3 Descriptive Statistics: Numerical Measures

CHAPTER CONTENTS

Statistics in Practice TV audience measurement

LEARNING OBJECTIVES

After studying this chapter and doing the exercises, you should be able to calculate and interpret the following statistical measures that help to describe the central location, variability and shape of data sets.

1 The mean, median and mode.

2 Percentiles (including quartiles), the range, the interquartile range, the variance, the standard deviation and the coefficient of variation.

3 You should understand the concept of skewness of distribution. You should be able to calculate *z*-scores and understand their role in identifying data outliers.

4 You should understand the role of Chebyshev's theorem and of the empirical rule in estimating the spread of data sets.

5 Five-number summaries and box plots.

6 Scatter diagrams, covariance and Pearson's correlation coefficient.

7 Weighted means.

8 Estimates of mean and standard deviation for grouped data.

In Chapter 2 we discussed tabular and graphical data summaries. In this chapter, we present several numerical measures for summarizing data.

We start with numerical summary measures for 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 shall also examine measures of the relationship between the variables.

STATISTICS IN PRACTICE

TV audience measurement

Television audience levels and audience share are important issues for advertisers, sponsors and, in the case of public service broadcasting, governments. In recent years in many countries, the number of TV channels available has increased substantially because of the use of digital, satellite and cable services. The Broadcasters' Audience Research Board (BARB) in the UK, for example, lists over 250 channels in its 'multi-channel viewing summary'. Technology also now allows viewers to 'time-shift' their viewing. Accurate audience measurement thereby becomes a more difficult task.

The *Handbook on Radio and Television Audience Research*[*] has a section on data analysis, in which the author makes the point 'most audience research is quantitative'. He then goes on to describe the various measures that are commonly used in this field, including: 'ratings', 'gross rating points', 'viewing share', 'viewing hours' and 'reach'. Many of the measures involve the use of averages: for example, 'average weekly viewing per person'.

BARB publishes viewing figures on its website, www.barb.co.uk. Figures for the week ending 22 July 2012, for example, a week before the start of the 2012 London Olympics, showed that 'average daily reach' for the lead public broadcasting channel BBC1 was just over 26 million viewers. This represented about 45 per cent of the potential viewing audience. Average weekly viewing for BBC1 was estimated at slightly under five hours per person. Two weeks later, with the 2012 Olympics in full swing and TV coverage being provided by the BBC, average daily reach for BBC1 had risen to 32 million viewers, and average viewing time had more than doubled to over ten hours per person.

In this chapter, you will learn how to compute and interpret some of the statistical measures used in reports such as those presented by BARB. You will learn about the mean, median and mode, and 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.

[*] *Handbook on Radio and Television Audience Research*, by Graham Mytton, published by UNICEF/UNESCO/BBC World Service Training Trust, web edition (2007).

We introduce numerical measures of location, dispersion, shape and association. If they are computed for sample data, they are called **sample statistics**. If they are computed for data for a whole 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 shall discuss in more detail the process of point estimation. In the guides on the associated online platform, we show how EXCEL, IBM SPSS and MINITAB can be used to compute many of the numerical measures described in the chapter.

3.1 MEASURES OF LOCATION

Mean

The most commonly used measure of location is the **mean**. The mean provides a measure of central location for the data. If the data are from a sample, the mean is denoted by putting a bar over the data symbol, e.g. $\bar{x}$. If the data are from a population, the Greek letter μ (mu) is used to denote the mean. When people refer to the 'average' value, they are usually referring to the mean value.

In statistical formulae, it is customary to denote the value of variable X for the first sample observation by x_1, for the second sample observation by x_2 and so on. In general, the value of variable X for the ith observation is denoted by x_i. (As we shall see in Chapters 5 and 6, a common convention in statistics is to *name* variables using capital letters, e.g. X, but to refer to specific values of those variables using small letters, e.g. x.) For a sample with n observations, the formula for the sample mean is as follows:

Sample mean

$$\bar{x} = \frac{\sum x_i}{n} \qquad \textbf{(3.1)}$$

In equation (3.1), the numerator is the sum of the values of the n observations. That is:

$$\sum x_i = x_1 + x_2 + \ldots + x_n$$

The Greek letter Σ (upper case sigma) is the summation sign.

To illustrate the computation of a sample mean, consider the following class size data for a sample of five university 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.

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

To compute the sample mean, we can write:

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

The sample mean class size is 44 students.

Here is a second illustration. Suppose a university careers office has sent a questionnaire to a sample of business school graduates requesting information on monthly starting salaries. Table 3.1 shows the data collected. The mean monthly starting salary for the sample of 12 business school graduates is computed as:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \ldots + x_{12}}{12} = \frac{2020 + 2075 + \ldots + 2040}{12} = \frac{24840}{12} = 2070$$

SALARY

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. We denote the number of observations in a population by N, and the population mean as μ.

Population mean

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

Median

The **median** is another measure of central location for a variable. The median is the value in the middle when the data are arranged in ascending order (smallest value to largest value).

Median

Arrange the data in ascending order.

a. For an odd number of observations, the median is the middle value.

b. For even number of observations, the median is the average of the two middle values.

Let us compute the median class size for the sample of five university classes. We first arrange the data in ascending order.

32 42 46 46 54

Because $n = 5$ is odd, the median is the middle value. This data set contains two observations with values of 46 (the 3rd and 4th ordered observations). Each observation is treated separately when we arrange the data in ascending order. The median class size is 46 students (the 3rd ordered observation).

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

1955 1980 2020 2040 2040 2050 2060 2070 2075 2125 2165 2260

Middle two values

Because $n = 12$ is even, we identify the middle two values: 2050 and 2060. The median is the average of these values:

$$\text{Median} = \frac{2050 + 2060}{2} = 2055$$

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 example, suppose one of the graduates (see Table 3.1) had a starting salary of €5000 per month (perhaps his/her family owns the company). If we change the highest monthly starting salary in Table 3.1 from €2260 to €5000, the sample mean changes from €2070 to €2298. The median of €2055, however, is unchanged, because €2050 and €2060 are still the middle two values. With the extremely high starting salary included, the median provides a more robust measure of central location than the mean. When a data set contains extreme values, the median is often the preferred measure of central location.*

TABLE 3.1 Monthly starting salaries for a sample of 12 business school graduates

Graduate	Monthly starting salary (€)	Graduate	Monthly starting salary (€)
1	2020	7	2050
2	2075	8	2165
3	2125	9	2070
4	2040	10	2260
5	1980	11	2060
6	1955	12	2040

*Another measure sometimes used when extreme values are present is the *trimmed mean*. A percentage of the smallest and largest values are removed from a data set, and the mean of the remaining values is computed. For example, to get the 5 per cent trimmed mean, the smallest 5 per cent and the largest 5 per cent of the data values are removed, and the mean of the remaining values is computed. Using the sample with $n = 12$ starting salaries, $0.05(12) = 0.6$. Rounding this value to 1 indicates that the 5 per cent trimmed mean would remove the smallest data value and the largest data value. The 5 per cent trimmed mean using the 10 remaining observations is 2062.5.

Mode

A third measure of location is the **mode** (although the mode does not necessarily measure *central* location). The mode is defined as follows.

Mode

The mode is the value that occurs with the 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. This value occurs twice and consequently is the mode. In the sample of starting salaries for the business school graduates, the only monthly starting salary that occurs more than once is €2040, and therefore this value is the mode for that data set.

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 modes are almost never reported, because listing three or more modes would not be particularly helpful in describing a central location for the data.

The mode is an important measure of location for qualitative data. For example, the qualitative data set in Table 2.2 resulted in the following frequency distribution for new car purchases.

