Introduction to Statistics

22

After the invention of the steam engine in the late 1700s by the Scottish engineer James Watt, the production of machine-made goods became widespread during the 1800s. However, it was not until the 1920s that much attention was paid to the quality control of the goods being produced. In 1924, Walter Shewhart of Bell Telephone Laboratories used a statistical chart for controlling product variables; in the 1940s, quality control was used in much of wartime production.

Quality control is one of the modern uses of *statistics*, the branch of mathematics in which data are collected, displayed, analysed, and interpreted. Today it is nearly impossible to read a newspaper or watch television news without seeing some type of study, in areas such as medicine or politics, that involves statistics. Other areas in which statistical methods are used include biology, physics, psychology, sociology, reliability engineering, actuarial science, economics, business, and education, to name but a few.

The first significant use of statistics was made in the 1660s by John Graunt, and in the 1690s by Edmund Halley (of Halley's Comet), when each published some conclusions about the population in England based on mortality tables. There was little development of statistics until the 1800s, when statistical measures became more widely used. For example, important contributions were made by the scientist Francis Galton, who used statistics in the study of human heredity, and by the nurse Florence Nightingale, who used statistical graphs to show that more soldiers died in the Crimean War (in the 1850s) from unsanitary conditions than from combat wounds.

In using statistics, we generally collect and summarize data (using methods from *descriptive statistics*) to make inferences based on those data (using methods from *inferential statistics*). The first two sections of this chapter are dedicated to descriptive statistics. After a discussion regarding the normal distribution, we dedicate the rest of the chapter to introducing some basic concepts of inferential statistics, including confidence intervals, statistical process control, and regression.

David P. Lewis/Shutterstock

LEARNING OUTCOMES

After completion of this chapter, the student should be able to:

- Understand the basic concepts of population, sample, parameter, statistic, and variable
- Construct frequency, relative frequency, and cumulative frequency tables for data
- Draw a histogram, a frequency polygon, and an ogive
- Calculate measures of central tendency (mean, median, and mode) and measures of spread (range and standard deviation)
- Use Chebychev's theorem to draw conclusions about a data set
- Calculate relative frequencies using the normal distribution
- Construct large sample confidence intervals for a population mean or for a population proportion
- Plot an $\bar{x}$ control chart, an R control chart, and a p control chart
- Find the equation of the least-squares line that best fits a given set of data
- Find the equation of a curve that best fits a given set of data by transforming the independent variable and using linear least squares

◀ In Section 22.4 we see how statistical analysis was used in the design of the 12.9-km-long Confederation Bridge that joins New Brunswick and Prince Edward Island.

22.1 Tabular and Graphical Representation of Data

Population • Sample • Variables Array • Class • Frequency Distribution Table • Relative Frequency • Frequency Polygon • Histogram • Cumulative Frequency • Ogive

In statistics, a **population** is the complete collection of elements (people, DVDs, households, temperatures) that are of interest and about which information is desired. Typically, a researcher is interested in a numerical property of the population called a **parameter.** For example, if the population is all DVDs produced by a certain manufacturer over the course of a week, a parameter would be the proportion of defective DVDs in that lot. If the population is all Internet users, a parameter would be the average number of hours spent each week on the Internet.

Because of constraints on time, money, and other scarce resources, conclusions about the population are usually drawn after observing only a subset of the population, called a **sample.** Quantities computed from samples are called **statistics.** Statistics are used to estimate parameters in the population. For example, the average in a sample of Internet users can be used to estimate the average among all users. Similarly, the proportion of defectives in a sample can be used to estimate the proportion of defectives in the complete lot.

We are usually only interested in some of the characteristics that elements of the population have in common. A **variable** is any characteristic whose value changes from individual to individual in the population. A **quantitative variable** has a value that represents a numerical measurement. Examples of quantitative variables are weight, length, voltage, pressure, and number of children in a family. When the value of a variable is non-numerical, it is called a qualitative variable, or an **attribute.** Examples of attributes are colour, gender, and quality (measured as defective or nondefective).

Values of variables that have been recorded constitute data. *Data that have been collected but not yet organized are called* **raw data.** In order to obtain useful information from the data, it is necessary to organize it in some way. Normally, a first step in organizing the data is to arrange the numerical values in ascending (or descending) order, forming what is called an **array.**

EXAMPLE 1 Illustrating an array

Each user in a sample of 50 home computer users was asked to estimate carefully the number of hours they spent each week on-line on the Internet. Following are the estimates.

12, 20, 15, 14, 7, 10, 12, 25, 18, 5, 10, 24, 16, 3, 12, 14, 28, 8, 13, 18,
15, 8, 11, 15, 14, 22, 14, 19, 6, 10, 18, 4, 16, 24, 18, 5, 13, 20, 12, 12,
25, 11, 8, 12, 20, 5, 10, 15, 13, 8

As we can see, no clear pattern can be seen from the raw data. Arranging these in numerical order to form an array, we can summarize the array by showing the number of persons reporting each estimate as follows:

(hours-persons) 3-1, 4-1, 5-3, 6-1, 7-1, 8-4, 10-4, 11-2, 12-6,
13-3, 14-4, 15-4, 16-2, 18-4, 19-1, 20-3, 22-1, 24-2, 25-2, 28-1 ■

In Example 1, although a pattern is somewhat clearer from the summarized array than from the raw data, a still clearer pattern is found by *grouping the data.* In the process of grouping, the detail of the raw data is lost, but the advantage is that a much clearer overall pattern of the data can be obtained.

The grouping of data is done by first defining what values are to be included in each group and then tabulating the number of values that are in each group. *Each group is called a* **class,** and *the number of values in each class is called the* **frequency.** *The table is called a* **frequency distribution table.** This is illustrated in the following example.

EXAMPLE 2 Frequency distribution table

In Example 1, we note that the estimates vary from 3 h to 28 h. We see that if we form *classes* of 0–4 h, 5–9 h, etc., we will have five possible estimates in each class and that there will be six classes. This gives us the following frequency distribution table of values showing the number of persons (frequency) reporting the indicated estimate of the hours spent on the Internet.

Estimate (hours)	0–4	5–9	10–14	15–19	20–24	25–29
Frequency (persons)	2	9	19	11	6	3

This table shows us the frequency distribution. The 0, 5, 10, and so on are the *lower class limits,* and the 4, 9, 14, and so on are the *upper class limits.* Each class includes five values, which is the *class width.* As with the class limits we have chosen, it is generally preferable to have the same width for each class.

We can see from this frequency distribution table that the pattern of hours on the Internet by the persons responding to the survey is clearer. ■

Guidelines for Constructing Frequency Tables

- Make sure that classes are mutually exclusive, so that each observation belongs to one, and only one, class.
- Use between 5 and 20 classes. The principal consideration is that the relevant characteristics of the data should be clear.
- Ensure that all classes (except for open-ended classes) have the same width.
- Use class limits with convenient numbers. The first lower class limit is selected as the lowest value, or as a convenient number less than the lowest value.
- Be sure to include all classes, even if their frequency is zero.

At times, it is also helpful to know the **relative frequency** *of the class, which is the frequency of the class divided by the total frequency of all classes.* The relative frequency can be expressed as a fraction, decimal, or percent.

EXAMPLE 3 Relative frequency

The relative frequency of each class for the data in Example 2 can be shown as in the following table:

Estimated Hours on Internet	*Frequency*	*Relative Frequency* (%)
0–4	2	4
5–9	9	18
10–14	19	38
15–19	11	22
20–24	6	12
25–29	3	6
Total	50	100

$2/50 = 0.04 = 4\%$

■

Practice Exercise

1. Assuming the data in Example 1 are divided into classes of 0–3 h, 4–7 h, etc., for the 8–11 h class find:
(a) the frequency
(b) the relative frequency

Using graphs is a very convenient method of representing frequency distributions. There are several useful types of graphs for such distributions. *Among the most important of these are the* **histogram** *and the* **frequency polygon.**

In order to represent *grouped* data, where the raw data values are generally not all the same within a given class, we find it necessary to use a representative value for each class. For this, we use the **class mark,** *which is found by dividing the sum of the lower and upper class limits by* 2. The following examples illustrate the use of a class mark with a histogram and a frequency polygon.

■ A graphing calculator may be used to display histograms and frequency polygons.

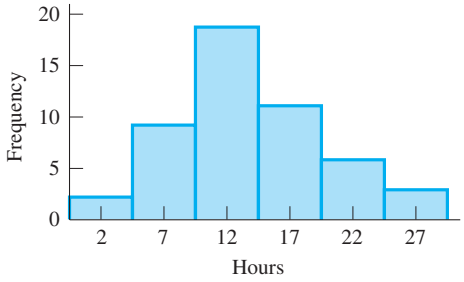

Fig. 22.1

■ Computer spreadsheets are very useful for this type of analysis.

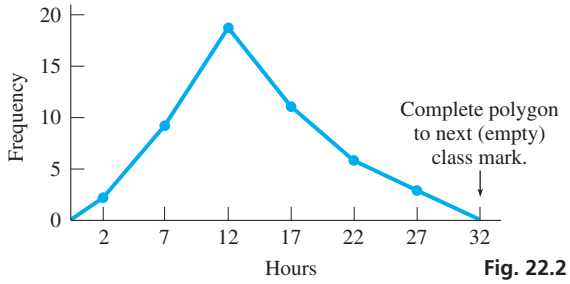

Fig. 22.2

EXAMPLE 4 Histogram

A histogram represents a particular set of data by displaying each class of the data as a rectangle. Each rectangle is labelled at the centre of its base by the *class mark.* The width of each rectangle represents the *class width,* and the height of the rectangle represents the *frequency* of the class.

For the data in Example 2 on estimated hours on the Internet, the class marks are $(0 + 4)/2 = 2$, $(5 + 9)/2 = 7$, and so on. Using these values, a histogram representing these data is shown in Fig. 22.1. ■

EXAMPLE 5 Frequency polygon

A frequency polygon is used to represent a set of data by plotting the class marks as abscissas (x-values) and the frequencies as ordinates (y-values). The resulting points are joined by straight-line segments.

A frequency polygon representing the data in Example 2 on estimated hours on the Internet is shown in Fig. 22.2.

If the polygon is not completed as shown in Fig. 22.2, and the figure starts at the first class mark, and ends at the last class mark, it is then referred to as a *broken-line graph*. ■

Another way of analysing data is to use *cumulative* totals. The way this is generally done is to change the frequency into a "less than" **cumulative frequency.** To do this, we *add the class frequencies,* starting at the lowest class boundary. The graphical display that is generally used for cumulative frequency is called an **ogive** (pronounced oh-jive).

EXAMPLE 6 Cumulative frequency—ogive

For the data on estimated hours on the Internet in Examples 2 and 3, the cumulative frequency is shown in the following table:

Estimated Hours on Internet	*Cumulative Frequency*
Less than 5	2
Less than 10	11
Less than 15	30
Less than 20	41
Less than 25	47
Less than 30	50

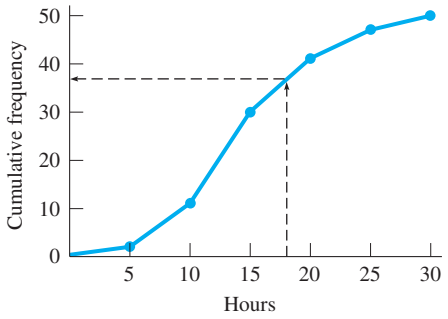

Fig. 22.3

The ogive showing the cumulative frequency for the values in this table is shown in Fig. 22.3. The vertical scale shows the *frequency,* and the horizontal scale shows the *class boundaries.*

One important use of an ogive is to determine the number of values above or below a certain value. For example, to *approximate* the number of respondents that use the Internet *less than* 18 hours per week, we draw a line from the horizontal axis to the ogive and then to the vertical axis as shown in Fig. 22.3. From this, we see that *about* 37 respondents use the Internet *less than* 18 hours per week. ■

If the data with which we are dealing have only a limited number of values and we do not divide the data into classes, we can still use the methods we have developed, using the specific values rather than *class* values. Consider the following example.

LEARNING TIP

Note that in a cumulative frequency table, the name of the class can vary. For example, the first class in Example 6 could be defined as less than 5, or up to but not including 5, or less than or equal to 4. What is important is that the classes be cumulative and that there be no overlap nor ambiguity between the classes.

EXAMPLE 7 Graphs using specific values

A test station measured the loudness of the sound of jet aircraft taking off from a certain airport. The decibel (dB) readings measured to the nearest 5 dB for the first 20 jets were as follows:

110, 95, 100, 115, 105, 110, 120, 110, 115, 105,
90, 95, 105, 110, 100, 115, 105, 120, 95, 110

Since there are only seven different values for the 20 readings, the best idea of the pattern is found by using these values, as shown in the following frequency table:

dB Reading	90	95	100	105	110	115	120
Frequency	1	3	2	4	5	3	2

The histogram for this table is shown in Fig. 22.4, and the frequency polygon is shown in Fig. 22.5. ■

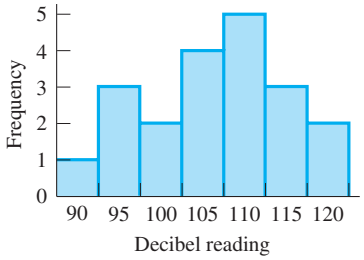

Fig. 22.4

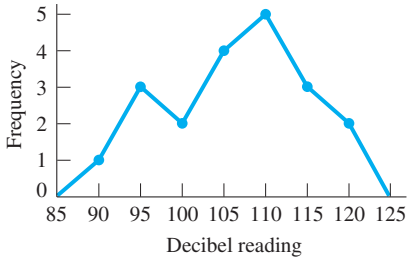

Fig. 22.5

EXERCISES 22.1

In Exercises 1–4, divide the data in Example 1 into five classes of hours (0–5, 6–11, etc.) on the Internet and then do the following:

1. Form a frequency distribution table.
2. Find the relative frequencies.
3. Draw a histogram.
4. Draw an ogive.

In Exercises 5–12, use the following data. An automobile company tested a new engine, and found the following results in twenty tests of the number of litres of gasoline used by a certain model for each 100 km travelled.

5.3, 5.8, 5.6, 5.4, 5.9, 5.4, 6.0, 5.8, 5.8, 5.4,
6.3, 5.6, 5.7, 5.6, 5.7, 5.9, 5.5, 6.1, 5.9, 5.8

5. Form an ordered array and summarize it by finding the frequency of each number of litres used.
6. Find the relative frequency of each number of litres used.
7. Form a frequency distribution table with five classes.
8. Find the relative frequencies for the data in the frequency distribution table in Exercise 7.
9. Draw a histogram for the data of Exercise 7.
10. Draw a frequency polygon for the data of Exercise 7.
11. Form a cumulative frequency table for the data of Exercise 7.
12. Draw an ogive for the data of Exercise 7.

