	52.15
50	Fri	Internet Explorer	13.4	3	98.75

of the week each transaction was made, the type of browser the customer used, the time spent on the website, the number of website pages viewed, and the amount spent by each of the 50 customers are contained in the file named Shoppers. A portion of the data are shown in Table 3.15.

Heavenly Chocolates would like to use the sample data to determine if online shoppers who spend more time and view more pages also spend more money during their visit to the website. The company would also like to investigate the effect that the day of the week and the type of browser have on sales.

Managerial Report

Use the methods of descriptive statistics to learn about the customers who visit the Heavenly Chocolates website. Include the following in your report.

1. Graphical and numerical summaries for the length of time the shopper spends on the website, the number of pages viewed, and the mean amount spent per transaction. Discuss what you learn about Heavenly Cholcolates' online shoppers from these numerical summaries.
2. Summarize the frequency, the total dollars spent, and the mean amount spent per transaction for each day of week. What observations can you make about Hevenly Chocolates' business based on the day of the week? Discuss.
3. Summarize the frequency, the total dollars spent, and the mean amount spent per transaction for each type of browser. What observations can you make about Heavenly Chocolate's business based on the type of browser? Discuss.
4. Develop a scatter diagram and compute the sample correlation coefficient to explore the relationship between the time spent on the website and the dollar amount spent. Use the horizontal axis for the time spent on the website. Discuss.
5. Develop a scatter diagram and compute the sample correlation coefficient to explore the relationship between the the number of website pages viewed and the amount spent. Use the horizontal axis for the number of website pages viewed. Discuss.
6. Develop a scatter diagram and compute the sample correlation coefficient to explore the relationship between the time spent on the website and the number of pages viewed. Use the horizontal axis to represent the number of pages viewed. Discuss.

Appendix 3.1 Descriptive Statistics Using Minitab

In this appendix, we describe how Minitab can be used to compute a variety of descriptive statistics and display box plots. We then show how Minitab can be used to obtain covariance and correlation measures for two variables.

Descriptive Statistics

Table 3.1 provided the starting salaries for 12 business school graduates. These data are available in the file StartSalary. Figure 3.12 shows the descriptive statistics for the salary data obtained by using Minitab. Definitions of the headings follow.

N	number of data values
N*	number of missing data values
Mean	mean
SE Mean	standard error of mean
StDev	standard deviation
Minimum	minimum data value
Q1	first quartile
Median	median
Q3	third quartile
Maximum	maximum data value

The label SE Mean refers to the *standard error of the mean.* It is computed by dividing the standard deviation by the square root of N. The interpretation and use of this measure are discussed in Chapter 7 when we introduce the topics of sampling and sampling distributions.

Although the numerical measures of range, interquartile range, variance, and coefficient of variation do not appear on the Minitab output, these values can be easily computed from the results in Figure 3.12 as follows.

$$\begin{aligned}\text{Range} &= \text{Maximum} - \text{Minimum}\\ \text{IQR} &= Q_3 - Q_1\\ \text{Variance} &= (\text{StDev})^2\\ \text{Coefficient of Variation} &= (\text{StDev/Mean}) \times 100\end{aligned}$$

Finally, note that Minitab's quartiles $Q_1 = 3457.5$ and $Q_3 = 3625$ are slightly different from the quartiles $Q_1 = 3465$ and $Q_3 = 3600$ computed in Section 3.1. The different conventions* used to identify the quartiles explain this variation. Hence, the values of Q_1 and Q_3 provided by one convention may not be identical to the values of Q_1 and Q_3 provided

FIGURE 3.12 DESCRIPTIVE STATISTICS PROVIDED BY MINITAB

N	N*	Mean	SEMean	StDev
12	0	3540.0	47.8	165.7
Minimum	**Q1**	**Median**	**Q3**	**Maximum**
3310.0	3457.5	3505.0	3625.0	3925.0

*With the n observations arranged in ascending order (smallest value to largest value), Minitab uses the positions given by $(n + 1)/4$ and $3(n + 1)/4$ to locate Q_1 and Q_3, respectively. When a position is fractional, Minitab interpolates between the two adjacent ordered data values to determine the corresponding quartile.

by another convention. Any differences tend to be negligible, however, and the results provided should not mislead the user in making the usual interpretations associated with quartiles.

WEB file
StartSalary

Let us show how the statistics in Figure 3.12 are generated. The starting salary data are in column C2 of the StartSalary worksheet. The following steps can be used to generate the descriptive statistics.

Step 1. Select the **Stat** menu
Step 2. Choose **Basic Statistics**
Step 3. Choose **Display Descriptive Statistics**
Step 4. When the Display Descriptive Statistics dialog box appears:
Enter C2 in the **Variables** box
Click **OK**

Box Plot

The following steps use the file StartSalary to generate the box plot for the starting salary data.