Car brand	*Frequency*
Audi	8
BMW	5
Mercedes	13
Opel	8
VW	19
Total	50

The mode, or most frequently purchased car brand, is VW. For this type of data it obviously makes no sense to speak of the mean or median. The mode provides the information of interest, the most frequently purchased car brand.

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 per cent of the observations have values less than the pth percentile; approximately $(100 - p)$ per cent 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 per cent of the observations are less than or equal to this value and *at least* $(100 - p)$ per cent of the observations are greater than or equal to this value.

Colleges and universities sometimes 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. It may not be readily apparent how this student performed in relation to other students taking the same test. However,

if the raw score of 54 corresponds to the 70th percentile, we know that approximately 70 per cent of the students scored lower than this individual and approximately 30 per cent of the students scored higher than this individual.

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

Calculating the *p*th percentile

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

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, consider the 85th percentile for the starting salary data in Table 3.1.

1 Arrange the data in ascending order.

1955 1980 2020 2040 2040 2050 2060 2070 2075 2125 2165 2260

2

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

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 2165.

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

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

Because i is an integer, step 3(b) states that the 50th percentile is the average of the sixth and seventh data values; that is (2050 + 2060)/2 = 2055. Note that the *50th percentile is also the median.*

Quartiles

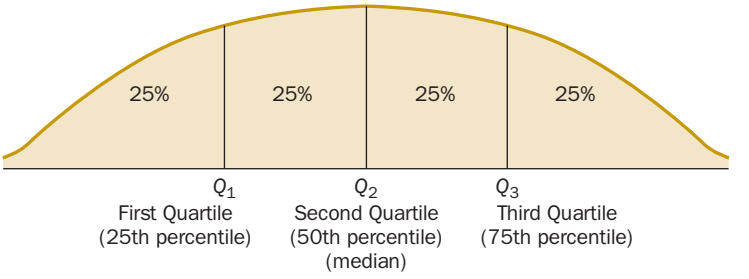
For the purposes of describing data distribution, it is often useful to consider the values that divide the data set into four parts, with each part containing approximately one-quarter (25 per cent) of the observations. Figure 3.1 shows a data distribution divided into four parts. The division points 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

FIGURE 3.1
Location of the quartiles



The starting salary data are again arranged in ascending order. We have already identified Q_2, the second quartile (median), as 2055.

1955 1980 2020 2040 2040 2050 2060 2070 2075 2125 2165 2260

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

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; hence:

$$Q_1 = (2020 + 2040)/2 = 2030$$

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; hence:

$$Q_3 = (2075 + 2125)/2 = 2100$$

1955 1980 2020 | 2040 2040 2050 | 2060 2070 2075 | 2125 2165 2260

$Q_1 = 2030$ $Q_2 = 2055$ (Median) $Q_3 = 2100$

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

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.

COMPLETE SOLUTIONS

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.

Applications

5. A sample of 30 Irish engineering graduates had the following starting salaries. Data are in thousands of euros.

36.8	34.9	35.2	37.2	36.2	35.8	36.8	36.1	36.7	36.6
37.3	38.2	36.3	36.4	39.0	38.3	36.0	35.0	36.7	37.9
38.3	36.4	36.5	38.4	39.4	38.8	35.4	36.4	37.0	36.4

a. What is the mean starting salary?
b. What is the median starting salary?
c. What is the mode?
d. What is the first quartile?
e. What is the third quartile?

ENGSAL

6. The following data were obtained for the number of minutes spent listening to recorded music for a sample of 30 individuals on one particular day.

88.3	4.3	4.6	7.0	9.2	0.0	99.2	34.9	81.7	0.0
85.4	0.0	17.5	45.0	53.3	29.1	28.8	0.0	98.9	64.5
4.4	67.9	94.2	7.6	56.6	52.9	145.6	70.4	65.1	63.6

COMPLETE SOLUTIONS

a. Compute the mean.
b. Compute the median.
c. Compute the first and third quartiles.
d. Compute and interpret the 40th percentile.

MUSIC

7. miniRank (www.minirank.com) rates the popularity of websites in most countries of the world, using a points system. The 25 most popular sites in South Africa as listed in July 2012 were as follows (the points scores have been rounded to one decimal place):

Website	*Points*	*Website*	*Points*
http://www.intoweb.co.za	253.1	www.dweb.co.za	118.2
http://www.weathersa.co.za	252.3	dweb.co.za	108.5
www.etraffic.co.za	212.4	www.webworx.org.za	107.6
www.gov.za	167.0	www.bacchus.co.za	105.2
www.intowebtraining.co.za	164.6	www.services.gov.za	103.3
www.capewebdesign.co.za	161.7	www.info.gov.za	102.2
www.saweather.co.za,	153.3	www.sars.co.za	95.6
www.web-inn.co.za	136.8	www.sars.gov.za	93.8
www.searchengine.co.za	136.1	www.mwebbusiness.co.za	93.6
www.saweather.co.za	133.6	www.dti.gov.za,	84.0
www.iol.co.za	132.5	www.jdconsulting.co.za	82.2
www.tradepage.co.za	128.6	www.linx.co.za	81.0
www.proudlysa.co.za	126.9		

RSA WWW

a. Compute the mean and median.
b. Do you think it would be better to use the mean or the median as the measure of central location for these data? Explain.
c. Compute the first and third quartiles.
d. Compute and interpret the 85th percentile.

8. Following is a sample of age data for individuals working from home by 'telecommuting'.

18	54	20	46	25	48	53	27	26	37
40	36	42	25	27	33	28	40	45	25

a. Compute the mean and the mode.
b. Suppose the median age of the population of all adults is 35.5 years. Use the median age of the preceding data to comment on whether the at-home workers tend to be younger or older than the population of all adults.
c. Compute the first and third quartiles.
d. Compute and interpret the 32nd percentile.

3.2 MEASURES OF VARIABILITY

In addition to measures of location, it is often desirable to consider measures of variability, or dispersion. For example, suppose you are a purchaser for a large manufacturing firm and 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 ten 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 ten 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 seven- or eight-day deliveries shown for J.C. Clark Distributors might be viewed favourably. However, a few of the slow 13- to 15-day deliveries could be disastrous in terms of keeping a workforce busy 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 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**.

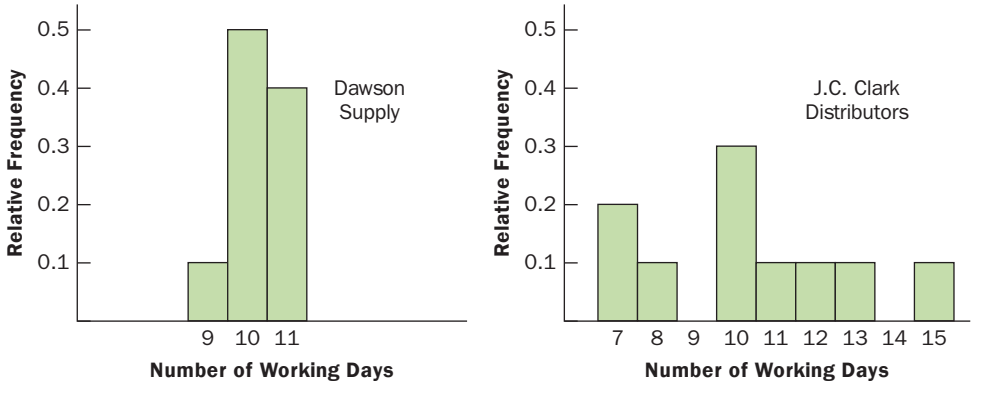


FIGURE 3.2
Historical data showing the number of days required to fill orders

Range

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

Refer to the data on starting salaries for business school graduates in Table 3.1. The largest starting salary is 2260 and the smallest is 1955. The range is $2260 - 1955 = 305$.