In Exercises 13–32, find the indicated quantities.

13. In testing a computer system, the number of instructions it could perform in 1 ns was measured at different points in a program. The numbers of instructions were recorded as follows:

    19, 21, 22, 25, 22, 20, 18, 21, 20, 19, 22, 21, 19, 23, 21

    Form a frequency distribution table for these values.
14. For the data of Exercise 13, draw a histogram.
15. For the data of Exercise 13, draw a frequency polygon.
16. For the data of Exercise 13, form a relative frequency distribution table.
17. A strobe light is designed to flash every 2.25 s at a certain setting. Sample bulbs were tested with the following results:

Time (s) *between Flashes* (class mark)	2.21	2.22	2.23	2.24
Number of Bulbs	2	7	18	41

Time (s)	2.25	2.26	2.27	2.28	2.29
No. Bulbs	56	32	8	3	3

    Draw a histogram for these data.
18. For the data of Exercise 17, draw a frequency polygon.
19. For the data of Exercise 17, form a cumulative frequency distribution table.
20. For the data of Exercise 17, draw an ogive.
21. In testing a braking system, the distance required to stop a car from 110 km/h was measured in 120 trials. The results are shown in the following distribution table:

Stopping Distance (m)	47–49	50–52	53–55	56–58
Times Car Stopped	2	15	32	36

Stopping Distance (m)	59–61	62–64	65–67
Times Car Stopped	24	10	1

    Form a relative frequency distribution table for these data.
22. For the data in Exercise 21, form a cumulative frequency distribution table.
23. For the data of Exercise 21, draw an ogive.
24. From the ogive in Exercise 23, estimate the number of cars that stopped in less than 57 m.

25. The dosage, in millisieverts (mSv), given by a particular X-ray machine, was measured 20 times, with the following readings:

0.425, 0.436, 0.396, 0.421, 0.444, 0.383, 0.437, 0.427, 0.433, 0.434, 0.415, 0.390, 0.441, 0.451, 0.418, 0.426, 0.429, 0.409, 0.436, 0.423

Form a histogram with six classes and the lowest class mark of 0.380 mSv.

26. For the data used for the histogram in Exercise 25, draw a frequency polygon.

27. The life of a certain type of battery was measured for a sample of batteries with the following results (in number of hours):

34, 30, 32, 35, 31, 28, 29, 30, 32, 25, 31, 30,
28, 36, 33, 34, 30, 31, 34, 29, 30, 32

Draw a frequency polygon using six classes.

28. For the data in Exercise 27, draw a cumulative frequency distribution table using six classes.

29. The diameters of a sample of fibre-optic cables were measured with the following results (diameters are class marks):

Diam. (mm)	0.0055	0.0056	0.0057	0.0058	0.0059	0.0060
No. Cables	4	15	32	36	59	64

Diam. (mm)	0.0061	0.0062	0.0063	0.0064	0.0065	0.0066
No. Cables	22	18	10	12	4	4

Draw a histogram for these data.

30. For the data of Exercise 29, draw a histogram with six classes. Compare the pattern of distribution with that of the histogram in Exercise 29.

31. Toss four coins 50 times and tabulate the number of heads that appear for each toss. Draw a frequency polygon showing the number of tosses for which 0, 1, 2, 3, or 4 heads appeared. Describe the distribution. (Is it about what should be expected?)

32. Most calculators can generate random numbers (between 0 and 1). On a calculator, display 50 random numbers and record the first digit. Draw a histogram showing the number of times for which each first digit $(0, 1, 2, \ldots, 9)$ appeared. Describe the distribution. (Is it about what should be expected?)

Answers to Practice Exercise

1. (a) 10 (b) 20%

22.2 Summarizing Data

Median • Arithmetic Mean • Mode • Range • Standard Deviation • Chebychev's Theorem

LEARNING TIP

Note that all measures of central tendency and spread are usually rounded off to one more decimal place than was present in the original data.

Tables and graphical representations give a general description of data. However, it is also useful and convenient to find representative values for the location of the centre of the distribution, and other numbers to give a measure of the deviation from this central value. In this way, we can obtain a numerical summary of the data. We study some measures of centre and deviation (or spread) in this section.

MEASURES OF CENTRAL TENDENCY

The task of a *measure of central tendency* is to describe with a single value the location of the centre of the distribution. Since there are different ways of defining what centre is, there are several measures of central tendency.

Median

The first of these measures of central tendency is the **median.** The median is the value that falls in the middle of an ordered array of data, leaving as many observations above it as it does below it. If there is no middle observation, the median is the number halfway between the two numbers nearest to the middle of the array.

EXAMPLE 1 Median—odd or even number of values

Given the numbers 5, 2, 6, 4, 7, 4, 7, 2, 8, 9, 4, 11, 9, 1, 3, we first arrange them in numerical order. This arrangement is

middle number

1, 2, 2, 3, 4, 4, 4, 5, 6, 7, 7, 8, 9, 9, 11

Since there are 15 numbers, the middle number is the eighth. Since the eighth number is 5, the median is 5.

If the number 11 is not included in this set of numbers and there are only 14 numbers in all, the median is that number halfway between the seventh and eighth numbers. Since the seventh is 4 and the eighth is 5, the median is 4.5. ■

Arithmetic Mean

Another very widely applied measure of central tendency is the **arithmetic mean** (often referred to simply as the **mean**). *The mean is calculated by finding the sum of all*

the values and then dividing by the number of values. (The mean is the number most people call the "average." However, in statistics the word *average* has the more general meaning of a measure of central tendency.)

EXAMPLE 2 Arithmetic mean

The arithmetic mean of the numbers given in Example 1 is determined by finding the sum of all the numbers and dividing by 15. Therefore, by letting $\bar{x}$ (read as "x bar") represent the mean, we have

$$\bar{x} = \frac{5 + 2 + 6 + 4 + 7 + 4 + 7 + 2 + 8 + 9 + 4 + 11 + 9 + 1 + 3}{15}$$

sum of values

$$= \frac{82}{15} = 5.5$$

number of values

Thus, the mean is 5.5. ■

If we wish to find the arithmetic mean of a large number of values, and if some of them appear more than once, the calculation can be simplified. The mean can be calculated by multiplying each value by its frequency, adding these results, and then dividing by the total number of values (the sum of the frequencies). Letting $\bar{x}$ represent the mean of the values $x_1, x_2, \ldots, x_n$, which occur with frequencies $f_1, f_2, \ldots, f_n$, respectively, we have

■ This is called a *weighted mean* since each value is given a weight based on the number of times it occurs.

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} \tag{22.1}$$

EXAMPLE 3 Arithmetic mean using frequencies

Using Eq. (22.1) to find the arithmetic mean of the numbers of Example 1, we first set up a table of values and their respective frequencies, as follows:

Values	1	2	3	4	5	6	7	8	9	11
Frequency	1	2	1	3	1	1	2	1	2	1

We now calculate the arithmetic mean $\bar{x}$ by using Eq. (22.1):

multiply each value by its frequency and add results

$$\bar{x} = \frac{1(1) + 2(2) + 3(1) + 4(3) + 5(1) + 6(1) + 7(2) + 8(1) + 9(2) + 11(1)}{1 + 2 + 1 + 3 + 1 + 1 + 2 + 1 + 2 + 1}$$

sum of frequencies

$$= \frac{82}{15} = 5.5$$

We see that this agrees with the result of Example 2. ■

Summations such as those in Eq. (22.1) occur frequently in statistics and other branches of mathematics. In order to simplify writing these sums, the symbol $\sum$ is used to indicate the process of summation. ($\sum$ is the Greek capital letter sigma.) $\sum x$ means the sum of the x's.

EXAMPLE 4 Summation symbol Σ

We can show the sum of the numbers $x_1, x_2, x_3, \ldots, x_n$ as

$$\sum x = x_1 + x_2 + x_3 + \cdots + x_n$$

If these numbers are 3, 7, 2, 6, 8, 4, and 9, we have

$$\sum x = 3 + 7 + 2 + 6 + 8 + 4 + 9 = 39$$ ■

Practice Exercises

For the following numbers, find the indicated value:
12, 17, 16, 12, 14, 18, 14, 12, 15, 18

1. The median **2.** The arithmetic mean

Using the summation symbol Σ, we can write Eq. (22.1) for the arithmetic mean as

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \cdots + x_n f_n}{f_1 + f_2 + f_3 + \cdots + f_n} = \frac{\sum xf}{\sum f} \qquad (22.1)$$

The summation notation $\sum x$ is an abbreviated form of the more general notation $\sum_{i=l}^{n} x_i$. This more general form can be used to indicate the sum of the first n numbers of a sequence or to indicate the sum of a certain set within the sequence. For example, for a set of at least 5 numbers, $\sum_{i=3}^{5} x_i$ indicates the sum of the third through the fifth of these numbers (in Example 4, $\sum_{i=3}^{5} x_i = 16$). We will use the abbreviated form $\sum x$ to indicate the sum of all the numbers being considered.

EXAMPLE 5 Arithmetic mean using frequencies

■ The arithmetic mean is one of a number of statistical measures that can be found on a calculator.

We find the arithmetic mean of the Internet hours in Example 1 of Section 22.1 (page 616) by

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{3(1) + 4(1) + 5(3) + \cdots + 12(6) + \cdots + 28(1)}{50}$$

$$= \frac{687}{50} = 13.7 \text{ h} \quad \text{(rounded off to tenths)}$$ ■

Mode

Another measure of central tendency is *the* **mode,** *which is the value that appears most frequently.* If two or more values appear with the same greatest frequency, each is a mode. If no value is repeated, there is no mode.

EXAMPLE 6 Mode

(a) The mode of the numbers in Example 1 is 4, since it appears three times, and no other value appears more than twice.

(b) The modes of the numbers

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

are 2 and 5, since each appears twice and no other number is repeated.

(c) There is no mode for the values

1, 2, 5, 6, 7, 9

since none of the values is repeated. ■

EXAMPLE 7 Measures of central tendency

To find the frictional force between two specially designed surfaces, the force to move a block with one surface along an inclined plane with the other surface is measured ten times. The results, with forces in newtons, are

2.2, 2.4, 2.1, 2.2, 2.5, 2.2, 2.4, 2.7, 2.1, 2.5

Find the mean, median, and mode of these forces.

LEARNING TIP

- The mean is useful for many statistical methods and is used extensively. Nevertheless, be aware that the mean is very sensitive to extreme observations so that a single extreme observation can change the value of the mean dramatically and give the wrong impression about the data.
- The median is not affected by extreme observations. Therefore, it is a good choice as a measure of centre in the presence of extreme values.
- The mean, the median, and the mode coincide when the distribution of data is symmetric. In those cases, the median or the mode (which are easy to calculate) can be used as estimates of the mean.

To find the mean, we sum the values of the forces and divide this total by 10. This gives

$$\bar{F} = \frac{\sum F}{10} = \frac{2.2 + 2.4 + 2.1 + 2.2 + 2.5 + 2.2 + 2.4 + 2.7 + 2.1 + 2.5}{10}$$

$$= \frac{23.3}{10} = 2.33 \text{ N}$$

The median is found by arranging the values in order and finding the middle value. The values in order are

$$2.1, 2.1, 2.2, 2.2, 2.2, 2.4, 2.4, 2.5, 2.5, 2.7$$

Since there are 10 values, we see that the fifth value is 2.2 and the sixth is 2.4. The value midway between these is 2.3, which is the median. Therefore, the median force is 2.3 N.

The mode is 2.2 N, since this value appears three times, which is more than any other value. ■

MEASURES OF SPREAD

Measures of central tendency on their own are not very informative. They do not tell us whether values are grouped closely together or how spread out they are. Therefore, we also need some measure of the deviation, or spread, of the values from the centre. If the spread is small and the numbers are grouped closely together, the measure of central tendency is more reliable and descriptive of the data than in the case in which the spread is greater.

Range

In statistics, there are several measures of spread that may be defined. The simplest one is the **range,** which is the difference between the highest value and the lowest value in the data set. For example, the range of the data in Example 7 is $2.7 - 2.1 = 0.6$. We will see how the range is applied to statistical process control in Section 22.5.

Standard Deviation

The most widely used measure of spread is the *standard deviation*. The **standard deviation** *of a set of* **sample** *values is defined by the equation*

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \tag{22.2}$$

The definition of s shows that the following steps are used in computing its value.

LEARNING TIP

The standard deviation s is a positive number. It is a *deviation from the mean,* regardless of whether the individual numbers are greater than or less than the mean. Numbers close together will have a small standard deviation, whereas numbers further apart have a larger standard deviation. Therefore, *the standard deviation becomes larger as the spread of data increases.*

Steps for Calculating Standard Deviation

1. *Find the arithmetic mean* $\bar{x}$ *of the numbers of the set.*
2. *Subtract the mean from each number of the set.*
3. *Square these differences.*
4. *Find the sum of these squares.*
5. *Divide this sum by* $n - 1$.
6. *Find the square root of this result.*

Following the steps shown above, we use Eq. (22.2) for the calculation of standard deviation in the following examples.

EXAMPLE 8 Standard deviation—using Eq. (22.2)

Find the standard deviation of the following numbers: 1, 5, 4, 2, 6, 2, 1, 1, 5, 3.

A table of the necessary values is shown below, and steps 1–6 are indicated:

step 2 step 3

x	$x - \bar{x}$	$(x - \bar{x})^2$
1	−2	4
5	2	4
4	1	1
2	−1	1
6	3	9
2	−1	1
1	−2	4
1	−2	4
5	2	4
3	0	0
30		32

step 4

$$\bar{x} = \frac{30}{10} = 3 \qquad \text{step 1}$$

$$\frac{\sum (x - \bar{x})^2}{n - 1} = \frac{32}{10 - 1} = \frac{32}{9} \qquad \text{step 5}$$

$$s = \sqrt{\frac{32}{9}} = 1.9 \qquad \text{step 6 (rounded off to tenths)}$$

■

LEARNING TIP

The *population standard deviation*, represented by the Greek letter σ (read "sigma"), is computed using n in the denominator of Eq. (22.2) instead of $n - 1$. The $n - 1$ in the denominator of Eq. (22.2) adjusts s so that it gives good estimates of the parameter σ when the standard deviation can only be measured from a sample.

Since we generally use data coming from samples, in this text we will always use Eq. (22.2), and we will refer to the sample standard deviation simply as the standard deviation.

If some of the values in the data are repeated, we can use the frequency of those values that occur more than once in calculating the standard deviation. This is illustrated in the following example.

EXAMPLE 9 Standard deviation using frequencies

Find the standard deviation of the numbers in Example 1.