Step 1. Select the **Graph** menu
Step 2. Choose **Boxplot**
Step 3. Select **Simple** and click **OK**
Step 4. When the Boxplot-One Y, Simple dialog box appears:
Enter C2 in the **Graph variables** box
Click **OK**

Covariance and Correlation

WEB file
Stereo

Table 3.6 provided for the number of commercials and the sales volume for a stereo and sound equipment store. These data are available in the file Stereo, with the number of commercials in column C2 and the sales volume in column C3. The following steps show how Minitab can be used to compute the covariance for the two variables.

Step 1. Select the **Stat** menu
Step 2. Choose **Basic Statistics**
Step 3. Choose **Covariance**
Step 4. When the Covariance dialog box appears:
Enter C2 C3 in the **Variables** box
Click **OK**

To obtain the correlation coefficient for the number of commercials and the sales volume, only one change is necessary in the preceding procedure. In step 3, choose the **Correlation** option.

Appendix 3.2 Descriptive Statistics Using Excel

Excel can be used to generate the descriptive statistics discussed in this chapter. We show how Excel can be used to generate several measures of location and variability for a single variable and to generate the covariance and correlation coefficient as measures of association between two variables.

Using Excel Functions

Excel provides functions for computing the mean, median, mode, sample variance, and sample standard deviation. We illustrate the use of these Excel functions by computing the mean, median,

FIGURE 3.13 USING EXCEL FUNCTIONS FOR COMPUTING THE MEAN, MEDIAN, MODE, VARIANCE, AND STANDARD DEVIATION

	A	B	C	D	E	F
1	Graduate	Starting Salary		Mean	=AVERAGE(B2:B13)	
2	1	3450		Median	=MEDIAN(B2:B13)	
3	2	3550		Mode	=MODE(B2:B13)	
4	3	3650		Variance	=VAR(B2:B13)	
5	4	3480		Standard Deviation	=STDEV(B2:B13)	
6	5	3355				
7	6	3310				
8	7	3490				
9	8	3730				
10	9	3540				
11	10	3925				
12	11	3520				
13	12	3480				
14						

	A	B	C	D	E	F
1	Graduate	Starting Salary		Mean	3540	
2	1	3450		Median	3505	
3	2	3550		Mode	3480	
4	3	3650		Variance	27440.91	
5	4	3480		Standard Deviation	165.65	
6	5	3355				
7	6	3310				
8	7	3490				
9	8	3730				
10	9	3540				
11	10	3925				
12	11	3520				
13	12	3480				
14						

WEB file
StartSalary

mode, sample variance, and sample standard deviation for the starting salary data in Table 3.1. Refer to Figure 3.13 as we describe the steps involved. The data are entered in column B.

Excel's AVERAGE function can be used to compute the mean by entering the following formula into cell E1:

=AVERAGE(B2:B13)

Similarly, the formulas =MEDIAN(B2:B13), =MODE(B2:B13), =VAR(B2:B13), and =STDEV(B2:B13) are entered into cells E2:E5, respectively, to compute the median, mode, variance, and standard deviation. The worksheet in the foreground shows that the values computed using the Excel functions are the same as we computed earlier in the chapter.

Excel also provides functions that can be used to compute the covariance and correlation coefficient. You must be careful when using these functions because the covariance function treats the data as a population and the correlation function treats the data as a sample. Thus, the result obtained using Excel's covariance function must be adjusted to provide the sample covariance. We show here how these functions can be used to compute the sample covariance and the sample correlation coefficient for the stereo and sound equipment store data in Table 3.7. Refer to Figure 3.14 as we present the steps involved.

WEB file
Stereo

Excel's covariance function, COVAR, can be used to compute the population covariance by entering the following formula into cell F1:

=COVAR(B2:B11,C2:C11)

Similarly, the formula =CORREL(B2:B11,C2:C11) is entered into cell F2 to compute the sample correlation coefficient. The worksheet in the foreground shows the values computed

FIGURE 3.14 USING EXCEL FUNCTIONS FOR COMPUTING COVARIANCE AND CORRELATION

	A	B	C	D	E	F	G
1	Week	Commercials	Sales		Population Covariance	=COVAR(B2:B11,C2:C11)	
2	1	2	50		Sample Correlation	=CORREL(B2:B11,C2:C11)	
3	2	5	57				
4	3	1	41				
5	4	3	54				
6	5	4	54				
7	6	1	38				
8	7	5	63				
9	8	3	48				
10	9	4	59				
11	10	2	46				
12							

	A	B	C	D	E	F	G
1	Week	Commercials	Sales		Population Covariance	9.90	
2	1	2	50		Sample Correlation	0.93	
3	2	5	57				
4	3	1	41				
5	4	3	54				
6	5	4	54				
7	6	1	38				
8	7	5	63				
9	8	3	48				
10	9	4	59				
11	10	2	46				
12							

using the Excel functions. Note that the value of the sample correlation coefficient (.93) is the same as computed using equation (3.12). However, the result provided by the Excel COVAR function, 9.9, was obtained by treating the data as a population. Thus, we must adjust the Excel result of 9.9 to obtain the sample covariance. The adjustment is rather simple. First, note that the formula for the population covariance, equation (3.11), requires dividing by the total number of observations in the data set. But the formula for the sample covariance, equation (3.10), requires dividing by the total number of observations minus 1. So, to use the Excel result of 9.9 to compute the sample covariance, we simply multiply 9.9 by $n/(n-1)$. Because $n = 10$, we obtain

$$s_{xy} = \left(\frac{10}{9}\right)9.9 = 11$$

Thus, the sample covariance for the stereo and sound equipment data is 11.