Although the range is the easiest of the measures of variability to compute, it is seldom used as the only measure. The range is based on only two of the observations and so is highly influenced by extreme values. Suppose one of the graduates received a starting salary of €5000 per month. In this case, the range would be $5000 - 1955 = 3045$ rather than 305. This would not be especially descriptive of the variability in the data because 11 of the 12 starting salaries are relatively closely grouped between 1955 and 2165.

Interquartile range

A measure of variability that overcomes the dependency on extreme values is the **interquartile range (IQR)**. This measure of variability is simply 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 per cent of the data.

Interquartile range

$$IQR = Q_3 - Q_1 \qquad \textbf{(3.3)}$$

For the data on monthly starting salaries, the quartiles are $Q_3 = 2100$ and $Q_1 = 2030$. The interquartile range is $2100 - 2030 = 70$.

Variance

The **variance** is a measure of variability that uses all the data. The variance is based on the difference between the value of each data value and the mean. This difference 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 (sigma squared). For a population of N observations and with μ denoting the population mean, the definition of the population variance is as follows:

Population variance

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

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 $n - 1$, not by n, the resulting sample variance provides an unbiased estimate of the population variance (a formal definition of unbiasedness is given in Chapter 7).

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
Totals		**0**	**256**
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

For this reason, the *sample variance*, denoted by s^2, is defined as follows:

Sample variance

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

Consider the data on class size for the sample of five university classes (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:

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

The units associated with the sample variance can cause confusion. Because the values 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 sample variance for the class size data is $s^2 = 64$ (students)^2. The squared units make it difficult to obtain an intuitive understanding and interpretation 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 comparable variables. The one with the larger variance will show the greater variability.

As another illustration, consider the starting salaries in Table 3.1 for the 12 business school graduates. In Section 3.1, we showed that the sample mean starting salary was 2070. The computation of the sample variance ($s^2 = 6754.5$) is shown in Table 3.3.

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

An alternative formula for the computation of the sample variance:

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

where:

$$\Sigma x_i^2 = x_1^2 + x_2^2 + \ldots + x_n^2$$

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.

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 $(x_i - \bar{x})^2$
2020	207	−50	2500
2075	207	5	25
2125	207	55	3025
2040	207	−30	900
1980	207	−90	8100
1955	207	−115	13225
2050	207	−20	400
2165	207	95	9025
2070	207	0	0
2260	207	190	36100
2060	207	−10	100
2040	207	−30	900
Totals		**0**	**74300**

Using equation (3.5)

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{74300}{11} = 6754.5$$

The standard deviation is derived from the variance as shown in equations (3.6) and (3.7).

Standard deviation

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} \quad \textbf{(3.6)}$$

$$\text{Sample standard deviation} = s = \sqrt{s^2} \quad \textbf{(3.7)}$$

Recall that the sample variance for the sample of class sizes in five university classes is $s^2 = 64$. Hence the sample standard deviation is:

$$s = \sqrt{64} = 8$$

For the data on starting salaries, the sample standard deviation is:

$$s = \sqrt{6754.5} = 82.2$$

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 $s^2 = 6754.5$ (€)^2. Because the standard deviation is the square root of the variance, the units are euros for the standard deviation, s = €82.2. In other words, the standard deviation is measured in the same units as the original data. The standard deviation is therefore more easily compared to the mean and other statistics measured in the same units as the original data.

Coefficient of variation

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

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

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) × 100% = 18.2%. The coefficient of variation tells us that the sample standard deviation is 18.2 per cent of the value of the sample mean. For the starting salary data with a sample mean of 2070 and a sample standard deviation of 82.2, the coefficient of variation, (82.2/2070) × 100% = 4.0%, tells us the sample standard deviation is only 4.0 per cent 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.

EXERCISES

Methods

COMPLETE SOLUTIONS

9. Consider a sample with data values of 10, 20, 12, 17 and 16. Calculate the range and interquartile range.

10. Consider a sample with data values of 10, 20, 12, 17 and 16. Calculate the variance and standard deviation.

11. Consider a sample with data values of 27, 25, 20, 15, 30, 34, 28 and 25. Calculate the range, interquartile range, variance and standard deviation.

Applications

COMPLETE SOLUTIONS

12. The goals scored in six handball matches were 41, 34, 42, 45, 35 and 37. Using these data as a sample, compute the following descriptive statistics.

a. Range.
b. Variance.
c. Standard deviation.
d. Coefficient of variation.

13. Dinner bill amounts for set menus at a Dubai restaurant, Al Khayam, show the following frequency distribution. The amounts are in AED (Emirati Dirham). Compute the mean, variance and standard deviation.

Dinner bill (AED)	*Frequency*
30	2
40	6
50	4
60	4
70	2
80	2
Total	**20**

14. The following data were used to construct the histograms of the number of days required to fill orders for Dawson Supply and for 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.

CRIME

15. Police records show the following numbers of daily crime reports for a sample of days during the winter months and a sample of days during the summer months.

Winter:	18	20	15	16	21	20	12	16	19	20
Summer:	28	18	24	32	18	29	23	38	28	18

a. Compute the range and interquartile range for each period.
b. Compute the variance and standard deviation for each period.
c. Compute the coefficient of variation for each period.
d. Compare the variability of the two periods.

16. A production department uses a sampling procedure to test the quality of newly produced items. The department employs the following decision rule at an inspection station: if a sample of 14 items has a variance of more than 0.005, the production line must be shut down for repairs. Suppose the following data have just been collected:

3.43	3.45	3.43	3.48	3.52	3.50	3.39
3.48	3.41	3.38	3.49	3.45	3.51	3.50

Should the production line be shut down? Why or why not?

3.3 MEASURES OF DISTRIBUTIONAL SHAPE, RELATIVE LOCATION AND DETECTING OUTLIERS

We described several measures of location and variability for data distributions. It is also often important to have a measure of the shape of a distribution. In Chapter 2 we noted that a histogram offers an excellent graphical display showing the shape of a distribution. An important numerical measure of the shape of a distribution is **skewness**.

Distributional shape

Four histograms constructed from relative frequency distributions are shown in Figure 3.3. The histograms in Panels A and B are moderately skewed. The one in Panel A is skewed to the left: its skewness is –0.85 (negative skewness). The histogram in Panel B is skewed to the right: its skewness is +0.85 (positive skewness). The histogram in Panel C is symmetrical: 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.* However, the skewness can be easily computed using statistical software (see the software guides on the online platform).