Since several of the numbers appear more than once, it is helpful to use the frequency of each number in the table, as follows:

step 1

step 2 step 3

x	f	xf	$x - \bar{x}$	$(x - \bar{x})^2$	$(x - \bar{x})^2 f$
1	1	1	−4.5	20.25	20.25
2	2	4	−3.5	12.25	24.50
3	1	3	−2.5	6.25	6.25
4	3	12	−1.5	2.25	6.75
5	1	5	−0.5	0.25	0.25
6	1	6	0.5	0.25	0.25
7	2	14	1.5	2.25	4.50
8	1	8	2.5	6.25	6.25
9	2	18	3.5	12.25	24.50
11	1	11	5.5	30.25	30.25
	15	82			123.75

step 4

$$\bar{x} = \frac{82}{15} = 5.5$$

$$\frac{\sum (x - \bar{x})^2 f}{n - 1} = \frac{123.75}{15 - 1} = \frac{123.75}{14} \qquad \text{step 5}$$

$$s = \sqrt{\frac{123.75}{14}} = 3.0 \qquad \text{step 6}$$

■

It is possible to reduce the computational work required to find the standard deviation. Algebraically, it can be shown (although we will not do so here) that the following equation is another form of Eq. (22.2) and therefore gives the same results.

$$s = \sqrt{\frac{n\left(\sum x^2\right) - \left(\sum x\right)^2}{n(n-1)}} \qquad \textbf{(22.3)}$$

Although the form of this equation appears more involved, it does reduce the amount of calculation that is necessary. Consider the following examples.

EXAMPLE 10 Standard deviation—using Eq. (22.3)

Using Eq. (22.3), find s for the numbers in Example 8.

x	x^2
1	1
5	25
4	16
2	4
6	36
2	4
1	1
1	1
5	25
3	9
30	122

$$n = 10$$
$$\Sigma x^2 = 122$$
$$(\Sigma x)^2 = 30^2 = 900$$
$$s = \sqrt{\frac{10(122) - 900}{10(9)}} = 1.9$$

■

Practice Exercise

3. Find the standard deviation of the first eight numbers in Example 8.

COMMON ERROR

It is a common error to confuse Σx^2 and $(\Sigma x)^2$ in Eq. (22.3). Note that for Σx^2, we square the x values and then add the squares, whereas for $(\Sigma x)^2$, we first add the x values and then square the sum.

EXAMPLE 11 Standard deviation—application

An ammeter measures the electric current in a circuit. In an ammeter, two resistances are connected in parallel, with most of the current passing through a very low resistance called the *shunt*. The resistance of each shunt in a sample of 100 shunts was measured. The results were grouped, and the class mark and frequency for each class are shown in the following table. Calculate the arithmetic mean and the standard deviation of the resistances of the shunts.

■ Statistical measures such as $\bar{x}$, Σx, Σx^2, s_x, σ_x, and n can be obtained directly on a scientific or a graphing calculator.

R (ohms)	f	Rf	R^2f
0.200	1	0.200	0.0400
0.210	3	0.630	0.1323
0.220	5	1.100	0.2420
0.230	10	2.300	0.5290
0.240	17	4.080	0.9792
0.250	40	10.000	2.5000
0.260	13	3.380	0.8788
0.270	6	1.620	0.4374
0.280	3	0.840	0.2352
0.290	2	0.580	0.1682
	100	24.730	6.1421

$$\bar{R} = \frac{24.730}{100} = 0.2473\ \Omega$$
$$n = 100$$
$$\Sigma R^2 = 6.1421$$
$$(\Sigma R)^2 = 24.730^2$$
$$s = \sqrt{\frac{100(6.1421) - 24.730^2}{100(99)}} = 0.0163$$

The arithmetic mean of the resistances is 0.2473 Ω, with a standard deviation of 0.0163 Ω. ■

EXAMPLE 12 Standard deviation—application

Find the standard deviation of the estimated hours on the Internet as grouped in Example 2 of Section 22.1 (page 617). In doing this, we assume that each value in the class is the same as the class mark. The method is not exact, but with a large set of numbers, it provides a good approximation with less arithmetic work.

Interval	x	f	xf	x^2f
0–4	2	2	4	8
5–9	7	9	63	441
10–14	12	19	228	2736
15–19	17	11	187	3179
20–24	22	6	132	2904
25–29	27	3	81	2187
		50	695	11 455

$$n = 50$$
$$\Sigma x^2 = 11\,455$$
$$(\Sigma x)^2 = 695^2$$
$$s = \sqrt{\frac{50(11\,455) - 695^2}{50(49)}} = 6.1$$

Thus, $s = 6.1$ h. ■

The mean and the standard deviation together can help us draw conclusions about the values in a data set. Thanks to a result known as **Chebychev's theorem,** we can state the percentage of data values that must be within a specific number of standard deviations from the mean.

Chebychev's Theorem

For **any** data set (population or sample), the proportion of observations that must be within k standard deviations of the mean is always at least $1 - \frac{1}{k^2}$ $(k > 1)$.

For the particular values $k = 2$, 3, and 4, here is what the statement of the theorem implies:

- At least 75% of observations are within two standard deviations of the mean.
- At least 89% of observations are within three standard deviations of the mean.
- At least 94% of observations are within four standard deviations of the mean.

Note that since Chebychev's theorem is so general, it will underestimate the percentages for some distributions. In Section 22.3, we will obtain more precise percentages for the important case of the normal distribution.

EXAMPLE 13 An application of Chebychev's theorem

A sample of computers of a certain brand had a mean time of 38 months without a hardware malfunction, with a standard deviation of 2.5 months. What percentage of the computers in the sample lasted between 33 and 43 months without a hardware malfunction?

We can write $33 = 38 - 2(2.5)$, and $43 = 38 + 2(2.5)$, so 33 and 43 are 2 standard deviations away from the mean, and we use Chebychev's theorem with $k = 2$. Therefore, at least 75% of the computers in the sample lasted between 33 and 43 months without a hardware malfunction. ■

In using the statistical measures we have discussed, we must be careful in using and interpreting such measures. Consider the following example.

EXAMPLE 14 Interpreting statistical measures

(a) The numbers 1, 2, 3, 4, 5 have a mean of 3, a median of 3, and a standard deviation of 1.6. These values fairly well describe the centre and distribution of the numbers in the set.

(b) The numbers 1, 2, 3, 4, 100 have a mean of 22, a median of 3, and a standard deviation of 44. The large difference between the median and the mean and the very large range of values within one standard deviation of the mean (-22 to 66) indicate that this set of measures does not describe this set of numbers well. In a case like this, the 100 should be checked to see if it is in error. ■

Example 14 illustrates how statistical measures can be misleading in the presence of extreme values. Misleading statistics can also come from the process of data collection.

Consider the probable results of a survey to find the percent of persons in favour of raising income taxes for the wealthy if the survey is taken at the entrance to a welfare office or if it is taken at the entrance to a stock brokerage firm. There are many other considerations in the proper use and interpretation of statistical measures.

EXERCISES 22.2

In Exercises 1–4, delete the 5 from the data numbers given for Example 1 and then do the following with the resulting data.

1. Find the median.

2. Find the arithmetic mean using the definition, as in Example 2.

3. Find the arithmetic mean using Eq. (22.1), as in Example 3.

4. Find the mode, as in Example 6.

In Exercises 5 and 6, use the data in Example 8. Change the first 1 to 6 and the first 2 to 7 and then find the standard deviation of the resulting data as directed.

5. Find s from the definition, as in Example 8.

6. Find s using Eq. 22.3, as in Example 10.

7. In Example 13, change 33 to 28 and 43 to 48 and then find the percentage.

In Exercises 8–15, use the following sets of numbers.

A: 3, 6, 4, 2, 5, 4, 7, 6, 3, 4, 6, 4, 5, 7, 3

B: 25, 26, 23, 24, 25, 28, 26, 27, 23, 28, 25

C: 0.48, 0.53, 0.49, 0.45, 0.55, 0.49, 0.47, 0.55, 0.48, 0.57, 0.51, 0.46, 0.53, 0.50, 0.49, 0.53

D: 105, 108, 103, 108, 106, 104, 109, 104, 110, 108, 108, 104, 113, 106, 107, 106, 107, 109, 105, 111, 109, 108

In Exercises 8–11, determine (a) the mean, (b) the median, and (c) the mode of the numbers of the given set.

8. Set *A* **9.** Set *B* **10.** Set *C* **11.** Set *D*

In Exercises 12–15, find the standard deviation s for the indicated set of numbers (a) using Eq. (22.2), and *(b) using Eq. (22.3).*

12. Set *A* **13.** Set *B* **14.** Set *C* **15.** Set *D*

In Exercises 16–30, the required data are those in Exercises 22.1.

16. Find the mean, the median, and the mode of L/100 km of fuel usage in Exercise 5.

17. Find the standard deviation of L/100 km of fuel usage in Exercise 5 using Eq. (22.2).

18. Find the standard deviation of L/100 km of fuel usage in Exercise 5 using Eq. (22.3).

19. Find the mean, the median, and the mode of computer instructions in Exercise 13.

20. Find the range and standard deviation of computer instructions in Exercise 13.

21. Find the mean and the median of strobe light times in Exercise 17.

22. Find the standard deviation of strobe light times in Exercise 17.

23. Find the mean and median of stopping distances in Exercise 21. (Use the class mark for each class.)

24. Find the standard deviation of stopping distances in Exercise 21. (Use the class mark for each class.)

25. Find the mean, the median, and the mode of X-ray dosages in Exercise 25.

26. Find the range and the standard deviation of X-ray dosages in Exercise 25.

27. Find the mean, the median, and the mode of battery lives in Exercise 27.

28. Find the standard deviation of battery lives in Exercise 27.

29. Find the mean and the median of cable diameters in Exercise 29.

30. Find the standard deviation of cable diameters in Exercise 29.

In Exercises 31–47, find the indicated measure of central tendency or of spread.

31. The weekly salaries (in dollars) for the workers in a small factory are as follows:
600, 750, 625, 575, 525, 700, 550,
750, 625, 800, 700, 575, 600, 700

Find the median and the mode of the salaries.

32. Find the mean salary for the salaries in Exercise 31.

33. Find the range and the standard deviation of the salaries in Exercise 31.

34. In a particular month, the electrical usage, rounded to the nearest 400 MJ, of 1000 homes in a certain city was summarized as follows:

Usage	2000	2400	2800	3200	3600	4000	4400	4800
No. Homes	22	80	106	185	380	122	90	15

Find the mean of the electrical usage.

35. Find the median and mode of electrical usage in Exercise 34.

36. Find the standard deviation of electrical usage in Exercise 34.

37. A test of air pollution in a city gave the following readings of the concentration of sulfur dioxide (in parts per million) for 18 consecutive days:

0.14, 0.18, 0.27, 0.19, 0.15, 0.22, 0.20, 0.18, 0.15,
0.17, 0.24, 0.23, 0.22, 0.18, 0.32, 0.26, 0.17, 0.23

Find the median and the mode of these readings.

38. Find the mean of the readings in Exercise 37.

39. Find the range and the standard deviation of air pollution data in Exercise 37.

40. The following data give the mean number of days of rain for Vancouver, British Columbia, for the 12 months of the year.

20, 17, 17, 14, 12, 11, 7, 8, 9, 16, 19, 22

Find the standard deviation.

41. The *midrange,* another measure of central tendency, is found by finding the sum of the lowest and the highest values and dividing this sum by 2. Find the midrange of the salaries in Exercise 31.

42. Find the midrange of the sulfur dioxide readings in Exercise 37. (see Exercise 41.)

43. Add \$100 to each of the salaries in Exercise 31. Then find the median, mean, and mode of the resulting salaries. State any conclusion that might be drawn from the results.

44. Multiply each of the salaries in Exercise 31 by 2. Then find the median, mean, and mode of the resulting salaries. State any conclusion that might be drawn from the results.

45. Change the final salary in Exercise 31 to \$4000, with all other salaries being the same. Then find the mean of these salaries. State any conclusion that might be drawn from the result. (The \$4000 here is called an *outlier,* which is an extreme value.)

46. Find the median and mode of the salaries indicated in Exercise 45. State any conclusion that might be drawn from the results.

47. The $k\%$ *trimmed mean* is a measure of central tendency that avoids the influence of extreme observations while still using most of the observations in the data set. It is computed by finding the mean of the data after the smallest $k\%$ and the largest $k\%$ of the data have been discarded. Find the 10% trimmed mean of X-ray dosages in Exercise 25 of Section 22.1.

In Exercises 48–50, solve the given problems.

48. Use Chebychev's theorem to find the percentage of values that are between 175 and 195 in a data set with mean 185 and standard deviation 5.

49. Use Chebychev's theorem to find the percentage of values that are between 55.7 and 68.3 in a data set with mean 62 and standard deviation 2.1.

50. The mean compressive strength of a sample of steel beams was 40 000 N/cm^2, with a standard deviation of 450 N/cm^2. What percent of the beams had compressive strength between 38 650 and 41 350 N/cm^2?

Answers to Practice Exercises

1. 14.5 **2.** 14.8 **3.** 2.0

22.3 Normal Distributions

Normal Distribution of a Population • Standard Normal Distribution • Standard Score (z-Score) • Sampling Distributions

■ The first derivation of the normal distribution is due to Abraham de Moivre (1667–1754), who was interested in approximating quantities arising in gambling problems. It was also derived independently by Pierre-Simon Laplace (1749–1847), and by Carl Friedrich Gauss (1777–1855), both in the context of measurement errors.

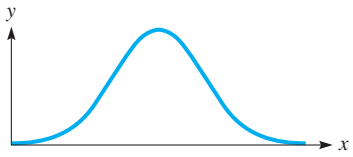

Fig. 22.6

In this section we discuss the **normal distribution**, the most important and most widely used distribution in statistics. The normal distribution is a continuous distribution, so we begin by discussing some generalities of continuous distributions.

In Section 22.1 we learned that for variables that take a limited number of values, the relative frequency of a value is obtained by dividing the frequency of that value by the total frequency of all values. Let us now consider variables that can be regarded as having an infinite number of possible values, such as weights, lengths, or durations for a very large population. (Such variables are said to be *continuous*.) When using the same procedure as before, the denominator becomes infinite, giving a relative frequency of zero for all values. How are we then to compute the relative frequency of intervals, if the relative frequency of all values is zero?

The answer lies in establishing a correspondence between relative frequency and *area*. To each continuous variable we associate a function, which we can graph as a curve on the plane. The relative frequency of a particular interval corresponds to the area under the curve in that interval. Because the relative frequency for the complete population must be 1, the total area under the curve must be 1 (100% of the data).

The **normal distribution** is associated with the symmetric, bell-shaped curve shown in Fig. 22.6. Using advanced methods, its equation is found to be

$$y = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \tag{22.4}$$

Here, μ is the *population mean* and σ is the *population standard deviation,* and π and e are the familiar numbers first used in Chapters 2 and 12, respectively.