Using Excel's Descriptive Statistics Tool

WEB file
StartSalary

As we already demonstrated, Excel provides statistical functions to compute descriptive statistics for a data set. These functions can be used to compute one statistic at a time (e.g., mean, variance, etc.). Excel also provides a variety of Data Analysis Tools. One of these tools, called Descriptive Statistics, allows the user to compute a variety of descriptive statistics at once. We show here how it can be used to compute descriptive statistics for the starting salary data in Table 3.1.

Step 1. Click the **Data** tab on the Ribbon
Step 2. In the **Analysis** group, click **Data Analysis**
Step 3. When the Data Analysis dialog box appears:
Choose **Descriptive Statistics**
Click **OK**

FIGURE 3.15 EXCEL'S DESCRIPTIVE STATISTICS TOOL OUTPUT

	A	B	C	D	E	F
1	**Graduate**	**Starting Salary**		***Starting Salary***		
2	1	3450				
3	2	3550		**Mean**	3540	
4	3	3650		Standard Error	47.82	
5	4	3480		**Median**	3505	
6	5	3355		**Mode**	3480	
7	6	3310		**Standard Deviation**	165.65	
8	7	3490		**Sample Variance**	27440.91	
9	8	3730		Kurtosis	1.7189	
10	9	3540		**Skewness**	1.0911	
11	10	3925		**Range**	615	
12	11	3520		**Minimum**	3310	
13	12	3480		**Maximum**	3925	
14				**Sum**	42480	
15				**Count**	12	
16						

Step 4. When the Descriptive Statistics dialog box appears:
Enter B1:B13 in the **Input Range** box
Select **Grouped By Columns**
Select **Labels in First Row**
Select **Output Range**
Enter D1 in the **Output Range** box (to identify the upper left-hand corner of the section of the worksheet where the descriptive statistics will appear)
Select **Summary statistics**
Click **OK**

Cells D1:E15 of Figure 3.15 show the descriptive statistics provided by Excel. The boldface entries are the descriptive statistics we covered in this chapter. The descriptive statistics that are not boldface are either covered subsequently in the text or discussed in more advanced texts.

Appendix 3.3 Descriptive Statistics Using StatTools

In this appendix, we describe how StatTools can be used to compute a variety of descriptive statistics and also display box plots. We then show how StatTools can be used to obtain covariance and correlation measures for two variables.

Descriptive Statistics

WEB file
StartSalary

We use the starting salary data in Table 3.1 to illustrate. Begin by using the Data Set Manager to create a StatTools data set for these data using the procedure described in the appendix in Chapter 1. The following steps will generate a variety of descriptive statistics.

Step 1. Click the **StatTools** tab on the Ribbon
Step 2. In the **Analyses Group,** click **Summary Statistics**
Step 3. Choose the **One-Variable Summary** option

Step 4. When the One-Variable Summary Statistics dialog box appears:
In the **Variables** section, select **Starting Salary**
Click **OK**

A variety of descriptive statistics will appear.

Box Plots

We use the starting salary data in Table 3.1 to illustrate. Begin by using the Data Set Manager to create a StatTools data set for these data using the procedure described in the appendix in Chapter 1. The following steps will create a box plot for these data.

WEB file
StartSalary

Step 1. Click the **StatTools** tab on the Ribbon
Step 2. In the **Analyses Group,** click **Summary Graphs**
Step 3. Choose the **Box-Whisker Plot** option
Step 4. When the StatTools—Box-Whisker Plot dialog box appears:
In the Variables section, select **Starting Salary**
Click **OK**

The symbol □ is used to identify an outlier and x is used to identify the mean.

Covariance and Correlation

We use the stereo and sound equipment data in Table 3.7 to demonstrate the computation of the sample covariance and the sample correlation coefficient. Begin by using the Data Set Manager to create a StatTools data set for these data using the procedure described in the appendix in Chapter 1. The following steps will provide the sample covariance and sample correlation coefficient.

WEB file
Stereo

Step 1. Click the **StatTools** tab on the Ribbon
Step 2. In the **Analyses Group,** click **Summary Statistics**
Step 3. Choose the **Correlation and Covariance** option
Step 4. When the StatTools—Correlation and Covariance dialog box appears:
In the **Variables** section
Select **No. of Commercials**
Select **Sales Volume**
In the **Tables to Create** section,
Select **Table of Correlations**
Select **Table of Covariances**
In the **Table Structure** section select **Symmetric**
Click **OK**

A table showing the correlation coefficient and the covariance will appear.