*The formula for the skewness of sample data is:

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

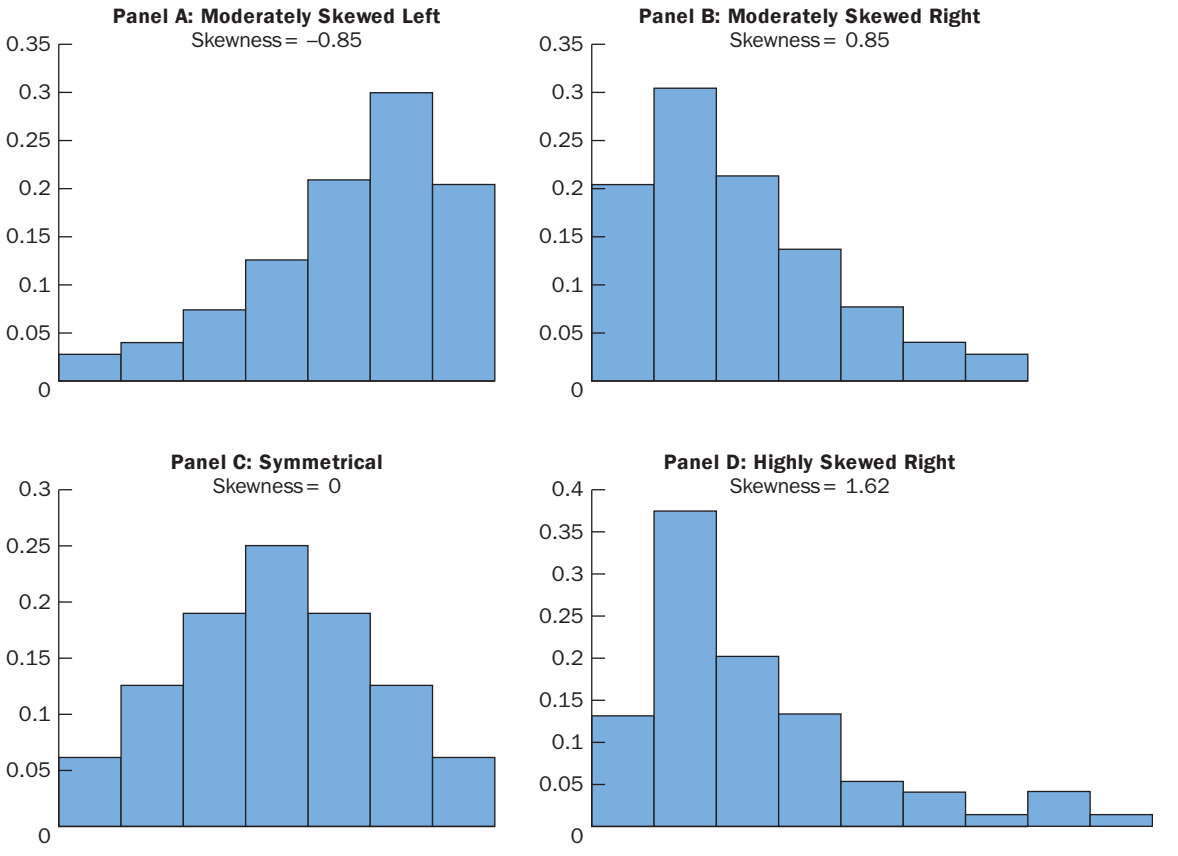


FIGURE 3.3
Histograms showing the skewness for four distributions

For a symmetrical 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 fashion store. The mean purchase amount is €77.60 and the median purchase amount is €59.70. The few large purchase amounts pull up the mean, but the median remains unaffected. The median provides a better measure of typical values when the data are highly skewed.

z-Scores

In addition to measures of location, variability and shape for a data set, we are often also interested in the relative location of data items within a data set. Such measures can help us determine whether a particular item is close to the centre of a data set or far out in one of the tails.

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

***z*-score**

$$z_i = \frac{x_i - \bar{x}}{s} \qquad \textbf{(3.9)}$$

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

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 $= \frac{x_i - \bar{x}}{s}$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	−2	−2/8 = −0.25
46	2	2/8 = 0.25
32	−12	−12/8 = −1.50

The z-score is often called the *standardized value* or the *standard score*. The z-score, z_i, represents 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 higher than the sample mean. Similarly, $z_2 = -0.5$ would indicate that x_2 is 0.5, or 1/2, standard deviation lower than the sample mean. Data values above the mean have a z-score greater than zero. Data values below the mean have a z-score less than zero. A z-score of zero indicates that the data value is equal to the sample mean.

The z-score is a measure of the relative location of the observation in a data set. Hence, 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 lie within a specified number of standard deviations of the mean.

Chebyshev's theorem

At least $(1 - 1/z^2) \times 100\%$ 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 per cent of the data values must be within $z = 2$ standard deviations of the mean.
- At least 89 per cent of the data values must be within $z = 3$ standard deviations of the mean.
- At least 94 per cent of the data values must be within $z = 4$ standard deviations of the mean.

Suppose that the mid-term test scores for 100 students in a university 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 per cent of the observations must have values within two standard deviations of the mean. Hence, at least 75 per cent of the students must have scored between 60 and 80.

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

$$\left(1-\frac{1}{z^2}\right) = \left(1-\frac{1}{(2.4)^2}\right) = 0.826$$

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

Empirical rule

Chebyshev's theorem applies to any data set, regardless of the shape of the distribution. It could be used, for example, with any of the skewed distributions in Figure 3.3. In many practical applications, however, data sets exhibit a symmetrical 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. The empirical rule is based on the normal probability distribution, which will be discussed in Chapter 6. The normal distribution is used extensively throughout this book.

Empirical rule

For data with a bell-shaped distribution:

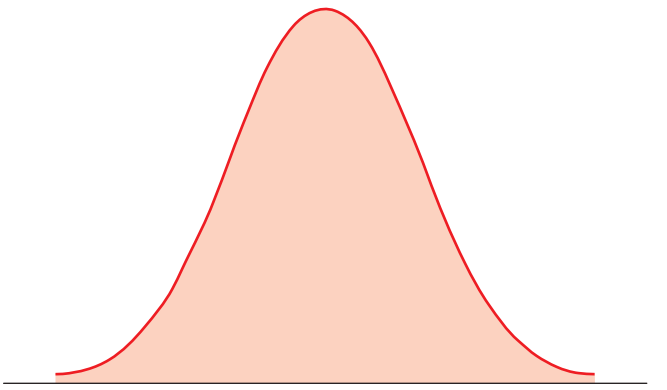
- Approximately 68 per cent of the data values will be within one standard deviation of the mean.
- Approximately 95 per cent of the data values will be within two standard deviations of the mean.
- Almost all of the data values will be within three standard deviations of the mean.

For example, the empirical rule allows us to say that *approximately* 95 per cent 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 per cent of the data values will be in that interval).

Consider liquid detergent cartons being filled automatically on a production line. Filling weights frequently have a bell-shaped distribution. If the mean filling weight is 500 grams and the standard deviation is 7 grams, we can use the empirical rule to draw the following conclusions:

- Approximately 68 per cent of the filled cartons will have weights between 493 and 507 grams (that is, within one standard deviation of the mean).
- Approximately 95 per cent of the filled cartons will have weights between 486 and 514 grams (that is, within two standard deviations of the mean).
- Almost all filled cartons will have weights between 479 and 521 grams (that is, within three standard deviations of the mean).

FIGURE 3.4
A symmetrical mound-shaped or bell-shaped distribution



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.

Standardized values (z-scores) can be used to identify outliers. 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, we recommend treating any data value with a z-score less than -3 or greater than $+3$ as an outlier, if the sample is small or moderately sized. 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 furthest from the mean. However, this standardized value is well within the -3 to $+3$ guideline for outliers. Hence, the z-scores give no indication that outliers are present in the class size data.

EXERCISES

Methods

COMPLETE SOLUTIONS

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

18. 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?

19. 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

20. 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

21. The results of a survey of 1154 adults showed that on average adults sleep 6.9 hours per day during the working week. 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 per day.
b. Use Chebyshev's theorem to calculate the percentage of individuals who sleep between 3.9 and 9.9 hours per day.
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)?

COMPLETE SOLUTIONS

22. Suppose that IQ scores have a bell-shaped distribution with a mean of 100 and a standard deviation of 15.

a. What percentage of people have an IQ score between 85 and 115?
b. What percentage of people have an IQ score between 70 and 130?
c. What percentage of people have an IQ score of more than 130?
d. A person with an IQ score greater than 145 is considered a genius. Does the empirical rule support this statement? Explain.