From Eq. (22.4), we can see that any particular normal distribution depends on the values of μ and σ. The horizontal location of the curve depends on μ, and the shape (how spread out the curve is) depends on σ, but the bell shape remains. This is illustrated in general in the following example.

EXAMPLE 1 Location and spread of a normal distribution

In Fig. 22.7, for the left curve, $\mu = 10$ and $\sigma = 5$, whereas for the right curve, $\mu = 20$ and $\sigma = 10$.

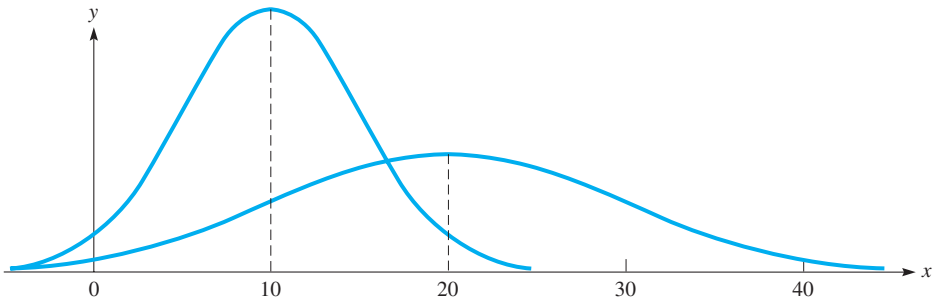

Fig. 22.7 ■

EXAMPLE 2 The normal distribution applied to reliability

■ See the chapter introduction.

The normal distribution has important applications in probabilistic reliability techniques for bridge design. For instance, load factors developed for the design of the Confederation Bridge are discussed in the article "Design criteria and load and resistance factors for the Confederation Bridge," by J. G. MacGregor et al. (*Can. J. Civ. Eng.*, 24, 882–897 (1997)). Quantities that were found to be normally distributed (or whose logarithms were found to be normally distributed) arose in the analysis of dead loads, live loads due to vehicles, wind loads, temperature loads, and ice loads. To give a specific example, after analysing records of daily average temperatures in the region for 46 years, it was concluded that the 3-day temperature drop that was equalled or exceeded 100 times in 100 years is distributed normally, with mean 26.9°C and standard deviation 3.2°C. This distribution is shown in Fig. 22.8. ■

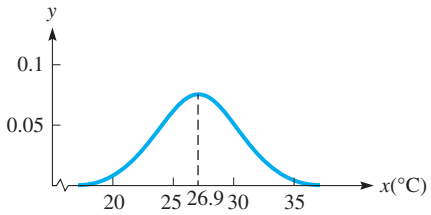

Fig. 22.8

Properties of the Normal Curve

- The curve is symmetric about the mean.
- The curve is always above the x-axis. (y is always positive.)
- The x-axis is a horizontal asymptote. (As x increases numerically, y becomes very small.)
- The total area under the curve is 1.
- The curve is bell-shaped, as seen in Figs. 22.6–22.8.

■ A more complete and rigorous treatment of the material covered in this and the remaining sections of this chapter would require the study of probability theory, which is beyond the scope of this introductory chapter.

Fig. 22.9 shows areas under a normal curve with mean μ and standard deviation σ for particular regions. Using the correspondence between relative frequency and area, we can use the given information to find the percentage of data values that fall within one, two, and three standard deviations from the mean, as summarized below.

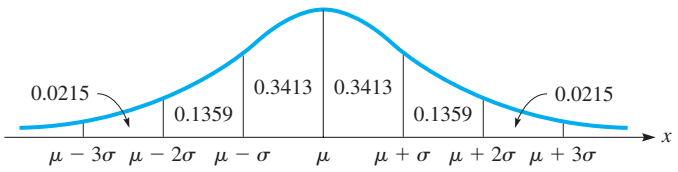

Fig. 22.9

Important Percentages for the Normal Curve

- About 68% of values are within one standard deviation from the mean–that is, between $\mu - \sigma$ and $\mu + \sigma$.
- About 95% of values are within two standard deviations from the mean–that is, between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- Almost all values (99.74%) are within three standard deviations from the mean–that is, between $\mu - 3\sigma$ and $\mu + 3\sigma$.

We can compare these percentages with those given by Chebychev's theorem in Section 22.2—namely, that at least 75% of observations are within two standard deviations from the mean and that at least 89% of observations are within three standard deviations from the mean. We see that Chebychev's theorem heavily underestimates the true percentages in the case of the normal distribution.

We can use the areas from Fig. 22.9 to calculate relative frequencies for other intervals. This is illustrated in the following example.

EXAMPLE 3 Relative frequencies and areas under the curve—application

Let us consider the normal distribution of 3–day temperature drops from Example 2, so that the mean and standard deviation are $\mu = 26.9°C$ and $\sigma = 3.2°C$, respectively. The percentage of values that lie between one standard deviation below the mean and two standard deviations above the mean is 81.85%, since the area between $\mu - \sigma$ and $\mu + 2\sigma$ is $0.3412 + 0.3413 + 0.1359 = 0.8185$. We have

$$\mu - \sigma = 26.9 - 3.2 = 23.7 \text{ and } \mu + 2\sigma = 26.9 + 2(3.2) = 33.3$$

Therefore, about 81.85% of values for this distribution are within 23.7°C and 33.3°C. ■

STANDARD NORMAL DISTRIBUTION

As we have just seen, there are innumerable possible normal distributions. However, there is one of particular interest. *The* **standard normal distribution** *is the normal distribution for which the mean is* 0 *and the standard deviation is* 1. Making these substitutions in Eq. (22.4), we have

$$y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{22.5}$$

as the equation of the standard normal distribution curve. All the properties of a normal curve are satisfied by the standard normal distribution. In particular, since the mean is 0, the curve is symmetric with respect to the y-axis. The curve is also bell–shaped, as seen in Fig. 22.10. As we discuss below, areas under the standard normal curve are used to find relative frequencies for all other normal curves.

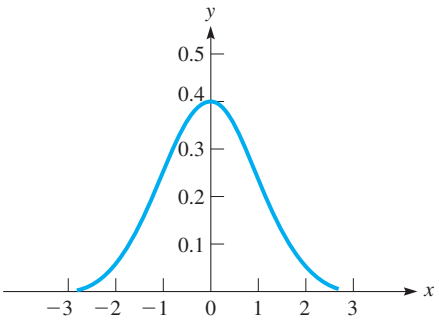

Fig. 22.10

We can find the relative frequency of values for any normal distribution by use of the **standard score** z (or z-score), which is defined as

$$z = \frac{x - \mu}{\sigma} \tag{22.6}$$

For the standard normal distribution, where $\mu = 0$, if we let $x = \sigma$, then $z = 1$. If we let $x = 2\sigma$, $z = 2$. Therefore, we can see that *a value of z tells us the number of standard deviations the given value of x is above or below the mean.* From the discussion above, we can see that the value of z can tell us the area under the curve between the mean and the value of x corresponding to that value of z. In turn, *this tells us the relative frequency of all values between the mean and the value of x.*

Table 22.1 gives the area under the standard normal distribution curve between zero and the given values of z. The table includes values only to $z = 3$ since nearly all of the area is between $z = -3$ and $z = 3$. Since the curve is symmetric to the y-axis, the values shown are also valid for negative values of z.

Table 22.1 Standard Normal (z) Distribution

z	*Area*	z	*Area*	z	*Area*
0.0	0.0000	1.0	0.3413	2.0	0.4772
0.1	0.0398	1.1	0.3643	2.1	0.4821
0.2	0.0793	1.2	0.3849	2.2	0.4861
0.3	0.1179	1.3	0.4032	2.3	0.4893
0.4	0.1554	1.4	0.4192	2.4	0.4918
0.5	0.1915	1.5	0.4332	2.5	0.4938
0.6	0.2257	1.6	0.4452	2.6	0.4953
0.7	0.2580	1.7	0.4554	2.7	0.4965
0.8	0.2881	1.8	0.4641	2.8	0.4974
0.9	0.3159	1.9	0.4713	2.9	0.4981
1.0	0.3413	2.0	0.4772	3.0	0.4987

The following examples illustrate the use of Eq. (22.6) and z-scores.

EXAMPLE 4 Normal score (z-score)

For a normal distribution curve based on values of $\mu = 20$ and $\sigma = 5$, find the area between $x = 24$ and $x = 32$. To find this area, we use Eq. (22.6) to find the corresponding values of z and then find the difference between the areas associated with these z-scores. These z-scores are

$$z = \frac{24 - 20}{5} = 0.8 \quad \text{and} \quad z = \frac{32 - 20}{5} = 2.4$$

For $z = 0.8$, the area is 0.2881, and for $z = 2.4$, the area is 0.4918. Therefore, the area between $x = 24$ and $x = 32$ (see Fig. 22.11) is

$$0.4918 - 0.2881 = 0.2037$$

This means that the relative frequency of the values between $x = 24$ and $x = 32$ is 20.37%. If we have a large set of measured values with $\mu = 20$ and $\sigma = 5$, we should expect that about 20% of them are between $x = 24$ and $x = 32$. ■

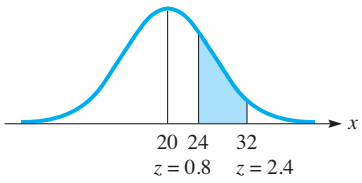

Fig. 22.11

Practice Exercise

1. For values of $\mu = 40$ and $\sigma = 8$, find the area between $x = 36$ and $x = 48$.

EXAMPLE 5 z-score—application

The lifetimes of a certain type of watch battery are normally distributed. The mean lifetime is 400 days, and the standard deviation is 50 days. For a sample of 5000 new batteries, determine how many batteries are expected to last **(a)** between 360 days and 460 days, **(b)** more than 320 days, and **(c)** less than 280 days.

(a) For this distribution, $\mu = 400$ days and $\sigma = 50$ days. Using Eq. (22.6), we find the z-scores for $x = 360$ days and $x = 460$ days. They are

$$z = \frac{360 - 400}{50} = -0.8 \quad \text{and} \quad z = \frac{460 - 400}{50} = 1.2$$

For $z = -0.8$, the area is to the left of the mean, and since the curve is symmetric about the mean, we use the $z = 0.8$ value of the area and add it to the area for $z = 1.2$. Therefore, the area is

$$0.2881 + 0.3849 = 0.6730$$

See Fig. 22.12. This means that 67.30% of the 5000 batteries, or 3365 of the batteries, are expected to last between 360 and 460 days. Because of variability within samples, not every sample will have exactly 3365 batteries that will last between

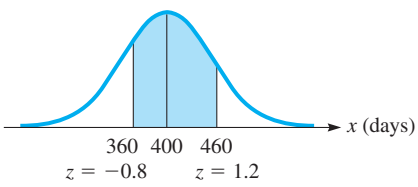

Fig. 22.12

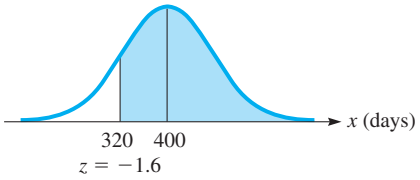

Fig. 22.13

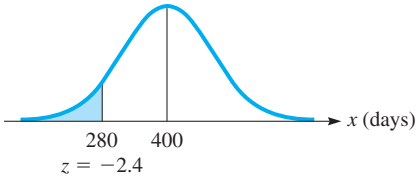

Fig. 22.14

360 and 460 days. On average, however, a sample of 5000 batteries will have 3365 that will last that amount of time.

(b) To determine the number of batteries that will last more than 320 days, we first find the z-score for $x = 320$. It is $z = (320 - 400)/50 = -1.6$. This means we want the total area to the right of $z = -1.6$. In this case, we add the area for $z = 1.6$ to the total area to the right of the mean. Since the total area under the curve is 1.0000, the total area on either side of the mean is 0.5000. Therefore, the area to the right of $z = -1.6$ is $0.4452 + 0.5000 = 0.9452$. See Fig. 22.13. This means that $0.9452 \times 5000 = 4726$ batteries are expected to last more than 320 days.

(c) To find the number of batteries that will last less than 280 days, we first find that $z = (280 - 400)/50 = -2.4$ for $x = 280$. Since we want the total area to the left of $z = -2.4$, we subtract the area for $z = 2.4$ from 0.5000, the total area to the left of the mean. Since the area for $z = 2.4$ is 0.4918, the total area to the left of $z = -2.4$ is $0.5000 - 0.4918 = 0.0082$. See Fig. 22.14. Therefore, $0.0082 \times 5000 = 41$ batteries are expected to last less than 280 days. ■

We now summarize the procedure for finding relative frequencies using z-scores.

Finding Relative Frequencies Using *z*-Scores

1. Sketch the normal curve, labeling the mean and the given x values. Identify the desired relative frequency as an area under the curve.
2. Use Eq. (22.6) to find the z-score for each x. Identify the desired relative frequency as an area under the standard normal curve.
3. Look up the absolute value of each z-score in Table 22.1 to find its associated area.
4. Depending on the situation, proceed as follows:

Area	Procedure
Between two z-scores of the same sign	Subtract the smaller area from the larger one
Between two z-scores of different sign	Add both areas together
To the right of a positive z-score or to the left of a negative z-score	Subtract the area from 0.5
To the right of a negative z-score or to the left of a positive z-score	Add the area to 0.5

SAMPLING DISTRIBUTIONS

In Example 4, we assumed that the lifetimes of the batteries were normally distributed. Of course, for any set of 5000 batteries, or any number of batteries for that matter, the lifetimes that actually occur will not follow a normal distribution *exactly*. There will be some variation from the normal distribution, but for a large sample, this variation should be small. The mean and the standard deviation for any sample will vary somewhat from that of the population. When we consider the relative frequency distribution of the sample means obtained from all possible samples of the same size, we obtain what is called the **sampling distribution** of the sample means.

In the study of probability, it is shown that if we select all possible samples of size n from a population with a mean μ and standard deviation σ, the mean of the sample means is also μ. Also, *the standard deviation of the sample means,* denoted by $\sigma_{\bar{x}}$, and called **the standard error of the mean,** is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{22.7}$$

Moreover, when n is large (*large* in this situation is usually considered to be over 30) the sampling distribution of the sample means is approximately normal. In other words, the normal curve approximates the relative frequency distribution of the sample means, so that areas under the normal curve can be used to approximate relative frequencies of the sample means.

The normal distribution also approximates the sampling distribution of quantities calculated from samples when the variable of interest is an attribute. For instance, suppose that we take samples of n items from a very large population containing a proportion p of defective items. If n is large (in this case, np and $n(1-p)$ **must both be at least 5**), then the sampling distribution of the sample proportion $\hat{p}$ of defective items in each sample is also approximately normal. In this situation, the mean of the sample proportions is p, and the standard error of $\hat{p}$ is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{22.8}$$

We can see from Eqs. (22.7) and (22.8) that as the sample size gets larger, the less variation there will be in the mean or the proportion obtained from a sample.

COMMON ERROR

It is a common error to forget the $\sqrt{n}$ term in the denominator of the standard error formulas. It is very important to include it since it is what guarantees that variation between samples decreases as the sample size increases.

EXAMPLE 6 Sampling distribution of the sample mean

For the sample of 5000 watch batteries in Example 5, we know that $\sigma = 50$ days. Therefore, the standard error of the mean $\bar{x}$ is $50/\sqrt{5000} = 0.7$ day. This means that of all samples of 5000 batteries, about 68% should have a mean lifetime of 400 ± 0.7 day (between 399.3 days and 400.7 days). Considering the significant digits of these values, 68% (within one standard deviation) or even 95% (within two standard deviations) of the sample values of $\bar{x}$ would not vary by more than 1 day. ■

EXAMPLE 7 Sampling distribution of the sample proportion

Of the items in a very large population, 10% are defective. Samples of size 200 are taken from this population, and the proportion of defectives in each sample is recorded. The normal approximation applies since $np = 200(0.1) = 20$ and $n(1-p) = 200(0.9) = 180$. Therefore, the sample proportion $\hat{p}$ is approximately normally distributed, with mean 0.1 and standard error $\sqrt{(0.10)(0.90)/200} = 0.021$, or 2.1%. This means that of all samples of size 200 taken from this population, about 68% would have a sample proportion of defectives of 0.1 ± 0.021 (i.e., between 7.9% and 12.1% of the sample would be defective in 68% of samples). ■

EXERCISES 22.3

In Exercises 1–4, make the given changes in the indicated examples of this section and then solve the indicated problem.

1. In Example 1, change the second σ from 10 to 5 and then describe the curve that would result in terms of either or both curves shown in Fig. 22.7.

2. In Example 3, change two to three and then find the resulting percent of values and the corresponding interval.

3. In Example 4, change $x = 32$ to $x = 33$ and then find the resulting area.

4. In Example 5(b), change 320 to 360 and then find the resulting number of batteries.

In Exercises 5–8, use a graphing calculator to display the indicated graph of Eq. (22.4).

5. Display the graph of the normal distribution of values for which $\mu = 10$ and $\sigma = 5$. Compare with the graph shown in Fig. 22.7.

6. Display the graph of the normal distribution of values for which $\mu = 20$ and $\sigma = 10$. Compare with the graph shown in Fig. 22.7.

7. Sketch a graph of a normal distribution of values for which $\mu = 100$ and $\sigma = 10$. Then compare with the graph displayed by a graphing calculator.

8. Sketch a graph of a normal distribution of values for which $\mu = 100$ and $\sigma = 30$. Then compare with the graph displayed by a graphing calculator. How does this graph differ from that of Exercise 7?

In Exercises 9–12, use the following data and refer to Fig. 22.9. A sample of 200 bags of cement are weighed as a quality check. Over a long period, it has been found that the mean value and standard deviation for this size bag are known and that the weights are normally distributed. Determine how many bags within this sample are expected to have weights that satisfy the following conditions.

9. Within one standard deviation of the mean

10. Within two standard deviations of the mean

11. Between the mean and two standard deviations above the mean

12. Between one standard deviation below the mean and three standard deviations above the mean

In Exercises 13–16, use the following data. Each AA battery in a sample of 500 batteries is checked for its voltage. It has been previously established for this type of battery (when newly produced) that the voltages are distributed normally with $\mu = 1.50$ V *and* $\sigma = 0.05$ V.

13. How many batteries are expected to have voltages between 1.45 V and 1.55 V?

14. How many batteries are expected to have voltages between 1.52 V and 1.58 V?

15. What percent of the batteries are expected to have voltages below 1.54 V?

16. What percent of the batteries are expected to have voltages above 1.64 V?

In Exercises 17–22, use the following data. The lifetimes of a certain type of automobile tire have been found to be distributed normally with a mean lifetime of 100 000 km *and a standard deviation of 10 000* km. *Answer the following questions for a sample of 5000 of these tires.*

17. How many tires are expected to last between 85 000 km and 100 000 km?

18. How many tires are expected to last between 95 000 km and 115 000 km?

19. How many tires are expected to last more than 118 000 km?

20. If the manufacturer guarantees to replace all tires that do not last 75 000 km, what percent of the tires may have to be replaced under this guarantee?

21. What is the standard error in the mean for all samples of 5000 of these tires? Explain the meaning of this result.

22. What percent of the samples of 5000 of these tires should have a mean lifetime of more than 100 282 km?

In Exercises 23–32, solve the given problems.

23. Find the standard error of the proportion of defective items in samples of size 500 taken from a very large population of which 12% of the items are defective. Explain the meaning of this result.

24. Of the 300 mL bottles filled by a certain filling machine, 1% contain less than 290 mL of juice. If samples of 600 bottles produced by this machine are selected, find the standard error of the sample proportion of bottles that contain less than 290 mL. Explain the meaning of this result.

25. With 75.8% of the area under the normal curve to the right of z, find the z-value.

26. With 21% of the area under the normal curve between z_1 and z_2, to the right of $z_1 = 0.8$, find z_2.

27. With 59% of the area under the normal curve between z_1 and z_2, to the left of $z_2 = 1.1$, find z_1.

28. With 5.8% of the area under the normal curve between z_1 and z_2, to the left of $z_2 = 2.0$, find z_1.

29. For the strobe light times in Exercise 17 of Section 22.1, find the percent of times within one standard deviation of the mean. From Exercises 21 and 22 of Section 22.2, we find that $\bar{x} = 2.248$ s and $s = 0.014$ s. Compare the results with that of a normal distribution.

30. Follow the same instructions as in Exercise 29 for the fibre-optic diameters in Exercise 29 of Section 22.1. From Exercises 29 and 30 of Section 22.2, $\bar{x} = 0.005\,95$ mm and $s = 0.000\,22$ mm.

31. Follow the same instructions as in Exercise 29 for the hours estimated on the Internet in Example 1 of Section 22.1. From Examples 5 and 12 of Section 22.2, we find that $\bar{x} = 13.7$ h and $s = 6.1$ h.

32. Discuss the results found in Exercises 29 and 31, considering the methods used to find the mean and the standard deviation.

Answer to Practice Exercise

1. $z = 0.5328$

22.4 Confidence Intervals

Estimators • Confidence Level • Confidence Intervals for Means • Margin of Error • Confidence Intervals for Proportions • Determining Sample Size

In this section we begin our study of inferential statistics, where information from a sample is used to make statements about a whole population. We focus on the problem of estimation of parameters, under the assumption that data are available for a random sample taken from a very large population.

When a parameter is being estimated, the estimate can be a single number (called a **point estimate**), or it can be a range of numbers (called a **confidence interval**). Consider the following example.

EXAMPLE 1 Two kinds of estimates

Consider the data on Internet hours in Example 1 of Section 22.1. If the sample of 50 users constitutes a random sample of a large population of users of home computers, then the information obtained from the sample can be used to estimate the population mean.

On the one hand, the sample mean $\bar{x} = 13.7$ is a point estimate of the population mean. Because it is a single number, this estimate does not convey information about the reliability of the estimation.

On the other hand, the interval $13.7 \pm 1.7 = (12.0\text{ h}, 15.4\text{ h})$ is a 95% confidence interval estimate of the population mean. This means that we are 95% confident that the true value of the population mean is between 12.0 h and 15.4 h. It is preferable to use a confidence interval because the length of the interval and the level of confidence attached to it give us an idea about the reliability of our estimation, in a sense that will be made precise below. ■

All confidence intervals are calculated by first selecting a **confidence level,** which measures the degree of certainty that the confidence interval will contain the population parameter. The most common values for the confidence level are 90%, 95%, and 99%, with the most common one being 95%. A confidence level of 95% means that, of all possible samples of size n taken from the same population, 95% of them will give an interval that will contain the population parameter, and 5% of them will not. For a particular sample, it is not possible to know whether it is one of the successful ones or not.

COMMON ERROR

It is a common error to interpret the level of confidence as measuring the likelihood that the parameter of interest will fall within a particular interval. There is nothing random about the parameter; its value is a constant (unfortunately unknown to us), and either our interval covers it or it does not. The randomness lies in the sample, so the level of confidence is the likelihood that a **sample of size *n*** will cover the parameter. At the 95% confidence level, 95% of the samples will, and 5% of the samples will not. We acknowledge that there is a 5% risk that the particular interval obtained will not cover the parameter.

The method for constructing a confidence interval depends on the parameter(s) being estimated and on the characteristics of the sample. We will concentrate our attention on *large sample confidence intervals for a single mean and a single proportion.* We will obtain general formulas for these two cases using the normal distribution theory developed in the previous section.

LARGE SAMPLE CONFIDENCE INTERVALS FOR THE MEAN

Suppose that we have a random sample of size n (n large) from a population with unknown mean μ. We construct a 95% confidence interval for the mean μ, under the assumption that the population standard deviation σ is known.

In Section 22.3, we learned that the sampling distribution of the sample mean is approximately normal, with mean μ and standard error $\sigma/\sqrt{n}$. It is found from tables that the area under the standard normal curve between -1.96 and 1.96 is 0.95, so that 95% of the observations from a normal distribution fall within 1.96 standard deviations of the mean. Applying this to the sampling distribution of the sample mean, we have that 95% of all samples of size n will have a sample mean $\bar{x}$ that falls within 1.96 standard errors of the true mean. In other words, in 95% of samples, the sample mean $\bar{x}$ satisfies the inequality

$$\mu - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

We can manipulate this inequality in order to transform it into a statement about the unknown population mean μ. We have

$$\mu - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \qquad \text{original inequality}$$

$$-1.96 \cdot \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < 1.96 \cdot \frac{\sigma}{\sqrt{n}} \qquad \text{subtract } \mu \text{ from each member}$$

$$-\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \qquad \text{subtract } \bar{x} \text{ from each member}$$

$$\bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} > \mu > \bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \qquad \text{multiply by } -1 \text{ (reverse the inequality)}$$

$$\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

This last inequality is equivalent to the original one and is therefore satisfied for 95% of samples of size n. In other words,

$$\left(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \tag{22.9}$$

is a 95% confidence interval for μ.

The endpoints of the interval in Eq. (22.9) are often written in the form

$$\bar{x} \pm E, \text{ with } E = 1.96 \cdot \frac{\sigma}{\sqrt{n}} \tag{22.10}$$

The quantity E is called the **margin of error,** and it represents the largest estimated difference between the estimate and the true value of the parameter.

EXAMPLE 2 A 95% confidence interval—σ known

A random sample of size $n = 100$ is taken from a population with $\sigma = 2.3$. Construct a 95% confidence interval for the population mean μ if the sample mean is $\bar{x} = 32.8$.

We substitute the given values of $\bar{x}$, σ, and n into Eq. (22.10). The resulting 95% confidence interval is

$$\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} = 32.8 \pm 1.96 \cdot \frac{2.3}{\sqrt{100}} = 32.8 \pm 0.451 = (32.3, 33.3)$$

Because the interval obtained is narrow, the estimation is quite precise. ■

When the standard deviation σ is unknown and the sample size is large, the sample standard deviation s can be used to estimate the population standard deviation σ. Therefore, a 95% confidence interval for μ when the standard deviation σ is unknown is given by

$$\bar{x} \pm E, \text{ with } E = 1.96 \cdot \frac{s}{\sqrt{n}} \tag{22.11}$$

Practice Exercise

1. Find a 95% confidence interval for a mean μ if a sample with $n = 45$ gives $\bar{x} = 97.6$ and $s = 3.2$.

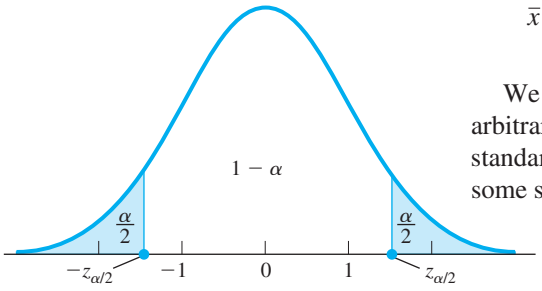

Fig. 22.15

Table 22.2 Common Values of $z_{\alpha/2}$

Confidence Level $(1 - \alpha)100\%$	$\alpha/2$	$z_{\alpha/2}$
90%	0.05	1.645
95%	0.025	1.96
99%	0.005	2.575

Practice Exercise

2. Find a 90% confidence interval for a mean μ if a sample with $n = 45$ gives $\bar{x} = 97.6$ and $s = 3.2$.

EXAMPLE 3 A 95% confidence interval—σ estimated

Let us derive the confidence interval for the mean stated in Example 1, obtained from the Internet hours data from Example 1 of Section 22.1. The sample mean and sample standard deviation for the 50 observations are $\bar{x} = 13.7$ h and $s = 6.1$ h, respectively. Substituting these values into Eq. (22.11), the resulting 95% confidence interval is

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}} = 13.7 \pm 1.96 \cdot \frac{6.1}{\sqrt{50}} = 13.7 \pm 1.69 = (12.0\text{ h}, 15.4\text{ h}) \quad \blacksquare$$

We now derive the formula for a large sample confidence interval for a mean for an arbitrary confidence level $1 - \alpha$. Here we have written the confidence level as is standard for general formulas, expressing it as a decimal whose value is $1 - \alpha$ for some small α (for example, $\alpha = 0.05$ for a 95% confidence interval).

Let $z_{\alpha/2}$ denote the score such that the area under the standard normal curve between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is $1 - \alpha$. See Fig. 22.15 and Table 22.2. By replacing 1.96 with $z_{\alpha/2}$ in the inequalities leading to Eq. (22.10), we obtain the following general formula.

Large Sample Confidence Interval for a Mean

A $(1 - \alpha)100\%$ confidence interval for the mean μ when the sample size is large $(n \geq 30)$ is given by

$$\bar{x} \pm E, \text{ where } \quad E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad \text{if } \sigma \text{ is known}$$

$$E = z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \quad \text{if } \sigma \text{ is unknown} \qquad (22.12)$$

EXAMPLE 4 A 99% confidence interval

Construct a 99% confidence interval for the Internet hours data from Example 1 of Section 22.1. Recall that $\bar{x} = 13.7$ h, $s = 6.1$ h, and $n = 50$.