23. Suppose the average charge for a seven-day hire of an economy-class car in Kuwait City is KWD (Kuwaiti dinar) 75.00, and the standard deviation is KWD20.00.

a. What is the *z*-score for a seven-day hire charge of KWD56.00?
b. What is the *z*-score for a seven-day hire charge of KWD153.00?
c. Interpret the *z*-scores in parts (a) and (b). Comment on whether either should be considered an outlier.

24. *Consumer Review* posts reviews and ratings of a variety of products on the Internet. The following is a sample of 20 speaker systems and their ratings, on a scale of 1 to 5, with 5 being best.

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
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 an exploratory data analysis technique. 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 (minimum).
2. First quartile (Q_1).
3. Median (Q_2).

4 Third quartile (Q_3).

5 Largest value (maximum).

The easiest way to construct 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.

1955 1980 2020|2040 2040 2050|2060 2070 2075|2125 2165 2260

$Q_1 = 2030$ $Q_2 = 2055$ (Median) $Q_3 = 2100$

The median of 2055 and the quartiles Q_1 = 2030 and Q_3 = 2100 were computed in Section 3.1. The smallest value is 1955 and the largest value is 2260. Hence the five-number summary for the salary data is 1955, 2030, 2055, 2100, 2260. Approximately one-quarter, or 25 per cent, of the observations are between adjacent numbers in a five-number summary.

Box plot

A **box plot** is a slightly elaborated graphical version of the five-number summary. Figure 3.5 shows the construction of a box plot for the monthly starting salary data.

1 A box is drawn with the ends of the box located at the first and third quartiles. For the salary data, Q_1 = 2030 and Q_3 = 2100. This box contains the middle 50 per cent of the data.

2 A vertical line is drawn in the box at the location of the median (2055 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 = $Q_3 - Q_1$ = 2100 − 2030 = 70. Hence, the limits are 2030 − 1.5(70) = 1925 and 2100 + 1.5(70) = 2205. 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. Hence the whiskers end at salary values of 1955 and 2165.

5 Finally, the location of each outlier is shown with a symbol, often *. In Figure 3.5 we see one outlier, 2260. (Note that box plots do not necessarily identify the same outliers as identifying *z*-scores less than −3 or greater than +3.)

Figure 3.5 includes the upper and lower limits, to show how these limits are computed and where they are located for the salary data. Although the limits are always computed, they are not generally drawn on the box plots. The MINITAB box plots in Figure 3.6 illustrate the usual appearance, and also demonstrate that box plots are an excellent graphical tool for making comparisons amongst two or more groups.

FIGURE 3.5
Box plot of the starting salary data with lines showing the lower and upper limits

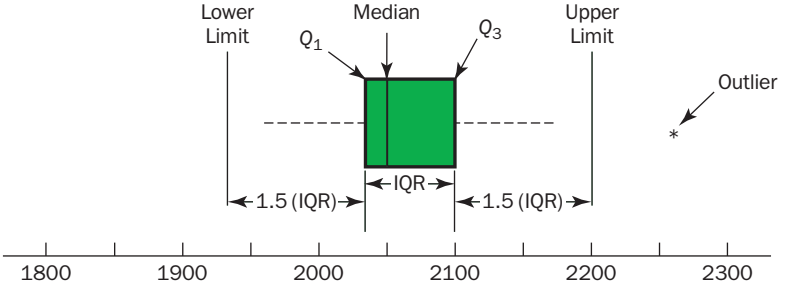


FIGURE 3.6
Box plot of monthly salary

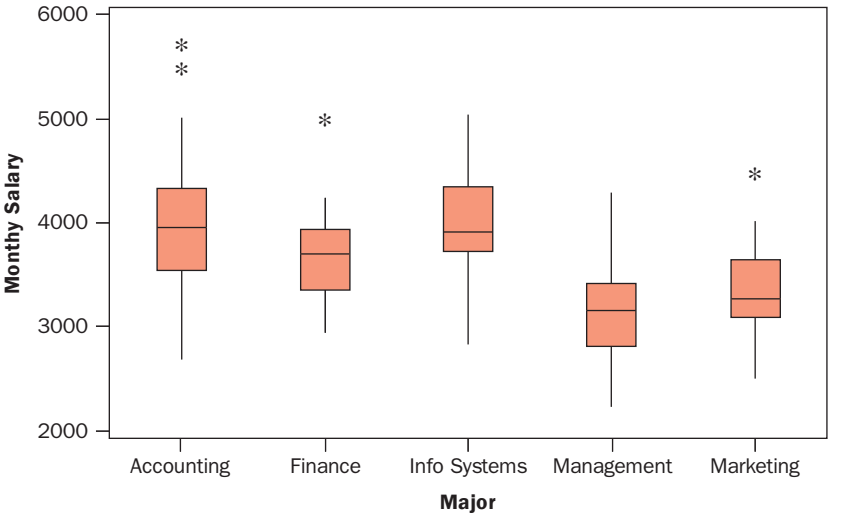


Figure 3.6 compares monthly starting salaries for a sample of 111 graduates, by major discipline. The major is shown on the horizontal axis and each box plot is arranged vertically above the relevant major label. The box plots in Figure 3.6 indicate that, for example:

- The highest median salary is in Accounting, the lowest in Management.
- Accounting salaries show the highest variation.
- There are high salary outliers for Accounting, Finance and Marketing.

EXERCISES

Methods

25. Consider a sample with data values of 27, 25, 20, 15, 30, 34, 28 and 25. Provide the five-number summary for the data.

26. Construct a box plot for the data in Exercise 25.

27. Prepare the five-number summary and the box plot for the following data: 5, 15, 18, 10, 8, 12, 16, 10, 6.

28. 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

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

8 408	1 374	1 872	8 879	2 459	11 413	608
14 138	6 452	1 850	2 818	1 356	10 498	7 478
4 019	4 341	739	2 127	3 653	5 794	8 305

a. Provide a five-number summary.
b. Compute the lower and upper limits (for the box plot).
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. Construct a box plot.

COMPLETE SOLUTIONS

30. A goal of management is to help their company earn as much as possible relative to the capital invested. One measure of success is return on equity – the ratio of net income to stockholders' equity. Return on equity percentages are shown here for 25 companies.

9.0	19.6	22.9	41.6	11.4	15.8	52.7	17.3	12.3	5.1
17.3	31.1	9.6	8.6	11.2	12.8	12.2	14.5	9.2	16.6
5.0	30.3	14.7	19.2	6.2					

EQUITY

a. Provide a five-number summary.
b. Compute the lower and upper limits (for the box plot).
c. Do the data contain any outliers? How would this information be helpful to a financial analyst?
d. Construct a box plot.

STOCK 2008

31. In 2008, stock markets around the world lost value. The website www.owneverystock.com listed the following percentage falls in stock market indices between the start of the year and the beginning of October.

Country	*% Fall*	*Country*	*% Fall*
New Zealand	27.05	Brazil	39.59
Canada	27.30	Japan	39.88
Switzerland	28.42	Sweden	40.35
Mexico	29.99	Egypt	41.57
Australia	31.95	Singapore	41.60
Korea	32.18	Italy	42.88
United Kingdom	32.37	Belgium	43.70
Spain	32.69	India	44.16
Malaysia	32.86	Hong Kong	44.52
Argentina	36.83	Netherlands	44.61
France	37.71	Norway	46.98
Israel	37.84	Indonesia	47.13
Germany	37.85	Austria	50.06
Taiwan	38.79	China	60.24

a. What are the mean and median percentage changes for these countries?
b. What are the first and third quartiles?
c. Do the data contain any outliers? Construct a box plot.
d. What percentile would you report for Belgium?