From Table 22.2 we get that $z_{\alpha/2} = 2.575$. We substitute the given values into Eq. (22.12). The desired 99% confidence interval is

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 13.7 \pm 2.575 \cdot \frac{6.1}{\sqrt{50}} = 13.7 \pm 2.22 = (11.5\text{ h}, 15.9\text{ h}) \quad \blacksquare$$

COMMON ERROR

Always state a confidence interval together with its confidence level. ***A confidence interval by itself is meaningless.***

Comparing the intervals in Example 3 and Example 4, we see that by increasing the confidence level, the margin of error increased as well, so that what we have gained in confidence, we have lost in precision. By examining the formula for the margin of error in Eq. (22.12), we can find the general relationship between confidence level, margin of error, and sample size.

Confidence Level, Margin of Error, and Sample Size Relationships

- For a fixed sample size, a higher confidence level implies a larger margin of error. What we gain in confidence we lose in precision.
- For a fixed sample size, a smaller margin of error implies a lower confidence level. What we gain in precision we lose in confidence.
- The only way we can increase the confidence level while at the same time decrease the margin of error is to increase the sample size.

The formula for E in Eq. (22.12) for known σ can be used to determine the sample size needed for a desired margin of error at a fixed confidence level. Suppose that a maximum margin of error E is desired with a confidence level $1 - \alpha$. We have

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \qquad \text{from Eq. (22.12)}$$

$$\sqrt{n} = \frac{z_{\alpha/2}\sigma}{E} \qquad \text{multiplying both sides by } \frac{\sqrt{n}}{E}$$

After squaring both sides, we get

$$n = \left[\frac{z_{\alpha/2}\sigma}{E}\right]^2 \tag{22.13}$$

When the result of Eq. (22.13) is not an integer, it must always be **rounded up** to the next integer in order to guarantee the prescribed margin of error.

EXAMPLE 5 Determining the sample size

An estimate of the mean direct-current output voltage of a certain kind of AC adaptor is desired. If it can be assumed that $\sigma = 0.04$ V, find the sample size necessary to estimate that mean with a margin of error of 0.01 V with 95% confidence.

Substituting $\sigma = 0.04$, $E = 0.01$, and $z_{\alpha/2} = 1.96$, we get

$$n = \left[\frac{1.96(0.04)}{0.01}\right]^2 = 61.5$$

Therefore, a sample of 62 adaptors is necessary. ■

LARGE SAMPLE CONFIDENCE INTERVALS FOR A PROPORTION

We now analyse the problem of estimating an unknown population proportion with confidence. We consider a large population such that a proportion p of its elements share a certain attribute. For example, p could be the proportion of defective items in a lot, or the proportion of incorrect entries in an account, or the proportion of components that will last a certain number of hours, or the proportion of voters who will vote for a certain candidate in the next election.

Suppose that a random sample of size n is taken. We calculate the sample proportion $\hat{p}$ by dividing the number of elements in the sample that share the attribute by the sample size n. We further assume that n is large, so that $np \geq 5$ and $n(1 - p) \geq 5$. (Since p is unknown, we require that $n\hat{p} \geq 5$ and $n(1 - \hat{p}) \geq 5$.)

As we saw in Section 22.3, under these conditions, the sampling distribution of $\hat{p}$ is approximately normal with mean p and standard error $\sqrt{p(1 - p)/n}$. Therefore, if $z_{\alpha/2}$ is the value that leaves an area of $1 - \alpha$ between $-z_{\alpha/2}$ and $z_{\alpha/2}$, then in $(1 - \alpha)100\%$ of samples, $\hat{p}$ satisfies

$$p - z_{\alpha/2}\sqrt{\frac{p(1 - p)}{n}} < \hat{p} < p + z_{\alpha/2}\sqrt{\frac{p(1 - p)}{n}}$$

Manipulating this inequality and eliminating some terms because of the large sample size gives the following general formula for a $(1 - \alpha)100\%$ confidence interval for p.

Large Sample Confidence Interval for a Proportion

Let $\hat{p}$ be the sample proportion obtained from a sample of size n such that $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$. A $(1-\alpha)100\%$ confidence interval for the population proportion p is given by

$$\hat{p} \pm E, \text{ where } E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{22.14}$$

EXAMPLE 6 A 90% confidence interval for *p*

A manufacturer wants to estimate the proportion of defective parts in a large lot produced by a particular machine. In a sample of 350 parts, 41 of them were found to be defective. Construct a 90% confidence interval for the proportion of defective parts in the lot.

We have

$$\hat{p} = \tfrac{41}{350},\ n\hat{p} = 41 \geq 5, \text{ and } n(1-\hat{p}) = 309 \geq 5$$

Therefore, we can apply Eq. (22.14) with $z_{\alpha/2} = 1.645$ (see Table 22.2). The desired 90% confidence interval is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{41}{350} \pm 1.645\sqrt{\frac{\frac{41}{350}\cdot\frac{309}{350}}{350}} = 0.117 \pm 0.028 = (0.089, 0.145)$$

■

As we did with the mean, we can obtain the required sample size for a desired margin of error at a fixed confidence level by using the formula for margin of error from Eq. (22.14) and solving for n. For a confidence level $1-\alpha$, we get

$$n = \hat{p}(1-\hat{p})\left[\frac{z_{\alpha/2}}{E}\right]^2 \tag{22.15}$$

Note that Eq. (22.15) requires an estimate for $\hat{p}$. (It can be obtained from past data or from a pilot study.) When no such estimate is available, we can use the fact that $\hat{p}(1-\hat{p})$ is maximized when $\hat{p} = \frac{1}{2}$, so we use this worst-case scenario estimate in Eq. (22.15). The required sample size for a confidence level $1-\alpha$ becomes

$$n = \frac{1}{4}\left[\frac{z_{\alpha/2}}{E}\right]^2 \tag{22.16}$$

EXAMPLE 7 Determining sample size

Suppose that the manufacturer from Example 6 wishes to estimate the proportion of defectives with a maximum error of 0.025 with 90% confidence.

(a) How large a sample will he need if no information from the past is used?

For this case we use Eq. (22.16) with $E = 0.025$ and $z_{\alpha/2} = 1.645$. The required sample size is $n = \frac{1}{4}\left[\frac{z_{\alpha/2}}{E}\right]^2 = \frac{1}{4}\left[\frac{1.645}{0.025}\right]^2 = 1082.41$, so that 1083 parts must be sampled.

(b) How large a sample will he need if the information from the sample of 350 parts is used? (Recall that $\hat{p} = \frac{41}{350}$.)

If $\hat{p} = \frac{41}{350}$ is known from the past, Eq. (22.15) gives

$$n = \hat{p}(1-\hat{p})\left[\frac{z_{\alpha/2}}{E}\right]^2 = \frac{41}{350}\cdot\frac{309}{350}\left[\frac{1.645}{0.025}\right]^2 = 447.8$$

Therefore, 448 parts must be sampled. Note how information about the possible size of $\hat{p}$ substantially reduced the size of the required sample. ■

EXERCISES 22.4

In Exercises 1–5, make the given changes in the indicated examples of this section and then solve the resulting problems.

1. In Example 2, change $n = 100$ to $n = 140$ and then find the indicated confidence interval.
2. In Example 4, change the confidence level from 99% to 90% and then find the indicated confidence interval.
3. In Example 5, change $\sigma = 0.04$ to $\sigma = 0.03$ and find the required sample size.
4. In Example 6, change the 41 to a 62 and find the indicated confidence interval.
5. In Example 7(a), change 90% to 95% and find the required sample size.

In Exercises 6–8, use the following data. A random sample of size $n = 300$ is taken from a large population with $\sigma = 25.9$. The sample mean is $\bar{x} = 247.1$.

6. Construct a 95% confidence interval for the population mean μ.
7. Construct a 99% confidence interval for the population mean μ.
8. How large a sample must be taken so that a 95% confidence interval for μ will have a maximum margin of error $E = 2.4$?

In Exercises 9–12, use the following data. A random sample of size $n = 215$ is taken from a large population, and 38 are found to be defective.

9. Construct a 95% confidence interval for the population proportion of defectives p.
10. Construct a 99% confidence interval for the population proportion of defectives p.
11. How large a sample must be taken so that a 95% confidence interval for p will have a maximum margin of error of 4.5%? Assume that the information from the sample is used.
12. How large a sample must be taken so that a 95% confidence interval for p will have a maximum margin of error of 4.5%? Assume that no prior information is used.

In Exercises 13–15, use the following information. A random sample of size n is taken from a large population with $\sigma = 3.14$. The sample mean is $\bar{x} = 83.7$.

13. Find a 90% confidence interval if the sample size n is 50.
14. Find a 90% confidence interval if the sample size n is 80. Compare the interval with the interval obtained in Exercise 13. How does increasing the sample size affect the margin of error?
15. Find a 95% confidence interval if the sample size is 80. Compare the interval with the interval obtained in Exercise 14. How does increasing the confidence level affect the margin of error?

In Exercises 16–22, solve the given problems.

16. A sample of 70 washing machines of a certain brand had a mean replacement time of 9.1 years, with a standard deviation of 2.7 years. Find a 95% confidence interval for the mean replacement time of all washing machines of this brand.
17. A test station measured the loudness of a random sample of 45 jets taking off from a certain airport. The mean was found to be 107.2 dB, with a standard deviation of 9.2 dB. Find a 90% confidence interval for the mean loudness of all jets taking off from this airport.
18. An airline wishes to estimate the mean time passengers have to wait for their luggage when arriving at a large airport. How many passengers must be sampled so that a 95% confidence interval for the true mean waiting time μ will have a maximum margin of error of 30 seconds? A similar study done in the past had a standard deviation of 2.16 minutes.
19. A toy manufacturer wishes to estimate the mean time it takes an adult to assemble a certain "easy to assemble" toy. How many adults must be sampled so that a 99% confidence interval for the true mean assembly time μ will have a maximum margin of error of 2.0 minutes? The standard deviation of assembly time for a similar model is known to be 5.9 minutes.
20. From a random sample of 60 bicycle helmets subjected to an impact test, 13 helmets showed some damage from the test. Find a 95% confidence interval for the true proportion of helmets that would show damage from this test.
21. Suppose that we want to estimate the proportion of drivers that exceed the 100 km/h speed limit by more than 10 km in a certain stretch of highway. How large a sample must be taken so that a 95% confidence interval for the true proportion p will have a maximum margin of error of 4%? Assume that no prior information is used.
22. Following are two confidence interval estimates of the true mean contents of certain 306 mL jars of sauce:

    $$(306.2, 307.4) \quad (306.3, 307.3)$$

    The confidence level for one interval is 90%, and the confidence level for the other is 95%, with both intervals constructed from the same sample data. Which of the intervals is the 90% confidence interval? Explain.

Answers to Practice Exercises

1. (96.7, 98.5) with 95% confidence **2.** (96.8, 98.4) with 90% confidence

22.5 Statistical Process Control

Control Charts • Central Line • Range • Upper and Lower Control Limits

One of the most important uses of statistics in industry is statistical process control (SPC), which is used to maintain and improve product quality. Samples are tested during the production at specified intervals to determine whether the production process needs adjustment to meet quality requirements.

■ This is intended only as a brief introduction to this topic. A complete development requires at least a chapter in a statistics book.

A particular industrial process is considered to be *in control* if it is stable and predictable, and sample measurements fall within upper and lower control limits. The process is *out of control* if it has an unpredictable amount of variation and there are sample measurements outside the control limits due to special causes.

EXAMPLE 1 Process—in control—out of control

A manufacturer of 1.5-V batteries states that the voltage of its batteries is no less than 1.45 V or greater than 1.55 V and has designed the manufacturing process to meet these specifications.

■ Minor variations may be expected, for example, from very small fluctuations in voltage, temperature, or material composition. Special causes resulting in an out-of-control process could include line stoppage, material defect, or an incorrect applied pressure.

If all samples of batteries that are tested have voltages in the proper range with only expected minor variations, the production process is *in control*.

However, if some samples have batteries with voltages out of the proper range, the process is *out of control*. This would indicate some special cause for the problem, such as an improperly operating machine or an impurity getting into the process. The process would probably be halted until the cause is determined. ■

CONTROL CHARTS

An important device used in SPC is the *control chart.* It is used to show a trend of a production characteristic over time. In this section we study one type of **control chart for measurements** (when the variable involved is quantitative), and one type of **control chart for attributes** (when the variable involved is qualitative). In both cases, samples are observed at specified intervals of time to see if the sample values are within acceptable limits. The sample values are plotted on a chart to check for trends and abnormalities in the production process.

In making a control chart for measurements, we must determine what the mean should be. For a stable process for which previous data are known, it can be based on a production specification or on previous data. For a new or recently modified process, it may be necessary to use present data, although the value may have to be revised for future charts. On a control chart, *this value is used as the population mean,* μ.

It is also necessary to establish the upper and lower control limits. The standard generally used is that 99.7% of the sample measurements should fall within these control limits. This assumes a normal distribution, and we note that this is within three sample standard deviations of the population mean. We will establish these limits by use of a table or a formula that has been made using statistical measures developed in a more complete coverage of quality control. This does follow the normal practice of using a formula or a more complete table in setting up the control limits.

In Fig. 22.16, we show a sample control chart, and on the following pages, we illustrate how control charts are made.

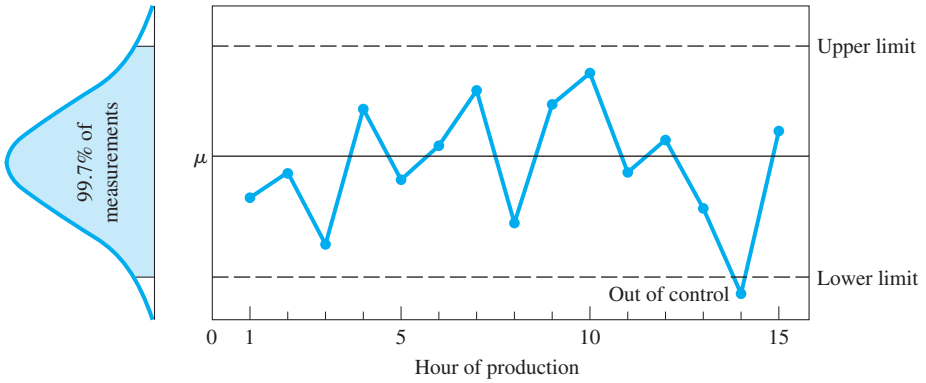

Fig. 22.16

EXAMPLE 2 Making $\bar{x}$ and R control charts

A pharmaceutical company makes a capsule of a prescription drug that contains 500 mg of the drug, according to the label. In a newly modified process of making the capsule, five capsules are tested every 15 min to check the amount of the drug in each capsule. Testing over a 5-h period gave the following results for the 20 subgroups of samples.