3.5 MEASURES OF ASSOCIATION BETWEEN TWO VARIABLES

We have examined numerical methods used to summarize *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.

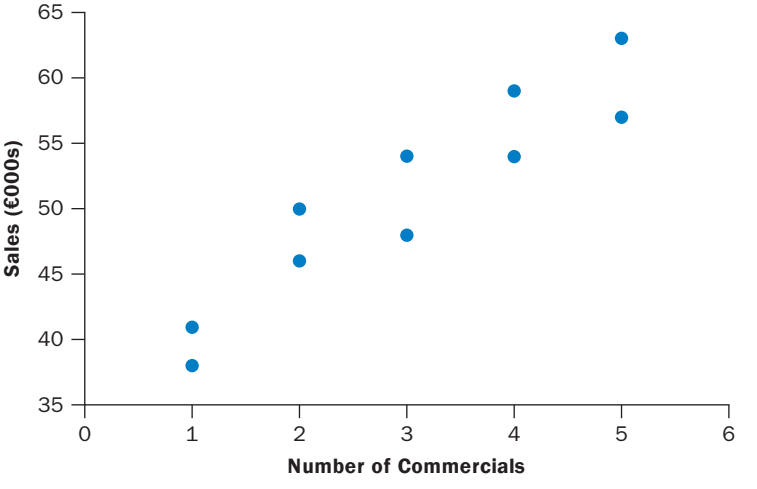
We begin by reconsidering the hi-fi equipment store discussed in Section 2.3. The store's manager wants to determine the relationship between the number of weekend television commercials shown and the sales at the store during the following week. Sample data with sales expressed in €000s were given in Table 2.12, and are repeated here in the first three columns of Table 3.5. It shows ten observations ($n = 10$), one for each week.

The scatter diagram in Figure 3.7 shows a positive relationship, with higher sales (vertical axis) associated with a greater number of commercials (horizontal axis). 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.

TABLE 3.5 Calculations for the sample covariance

Week	Number of commercials x_i	Sales volume (€000s) y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2	50	−1	−1	1
2	5	57	2	6	12
3	1	41	−2	−10	20
4	3	54	0	3	0
5	4	54	1	3	3
6	1	38	−2	−13	26
7	5	63	2	12	24
8	3	48	0	−3	0
9	4	59	1	8	8
10	2	46	−1	−5	5
Totals	**30**	**510**	**0**	**0**	**99**

FIGURE 3.7
Scatter diagram for the hi-fi equipment store



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

$$s_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} \quad \textbf{(3.10)}$$

This formula pairs each x_i with a corresponding 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 $n - 1$.

To measure the strength of the linear relationship between the number of commercials X and the sales volume Y in the hi-fi equipment store problem, we use equation (3.10) to compute the sample covariance. The calculations in Table 3.5 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:

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

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

$$\sigma_{XY} = \frac{\Sigma(x_i - \mu_X)(y_i - \mu_Y)}{N} \quad \textbf{(3.11)}$$

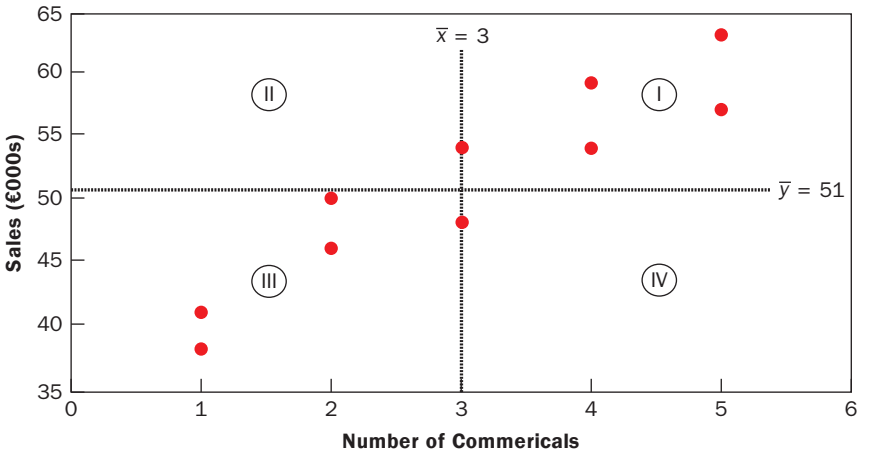
In equation (3.11) we use the notation μ_X for the population mean of X and μ_Y for the population mean of Y. The population covariance σ_{XY} is defined for a population of size N.

Interpretation of the covariance

To aid in the interpretation of the sample covariance, consider Figure 3.8. 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. Hence, the value of $(x_i - \bar{x})(y_i - \bar{y})$ is positive for points in quadrants I and III, negative for points in quadrants II and IV .

If the value of s_{XY} is positive, the points with the greatest influence on s_{XY} are 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 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 s_{XY} will be close to zero, indicating no linear association between X and Y. Figure 3.9 shows the values of s_{XY} that can be expected with three different types of scatter diagrams.

FIGURE 3.8
Partitioned scatter diagram for the hi-fi equipment store



Referring again to Figure 3.8, we see that the scatter diagram for the hi-fi equipment store follows the pattern in the top panel of Figure 3.9. As we 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 metres or centimetres (or feet). Measuring the height in centimetres, however, gives us much larger numerical values for $(x_i - \bar{x})$ than when we measure height in metres. Hence, with height measured in centimetres, 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. The **correlation coefficient** is measure of the relationship between two variables that is not affected by the units of measurement for X and Y.

Correlation coefficient

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

Pearson product moment correlation coefficient: sample data

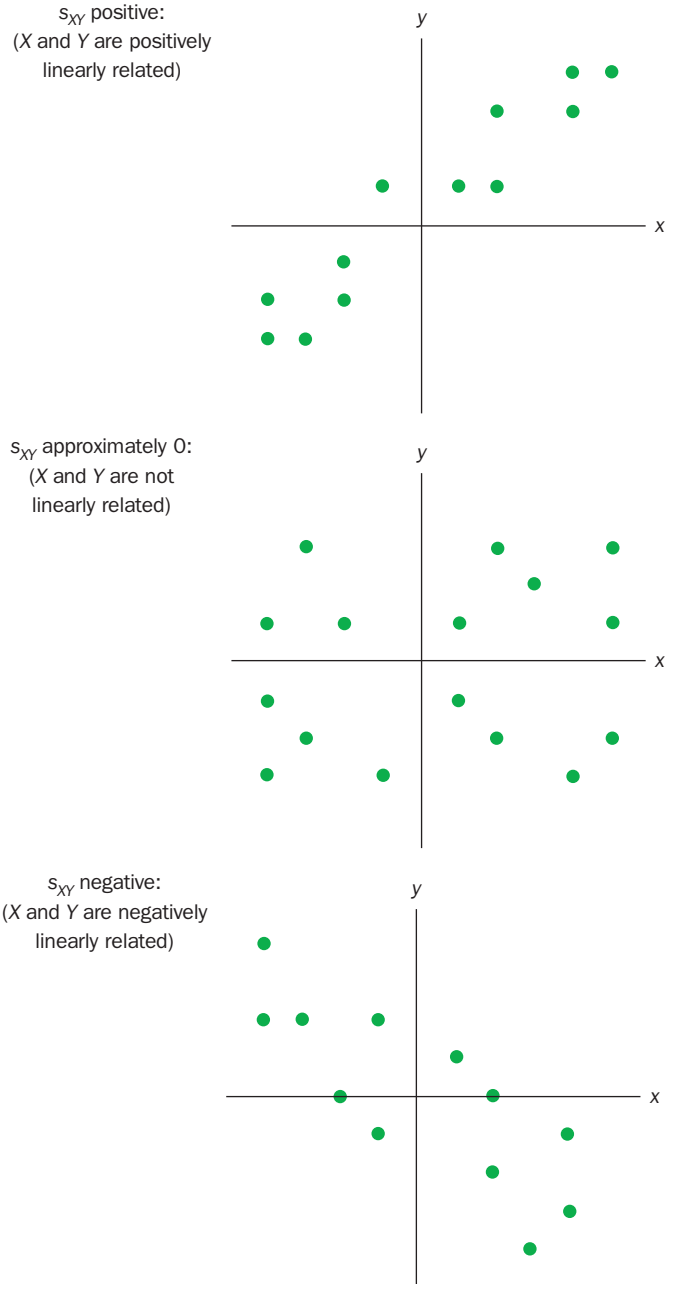
$$r_{XY} = \frac{s_{XY}}{s_X s_Y} \qquad \textbf{(3.12)}$$

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.

FIGURE 3.9
Interpretation of sample covariance



Let us now compute the sample correlation coefficient for the hi-fi equipment store. Using the data in Table 3.5, 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:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{11}{(1.49)(7.93)} = +0.93$$

The formula for computing the correlation coefficient for a population, denoted by the Greek letter ρ_{XY} (ρ is rho, pronounced 'row', to rhyme with 'go'), 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} 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 3.10 depicts the relationship between X and Y based on the following sample data.

x_i	y_i
5	10
10	30
15	50

The straight line drawn through 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 3.6.

FIGURE 3.10
Scatter diagram depicting a perfect positive linear relationship

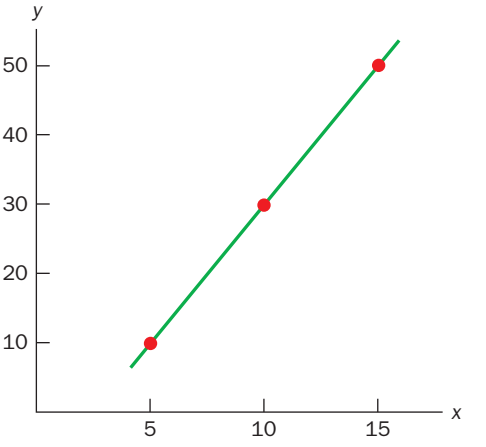


TABLE 3.6 Computations used in calculating the sample correlation coefficient

	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**
	$\bar{x} = 10$	$\bar{y} = 10$					

Using the results in Table 3.6, we find:

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

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

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

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{100}{5 \times 20} = +1$$

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. If the points in the data set fall on a straight line with a 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.

Suppose that a 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 closer and closer to zero. 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 hi-fi equipment store, recall that $r_{XY} = +0.93$. 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 one variable causes the other. For instance, we may find that a restaurant's quality rating and its typical meal price are positively correlated. However, increasing the meal price will not cause quality to increase.

EXERCISES

COMPLETE SOLUTIONS

Methods

32. Five observations taken for two variables follow.

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

a. Construct a scatter diagram with the x_i values 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.

33. Five observations taken for two variables follow.

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

a. Construct 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

34. Below are return on investment figures (%) and current ratios (current assets/current liabilities) for 15 German companies, for the year 2011 (file G_Comp on the online platform).


G_COMP

Company	*Return on investment (%)*	*Current ratio*
Adidas	8.15	1.50
BASF	14.66	1.64
Bayer	6.37	1.50
BMW	5.98	1.04
Continental	7.15	1.06
Daimler	5.70	1.11
Deutsche Bank	0.25	0.82
Deutsche Telekom	2.46	0.65
Fresenius	9.10	1.34
Henkel	9.16	1.58
Linde	5.60	0.89
SAP	20.53	1.54
Siemens	8.87	1.21
Tui	1.53	0.65
Volkswagen	7.46	1.05

a. Construct a scatter diagram with current ratio on the horizontal axis.
b. Is there any relationship between return on investment and current ratio? Explain.
c. Compute and interpret the sample covariance.
d. Compute and interpret the sample correlation coefficient.
e. What does the sample correlation coefficient tell you about the relationship between return on investment and current ratio?

35. Stock markets across the Eurozone tend to have mutual influences on each other. The index levels of the German DAX index and the French CAC 40 index for ten weeks beginning with 4 June 2012 are shown below (file 'DAX_CAC' on the online platform).


DAX_CAC

Date	*DAX*	*CAC 40*
04-Jun-12	6130.82	3051.69
11-Jun-12	6229.41	3087.62
18-Jun-12	6263.25	3090.90

Date	DAX	CAC 40
25-Jun-12	6416.28	3196.65
02-Jul-12	6410.11	3168.79
09-Jul-12	6557.10	3180.81
16-Jul-12	6630.02	3193.89
23-Jul-12	6689.40	3280.19
30-Jul-12	6865.66	3374.19
06-Aug-12	6967.95	3453.28

a. Compute the sample correlation coefficient for these data.
b. Are they poorly correlated, or do they have a close association?

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 re-stated as follows.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad \textbf{(3.14)}$$

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**. The weighted mean is computed as follows:

Weighted mean

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \qquad \textbf{(3.15)}$$

where:

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

For sample data, equation (3.15) provides the weighted sample mean. For population data, 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. Note that the cost per kilogram has varied from €2.80 to €3.40 and the quantity purchased has varied from 500 to 2750 kilograms.

Purchase	Cost per kilogram (€)	Number of kilograms
1	3.00	1200
2	3.40	500
3	2.80	2750
4	2.90	1000
5	3.25	800

Suppose a manager asked for information about the mean cost per kilogram of the raw material. Because the quantities ordered vary, we must use the formula for a weighted mean. The five cost-per-kilogram values are $x_1 = 3.00$, $x_2 = 3.40$, … etc. The weighted mean cost per kilogram is found by weighting each cost by its corresponding quantity. The weights are $w_1 = 1200$, $w_2 = 500$, … etc. Using equation (3.15), the weighted mean is calculated as follows:

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

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

The weighted mean computation shows that the mean cost per kilogram 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-kilogram 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 kilogram purchased.

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, in the context of the particular application.

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. We show how the weighted mean formula can be used to obtain approximations of the mean, variance and standard deviation for **grouped data**.

Recall from Section 2.2 the frequency distribution of times in days required to complete year-end audits for the small accounting firm of Sanderson and Clifford. It is shown again in the first two columns of Table 3.7 ($n = 20$ clients). 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 sample size n. That is, $\Sigma f_i = n$.

Hence, the equation for the sample mean for grouped data is as follows in equation (3.16).

Sample mean for grouped data

$$\bar{x} = \frac{\Sigma f_i M_i}{n} \qquad \textbf{(3.16)}$$

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 3.7 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 3.7. 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 given 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.

TABLE 3.7 Computation of the sample mean audit time for grouped data

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

$$\text{Sample mean } \bar{x} = \frac{\sum f_i M_i}{n} = \frac{180}{20} = 19 \text{ days}$$

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 $\sum f_i (M_i - \bar{x})^2$.

The term $n - 1$ rather than n appears in the denominator in order to make the sample variance an unbiased estimator of the population variance. The following formula is used to obtain the sample variance for grouped data.