Subgroup	*Amount of Drug (in mg) of Five Capsules*					*Mean* $\bar{x}$	*Range* R
1	503	501	498	507	502	502.2	9
2	497	499	500	495	502	498.6	7
3	496	500	507	503	502	501.6	11
4	512	503	488	500	497	500.0	24
5	504	505	500	508	502	503.8	8
6	495	495	501	497	497	497.0	6
7	503	500	507	499	498	501.4	9
8	494	498	497	501	496	497.2	7
9	502	504	505	500	502	502.6	5
10	500	502	500	496	497	499.0	6
11	502	498	510	503	497	502.0	13
12	497	498	496	502	500	498.6	6
13	504	500	495	498	501	499.6	9
14	500	499	498	501	494	498.4	7
15	498	496	502	501	505	500.4	9
16	500	503	504	499	505	502.2	6
17	487	496	499	498	494	494.8	12
18	498	497	497	502	497	498.2	5
19	503	501	500	498	504	501.2	6
20	496	494	503	502	501	499.2	9
					Sum	9998.0	174
					Mean	499.9	8.7

From this table of values, we can make an $\bar{x}$ control chart and an R control chart. The $\bar{x}$ chart maintains a check on the average quality level, whereas the R chart maintains a check on the dispersion of the production process. These two control charts are often plotted together and referred to as the $\bar{x}$–R chart.

In order to define the **central line** of the $\bar{x}$ chart, which ideally is equivalent to the value of the population mean μ, we use the mean of the sample means $\bar{\bar{x}}$. For the central line of the R chart, we use $\bar{R}$. From the table, we see that

$$\bar{\bar{x}} = 499.9 \text{ mg} \quad \text{and} \quad \bar{R} = 8.7 \text{ mg}$$

Table 22.3 Control Chart Factors

n	d_2	A	A_2	D_1	D_2	D_3	D_4
5	2.326	1.342	0.577	0.000	4.918	0.000	2.115
6	2.534	1.225	0.483	0.000	5.078	0.000	2.004
7	2.704	1.134	0.419	0.205	5.203	0.076	1.924

The **upper control limit** (UCL) and the **lower control limit** (LCL) for each chart are defined in terms of the mean range $\bar{R}$ and an appropriate constant taken from a table of control chart factors. These factors, which are related to the sample size n, are determined by statistical considerations found in a more complete coverage of quality control. At the left is a brief table of control chart factors (Table 22.3).

The UCL and LCL for the $\bar{x}$ chart are found as follows:

$$\text{UCL}(\bar{x}) = \bar{\bar{x}} + A_2\bar{R} = 499.9 + 0.577(8.7) = 504.9 \text{ mg} \quad \text{(using Table 22.3 with } n = 5\text{)}$$

$$\text{LCL}(\bar{x}) = \bar{\bar{x}} - A_2\bar{R} = 499.9 - 0.577(8.7) = 494.9 \text{ mg}$$

The UCL and LCL for the R chart are found as follows:

$$\text{LCL}(R) = D_3\overline{R} = 0.000(8.7) = 0.0 \text{ mg}$$
$$\text{UCL}(R) = D_4\overline{R} = 2.115(8.7) = 18.4 \text{ mg}$$

Using these central lines and control limit lines, we now plot the $\bar{x}$ control chart in Fig. 22.17 and the R control chart in Fig. 22.18.

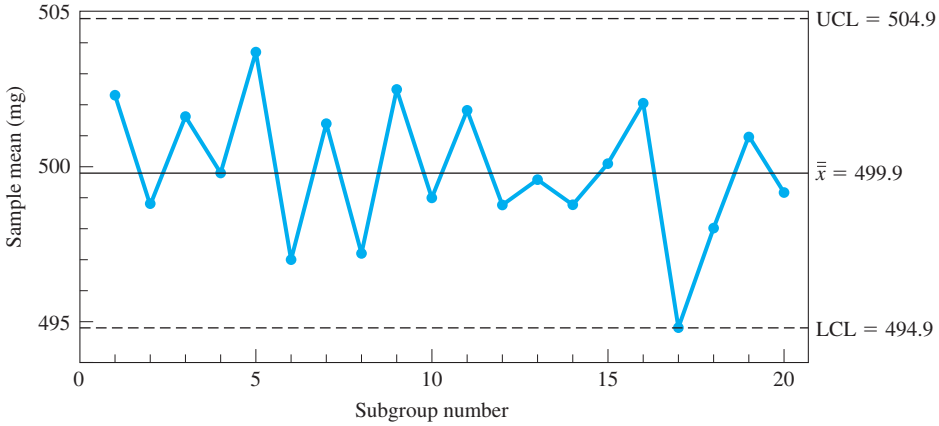

Fig. 22.17

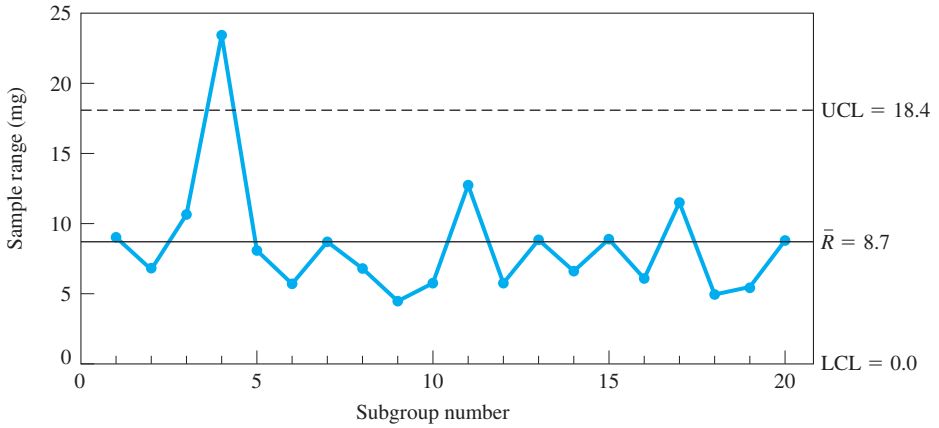

Fig. 22.18

This would be considered a *well-centred process* since $\bar{\bar{x}} = 499.9$ mg, which is very near the target value of 500.0 mg. We do note, however, that subgroup 17 was at a control limit and this might have been due to some special cause, such as the use of a substandard mixture of ingredients. We also note that the process was *out of control* due to some special cause since the range of subgroup 4 was above the upper control limit. We should keep in mind that there are numerous considerations, including human factors, that should be taken into account when making and interpreting control charts and that this is only a very brief introduction to this important industrial use of statistics. ■

We now discuss a control chart for attributes, for the case when each item tested is classified as being either acceptable or not acceptable. To monitor such an attribute in a production process, we obtain the *proportion* of defective parts by *dividing the number of defective parts in a sample by the total number of parts in the sample,* and then make a *p control chart.* This is illustrated in the following example.

Day	*Defective DVDs*	*Proportion Defective*
1	22	0.022
2	16	0.016
3	14	0.014
4	18	0.018
5	12	0.012
6	25	0.025
7	36	0.036
8	16	0.016
9	14	0.014
10	22	0.022
11	20	0.020
12	17	0.017
13	26	0.026
14	20	0.020
15	22	0.022
16	28	0.028
17	17	0.017
18	15	0.015
19	25	0.025
20	12	0.012
21	16	0.016
22	22	0.022
23	19	0.019
24	16	0.016
25	20	0.020
Sum	490	

Practice Exercise

1. In Example 3, change Day 9 datum from 14 to 24 defective parts. Then find UCL(p) and LCL(p).

EXAMPLE 3 Making a *p* control chart

A manufacturer of video discs has 1000 DVDs checked each day for defects (surface scratches, for example). The data for this procedure for 25 days are shown in the table at the left.

The central line for the p control chart is ideally equal to the true proportion of defectives in the population. This value is usually estimated from the data, using the average sample proportion $\bar{p}$, which in this case is

$$\bar{p} = \frac{490}{25\,000} = 0.0196$$

The control limits are each three standard deviations from $\bar{p}$. If n is the size of each sample, we obtain the standard error of $\bar{p}$ using Eq. (22.8) (with $\bar{p}$ in place of p). We get

$$\sigma_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{0.0196(1-0.0196)}{1000}} = 0.004\,38$$

Therefore, the control limits are

$$\text{UCL}(p) = 0.0196 + 3(0.004\,38) = 0.0327$$
$$\text{LCL}(p) = 0.0196 - 3(0.004\,38) = 0.0065$$

Using this central line and these control limit lines, we now plot the p control chart in Fig. 22.19.

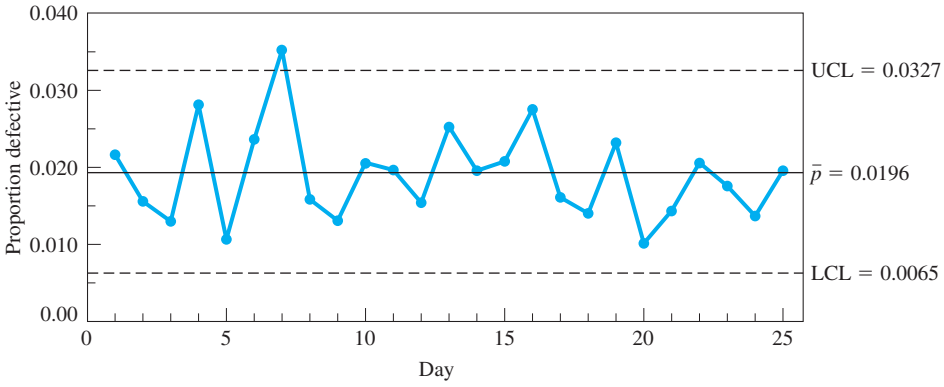

Fig. 22.19

According to the proportion mean of 0.0196, the process produces about 2% defective DVDs. We note that the process was out of control on Day 7. An adjustment to the production process was probably made to remove the special cause of the additional defective DVDs. ■

EXERCISES 22.5

In Exercises 1–4, in Example 2, change the first subgroup to 497, 499, 502, 493, and 498 and then proceed as directed.

1. Find UCL $(\bar{x})$ and LCL $(\bar{x})$.

2. Find LCL (R) and UCL (R).

3. How would the $\bar{x}$ control chart differ from Fig. 22.17?

4. How would the R control chart differ from Fig. 22.18?

In Exercises 5–8, use the following data.

Five automobile engines are taken from the production line each hour and tested for their torque (in N · m) when rotating at a constant frequency. The measurements of the sample torques for 20 h of testing are as follows:

Hour	*Torques (in* N · m*) of Five Engines*				
1	366	352	354	360	362
2	370	374	362	366	356
3	358	357	365	372	361
4	360	368	367	359	363
5	352	356	354	348	350
6	366	361	372	370	363
7	365	366	361	370	362
8	354	363	360	361	364
9	361	358	356	364	364
10	368	366	368	358	360
11	355	360	359	362	353
12	365	364	357	367	370
13	360	364	372	358	365
14	348	360	352	360	354
15	358	364	362	372	361
16	360	361	371	366	346
17	354	359	358	366	366
18	362	366	367	361	357
19	363	373	364	360	358
20	372	362	360	365	367

5. Find the central line, UCL, and LCL for the mean.

6. Find the central line, UCL, and LCL for the range.

7. Plot an $\bar{x}$ chart.

8. Plot an R chart.

In Exercise 9–12, use the following data.

Five AC adaptors that are used to charge batteries of a cellular phone are taken from the production line each 15 minutes and tested for their direct-current output voltage. The output voltages for 24 sample subgroups are as follows:

Subgroup	*Output Voltages of Five Adaptors*				
1	9.03	9.08	8.85	8.92	8.90
2	9.05	8.98	9.20	9.04	9.12
3	8.93	8.96	9.14	9.06	9.00
4	9.16	9.08	9.04	9.07	8.97
5	9.03	9.08	8.93	8.88	8.95
6	8.92	9.07	8.86	8.96	9.04
7	9.00	9.05	8.90	8.94	8.93
8	8.87	8.99	8.96	9.02	9.03
9	8.89	8.92	9.05	9.10	8.93
10	9.01	9.00	9.09	8.96	8.98
11	8.90	8.97	8.92	8.98	9.03
12	9.04	9.06	8.94	8.93	8.92
13	8.94	8.99	8.93	9.05	9.10
14	9.07	9.01	9.05	8.96	9.02
15	9.01	8.82	8.95	8.99	9.04
16	8.93	8.91	9.04	9.05	8.90
17	9.08	9.03	8.91	8.92	8.96
18	8.94	8.90	9.05	8.93	9.01
19	8.88	8.82	8.89	8.94	8.88
20	9.04	9.00	8.98	8.93	9.05
21	9.00	9.03	8.94	8.92	9.05
22	8.95	8.95	8.91	8.90	9.03
23	9.12	9.04	9.01	8.94	9.02
24	8.94	8.99	8.93	9.05	9.07

9. Find the central line, UCL, and LCL for the mean.

10. Find the central line, UCL, and LCL for the range.

11. Plot an $\bar{x}$ chart.

12. Plot an R chart.

In Exercises 13–16, use the following information.

For a production process for which there is a great deal of data since its last modification, the population mean μ and population standard deviation σ are assumed known. For such a process, we have the following values (using additional statistical analysis):

$\bar{x}$ chart: central line $= \mu$, UCL $= \mu + A\sigma$, LCL $= \mu - A\sigma$

R chart: central line $= d_2\sigma$, UCL $= D_2\sigma$, LCL $= D_1\sigma$

The values of A, d_2, D_2, and D_1 are found in the table of control chart factors in Example 2 (Table 22.3).

13. In the production of robot links and tests for their lengths, it has been found that $\mu = 2.725$ cm and $\sigma = 0.032$ cm. Find the central line, UCL, and LCL for the mean if the sample subgroup size is 5.

14. For the robot link samples of Exercise 13, find the central line, UCL, and LCL for the range.

15. After bottling, the volume of soft drink in six sample bottles is checked each 10 minutes. For this process $\mu = 750.0$ mL and $\sigma = 2.2$ mL. Find the central line, UCL, and LCL for the range.

16. For the bottling process of Exercise 15, find the central line, UCL, and LCL for the mean.

In Exercises 17 and 18, use the following data.

A telephone company rechecks the entries for 1000 of its new customers each week for name, address, and phone number. The data collected regarding the number of new accounts with errors, along with the proportion of these accounts with errors, is given in the following table for a 20-wk period:

Week	*Accounts with Errors*	*Proportion with Errors*
1	52	0.052
2	36	0.036
3	27	0.027
4	58	0.058
5	44	0.044
6	21	0.021
7	48	0.048
8	63	0.063
9	32	0.032
10	38	0.038
11	27	0.027
12	43	0.043
13	22	0.022
14	35	0.035
15	41	0.041
16	20	0.020
17	28	0.028
18	37	0.037
19	24	0.024
20	42	0.042
Total	738	

17. For a p chart, find the values for the central line, UCL, and LCL.

18. Plot a p chart.

In Exercises 19 and 20, use the following data.

A maker of electric fuses checks 500 fuses each day for defects. The number of defective fuses, along with the proportion of defective fuses for 24 days, is shown in the following table.