Sample variance for grouped data

$$s^2 = \frac{\sum f_i (M_i - \bar{x})^2}{n - 1} \qquad \textbf{(3.17)}$$

The calculation of the sample variance for audit times based on the grouped data from Table 3.7 is shown in Table 3.8. 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 $s = \sqrt{30} = 5.48$.

Note that formulae (3.16) and (3.17) are for a sample. Population summary measures are computed similarly in equations (3.18) and (3.19).

TABLE 3.8 Computation of the sample variance of audit times for grouped data

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
Total		**20**			**570**

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

$\sum f_i (M_i - \bar{x})^2$

Population mean for grouped data

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

Population variance for grouped data

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

EXERCISES

Methods

COMPLETE SOLUTIONS

36. Consider the following data and corresponding weights.

x_i	*Weight*
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.

37. 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

38. The assessment for a statistics module comprises a multiple-choice test, a data analysis project, an EXCEL test and a written examination. Scores for Jil and Ricardo on the four components are show below.

Assessment	*Jil*	*Ricardo*
Multiple-choice test	80%	48%
Data analysis project	60%	78%
EXCEL test	62%	60%
Written examination	57%	53%

a. Calculate weighted mean scores (%) for Jil and Ricardo assuming the respective weightings for the four components are 20, 20, 30, 30.
b. Calculate weighted mean scores (%) for Jil and Ricardo assuming the respective weightings for the four components are 10, 25, 15, 50.

39. A petrol station recorded the following frequency distribution for the number of litres of petrol sold per car in a sample of 680 cars.

Petrol (litres)	*Frequency*
1–15	74
16–30	192
31–45	280
46–60	105
61–75	23
76–90	6
Total	**680**

Compute the mean, variance and standard deviation for these grouped data. If the petrol station expects to serve petrol to about 120 cars on a given day, estimate the total number of litres of petrol that will be sold.

ONLINE RESOURCES

For the data files, online summary, additional questions and answers, and software section for Chapter 3, go to the online platform.

SUMMARY

In this chapter we introduced several descriptive statistics that can be used to summarize the location, variability and shape of a data distribution. 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. In statistical inference, the sample statistic is referred to as the point estimator of the population parameter. Some of the notation used for sample statistics and population parameters follow.

	Sample statistic	*Population parameter*
Mean	$\bar{x}$	μ
Variance	s^2	σ^2
Standard deviation	s	σ
Covariance	s_{XY}	ρ_{XY}
Correlation	r_{XY}	σ_{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 showed how to calculate *z*-scores, and indicated how they can be used to identify outlying observations. 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 construct a five-number summary and a box plot to provide simultaneous information about the location, variability and shape of the distribution.

Section 3.5 introduced covariance and the correlation coefficient as measures of association between two variables.

In Section 3.6, we showed how to compute a weighted mean and how to calculate a mean, variance and standard deviation for grouped data.

KEY TERMS

Box plot
Chebyshev's theorem
Coefficient of variation
Correlation coefficient
Covariance
Empirical rule
Five-number summary
Grouped data
Interquartile range (IQR)
Mean
Median
Mode
Outlier
Percentile
Point estimator
Population parameter
Quartiles
Range
Sample statistic
Skewness
Standard deviation
Variance
Weighted mean
***z*-score**

KEY FORMULAE

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

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

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

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} \tag{3.6}$$

$$\text{Sample standard deviation} = s = \sqrt{s^2} \tag{3.7}$$

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

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

CASE PROBLEM 1

COMPANIES 2012

Company Profiles

The file 'Companies 2012' on the online platform contains a data set compiled mid-year 2012. It comprises figures relating to samples of companies whose shares are traded on the stock exchanges in Germany, France, South Africa and Israel. The data contained in the file are:

- Name of company.
- Country of stock exchange where the shares are traded.
- Return on shareholders' funds in 2011 (%).
- Profit margin in 2011 (%).
- Return on total assets in 2011 (%).
- Current ratio, 2011.
- Solvency ratio, 2011.
- Price/earnings ratio, 2011.

The first few rows of data are shown below.

Company name	*Country*	*Return on share holders' funds, 2011 (%)*	*Profit margin 2011 (%)*	*Return on total assets 2011 (%)*	*Curent ratio, 2011*	*Solvency ratio, 2011*	*Price/ earnings ratio, 2011*
Adidas AG	Germany	17.40	6.85	8.15	1.50	46.81	15.72
Allianz SE	Germany	10.79	6.99	0.77		7.15	11.92
Altana AG	Germany	3.32	3.28	2.28	2.40	68.77	200.13
BASF SE	Germany	37.16	11.90	14.66	1.64	39.46	7.96
Bayer AG	Germany	17.50	9.04	6.37	1.50	36.41	16.47
BWW AG	Germany	27.31	10.69	5.98	1.04	21.91	6.52
Commerz bank	Germany	2.04	4.09	0.08	0.41	3.75	8.92
Continental AG	Germany	26.05	6.06	7.15	1.06	27.44	7.71
Daimler AG	Germany	21.32	7.84	5.70	1.11	26.75	6.35
Deutsche Bank AG	Germany	9.86	16.16	0.25	0.82	2.53	6.23

Managerial report

1. Produce summaries for each of the numerical vari-ables in the file using suitable descriptive statistics. For each variable, identify outliers as well as summarizing the overall characteristics of the data distribution.
2. Investigate whether there are any differences between countries in average profit margin. Similarly, investigate whether there are differences between countries in average current ratio and in average price/earnings ratio.
3. Investigate whether there is any relationship between return on investment and current ratio. Similarly, investigate whether there is any relationship between return on investment and price/earnings ratio.

The Johannesburg Stock Exchange

CASE PROBLEM 2

Chocolate Perfection Website Transactions

Chocolate Perfection manufactures and sells quality chocolate products in Dubai. 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 management is now considering strategies to increase sales even further. To learn more about the website customers, a sample of 50 Chocolate Perfection transactions was selected from the previous month's sales. Data showing the day 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'. Amount spent is in United Arab Emirates dirham (AED). (One Euro is around five AED.) A portion of the data is shown below.

Customer	*Day*	*Browser*	*Time (min)*	*Pages Viewed*	*Amount Spent (AED)*
1	Mon	Internet Explorer	12.0	4	200.09
2	Wed	Other	19.5	6	348.28
3	Mon	Internet Explorer	8.5	4	97.92
4	Tue	Firefox	11.4	2	164.16
5	Wed	Internet Explorer	11.3	4	243.21
6	Sat	Firefox	10.5	6	248.83
7	Sun	Internet Explorer	11.4	2	132.27
8	Fri	Firefox	4.3	6	205.37
9	Wed	Firefox	12.7	3	260.35

SHOPPERS

Chocolate Perfection 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 has on sales.

Managerial report

Use the methods of descriptive statistics to learn about the customers who visit the Chocolate Perfection 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 Chocolate Perfection's online shoppers from these numerical summaries.
2. Summarize the frequency, the total amount spent and the mean amount spent per transaction for each day of week. What observations can you make about Chocolate Perfection's business based on the day of the week? Discuss.
3. Summarize the frequency, the total amount spent and the mean amount spent per transaction for each type of browser. What observations can you make about Chocolate Perfection's business, based on the type of browser? Discuss.
4. Construct a scatter diagram and compute the sample correlation coefficient to explore the relationship between the time spent on the website and the amount spent. Use the horizontal axis for the time spent on the website. Discuss.
5. Construct a scatter diagram and compute the sample correlation coefficient to explore the relationship between the number of website pages viewed and the amount spent. Use the horizontal axis for the number of website pages viewed. Discuss.
6. Construct 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.