Day	*Number Defective*	*Proportion Defective*
1	26	0.052
2	32	0.064
3	37	0.074
4	16	0.032
5	28	0.056
6	31	0.062
7	42	0.084
8	22	0.044
9	31	0.062
10	28	0.056
11	24	0.048
12	35	0.070
13	30	0.060
14	34	0.068
15	39	0.078
16	26	0.052
17	23	0.046
18	33	0.066
19	25	0.050
20	25	0.050
21	32	0.064
22	23	0.046
23	34	0.068
24	20	0.040
Total	696	

19. For a p chart, find the values for the central line, UCL, and LCL.

20. Plot a p chart.

Answers to Practice Exercise

1. $\mathrm{UCL}(p) = 0.0333, \mathrm{LCL}(p) = 0.0067$

22.6 Linear Regression

Regression • Linear Regression • Method of Least Squares • Deviation • Least-Squares Line

We have considered statistical methods for dealing with one variable. We now discuss how to find an equation relating two variables for which a set of points is known.

In this section, we show a method of finding the equation of a straight line that passes through a set of data points, and in this way we *fit* the line to the points. In general, *the fitting of a curve to a set of points is called* **regression.** Fitting a straight line to a set of points is *linear regression,* and fitting some other type of curve is called *nonlinear regression.* We consider nonlinear regression in the next section.

Some of the reasons for using regression to find the equation of a curve that passes through a set of points, and thereby "fit" the curve to the points, are **(1)** to express a concise relationship between the variables, **(2)** to use the equation to predict certain fundamental results, **(3)** to determine the reliability of certain sets of data, and **(4)** to use the data for testing certain theoretical concepts.

For a given set of several (at least 5 or 6) points for representing pairs of data values, we cannot reasonably expect that the curve of any given equation will pass through all of the points *exactly.* Therefore, when we fit the curve of an equation to the points, we

are finding the curve that best approximates passing through the points. It is possible that the curve that best fits the data will not actually pass directly through any of the points, although it should come reasonably close to most of them. Consider the following example.

EXAMPLE 1 Fitting a line to a set of points

All the students enrolled in a mathematics course took an entrance test. To study the reliability of this test as an indicator of future success, an instructor tabulated the test scores of ten students (selected at random), along with their course averages at the end of the course, and made a graph of the data. See the table below and Fig. 22.20.

Student	*Entrance Test Score, Based on 40*	*Course Average, Based on 100*
A	29	63
B	33	88
C	22	77
D	17	67
E	26	70
F	37	93
G	30	72
H	32	81
I	23	47
J	30	74

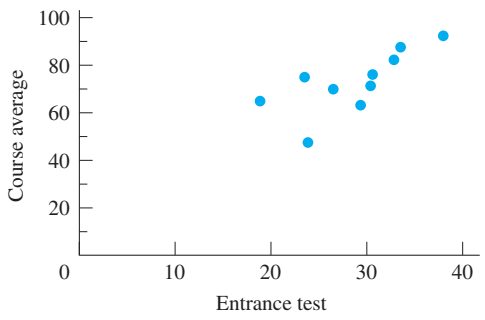

Fig. 22.20

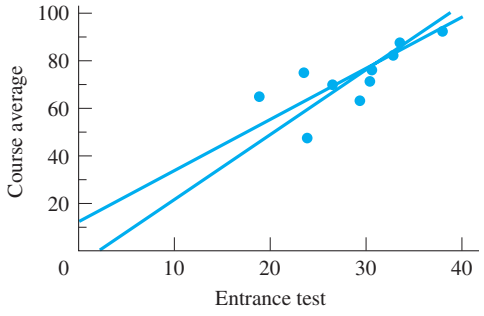

Fig. 22.21

We now ask whether there is a functional relationship between the test scores and the course grades. Certainly, no clear-cut relationship exists, but in general we see that the higher the test score, the higher the course grade. This leads to the possibility that there might be some straight line, from which none of the points would vary too significantly. If such a line could be found, then it could be the basis of predictions as to the possible success a student might have in the course, on the basis of his or her grade on the entrance test. Assuming that such a straight line exists, the problem is to find the equation of this line. Fig. 22.21 shows two such possible lines. ■

There are a number of different methods of determining the straight line that best fits the given data points. We employ the method that is most widely used: the **method of least squares.** *The basic principle of this method is that the sum of the squares of the deviations of all data points from the best line (in accordance with this method) has the least value possible. By* **deviation,** *we mean the difference between the y-value of the line and the y-value for the point (of original data) for a particular value of x.*

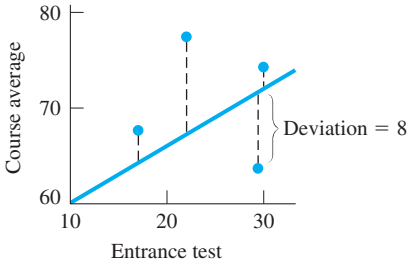

Fig. 22.22

EXAMPLE 2 Deviation

In Fig. 22.22, the deviations of some of the points of Example 1 are shown. The point $(29, 63)$ (student A of Example 1) has a deviation of 8 from the indicated line in the figure. Thus, we square the value of this deviation to obtain 64. In order to find the equation of the straight line that best fits the given points, the method of least squares requires that the sum of all such squares be a minimum. ■

In applying this method of least squares, it is necessary to use the equation of a straight line and the coordinates of the points of the data. The deviations of all of these data points are determined, and these values are then squared. It is then necessary to determine the constants for the slope m and the y-intercept b in the equation of a straight line $y = mx + b$ for which the sum of the squared values is a minimum. To do this requires certain methods of advanced mathematics. Using those methods, it is shown that *the equation of the* **least-squares line**

$$y = mx + b \tag{22.17}$$

can be found by calculating the values of the slope m and the y-intercept b by using the formulas

$$m = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} \tag{22.18}$$

and

$$b = \frac{\left(\sum x^2\right)\left(\sum y\right) - \left(\sum xy\right)\left(\sum x\right)}{n\sum x^2 - \left(\sum x\right)^2} \tag{22.19}$$

In Eqs. (22.18) and (22.19), ***the x's and y's are the values of the coordinates of the points in the given data,*** and n is the number of points of data. We can reduce the calculational work in finding the values of m and b by noting that the denominators in Eqs. (22.18) and (22.19) are the same. Therefore, in using a calculator, the value of this denominator can be stored in memory.

COMMON ERROR

It is a common error to confuse $\sum x^2$ and $(\sum x)^2$ in the denominators of Eqs. (22.18) and (22.19). Note that for $\sum x^2$, we square the x values and then add the squares, whereas for $(\sum x)^2$, we first add the x values and then square the sum.

EXAMPLE 3 Finding the equation of the least-squares line

Find the equation of the least-squares line for the points indicated in the following table. Graph the line and data points on the same graph.

x	1	2	3	4	5
y	3	6	6	8	12

We see from Eqs. (22.18) and (22.19) that we need the sums of x, y, xy, and x^2 in order to find m and b. Thus, we set up a table for these values, along with the necessary calculations, as follows:

x	y	xy	x^2
1	3	3	1
2	6	12	4
3	6	18	9
4	8	32	16
5	12	60	25
sums → 15	35	125	55
	$\sum x$	$\sum y$	$\sum xy$ $\sum x^2$

$$n = 5 \quad (5 \text{ points})$$

$$m = \frac{5(125) - (15)(35)}{5(55) - (15)^2} = \frac{100}{50} = 2$$

$$b = \frac{(55)(35) - (125)(15)}{50} = \frac{50}{50} = 1$$

This means that the equation of the least-squares line is $y = 2x + 1$. This line and the data points are shown in Fig. 22.23. ■

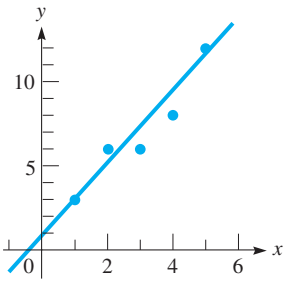

Fig. 22.23

EXAMPLE 4 Finding the equation of the least-squares line

Find the least-squares line for the data of Example 1.

Here, the x-values will be the entrance-test scores and the y-values are the course averages.

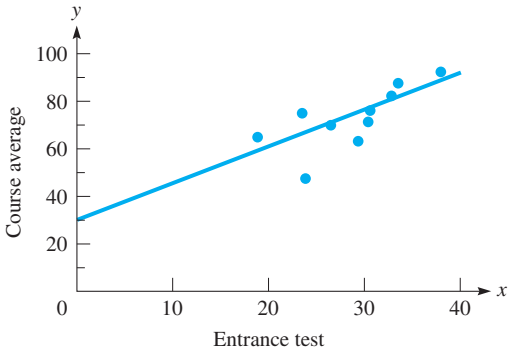

Fig. 22.24

x	y	xy	x^2
29	63	1827	841
33	88	2904	1089
22	77	1694	484
17	67	1139	289
26	70	1820	676
37	93	3441	1369
30	72	2160	900
32	81	2592	1024
23	47	1081	529
30	74	2220	900
279	732	20 878	8101

$$n = 10$$

$$m = \frac{10(20\,878) - 279(732)}{10(8101) - 279^2} = 1.44$$

$$b = \frac{8101(732) - 20\,878(279)}{10(8101) - 279^2} = 33.1$$

Thus, the equation of the least-squares line is $y = 1.44x + 33.1$. The line and data points are shown in Fig. 22.24. This line best fits the data, although the fit is obviously approximate. It can be used to predict the approximate course average that a student might be expected to attain, based on the entrance test. ■

■ A graphing calculator, a spreadsheet, or computer software can be used to find the slope and the intercept and to display the least-squares line.

EXAMPLE 5 Least-squares line—application

In a research project to determine the amount of a drug that remains in the bloodstream after a given dosage, the amounts y (in mg of drug/dL of blood) were recorded after t hours, as shown in the following table. Find the least-squares line for these data, expressing y as a function of t. Sketch the graph of the line and data points.

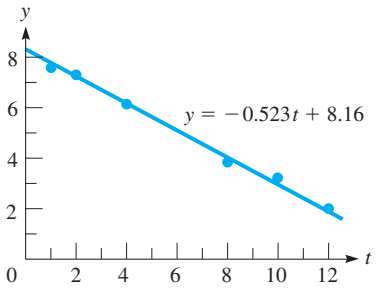

Fig. 22.25

The calculations are as follows:

t (h)	1.0	2.0	4.0	8.0	10.0	12.0
y (mg/dL)	7.6	7.2	6.1	3.8	2.9	2.0

t	y	ty	t^2
1.0	7.6	7.6	1.0
2.0	7.2	14.4	4.0
4.0	6.1	24.4	16.0
8.0	3.8	30.4	64.0
10.0	2.9	29.0	100
12.0	2.0	24.0	144
37.0	29.6	129.8	329

$$n = 6$$

$$m = \frac{6(129.8) - 37.0(29.6)}{6(329) - 37.0^2} = -0.523$$

$$b = \frac{(329)(29.6) - (129.8)(37.0)}{6(329) - 37.0^2} = 8.16$$

The equation of the least-squares line is $y = -0.523t + 8.16$. The line and the data points are shown in Fig. 22.25. This line is useful in determining the effectiveness of the drug. It can also be used to determine when additional medication may be administered. ■

EXERCISES 22.6

In Exercises 1–14, find the equation of the least-squares line for the given data. Graph the line and data points on the same graph.

1. In Example 3, replace the y-values with 3, 7, 9, 9, and 12. Then follow the instructions above.

2.

x	1	2	3	4	5	6	7
y	10	17	28	37	49	56	72

3.

x	20	26	30	38	48	60
y	160	145	135	120	100	90

4.

x	1	3	6	5	8	10	4	7	3	8
y	15	12	10	8	9	2	11	9	11	7

5. In Example 5, change the y (mg of drug/dL of blood) values to 8.7, 8.4, 7.7, 7.3, 5.7, and 5.2. Then proceed to find y as a function of t, as in Example 5.

6. The velocity v (in m/s) of a falling object was found each second by use of an electronic device, as shown in the following table. Find v as a function of t.

t (s)	1.00	2.00	3.00	4.00	5.00	6.00	7.00
v (m/s)	9.70	19.5	29.5	39.4	49.2	58.9	68.6

7. In an electrical experiment, the following data were found for the values of current and voltage for a particular element of the circuit. Find the voltage V as a function of the current i.

Current (mA)	15.0	10.8	9.30	3.55	4.60
Voltage (V)	3.00	4.10	5.60	8.00	10.50

8. A particular muscle was tested for its speed of shortening as a function of the force applied to it. The results appear below. Find the speed as a function of the force.

Force (N)	60.0	44.2	37.3	24.2	19.5
Speed (m/s)	1.25	1.67	1.96	2.56	3.05

9. The altitude h (in m) of a rocket was measured at several positions at a horizontal distance x (in m) from the launch site, shown in the table. Find the least-squares line for h as a function of x.

x (m)	0	500	1000	1500	2000	2500
h (m)	0	1130	2250	3360	4500	5600

10. In testing an air-conditioning system, the temperature T in a building was measured during the afternoon hours with the results shown in the table. Find the least-squares line for T as a function of the time t from noon.

t (h)	0.0	1.0	2.0	3.0	4.0	5.0
T (°C)	20.5	20.6	20.9	21.3	21.7	22.0

11. The pressure p was measured along an oil pipeline at different distances from a reference point, with results as shown. Find the least-squares line for p as a function of x.

x (m)	0	50	100	150	200
p (kPa)	4370	4240	4070	3970	3840

12. The heat loss L per hour through various thicknesses of a particular type of insulation was measured as shown in the table. Find the least-squares line for L as a function of t.

t (m)	3.0	4.0	5.0	6.0	7.0
L (MJ)	5.90	4.80	3.90	3.10	2.45

13. In an experiment on the photoelectric effect, the frequency of light being used was measured as a function of the stopping potential (the voltage just sufficient to stop the photoelectric effect) with the results given below. Find the least-squares line for V as a function of f. The frequency for $V = 0$ is known as the *threshold frequency*. From the graph determine the threshold frequency.

f (PHz)	0.550	0.605	0.660	0.735	0.805	0.880
V (V)	0.350	0.600	0.850	1.10	1.45	1.80

14. If gas is cooled under conditions of constant volume, it is noted that the pressure falls nearly proportionally as the temperature. If this were to happen until there was no pressure, the theoretical temperature for this case is referred to as *absolute zero*. In an elementary experiment, the following data were found for pressure and temperature under constant volume.

T (°C)	0.0	20	40	60	80	100
p (kPa)	133	143	153	162	172	183

Find the least-squares line for p as a function of T, and from the graph determine the value of absolute zero found in this experiment.

The linear coefficient of correlation, a measure of the strength of the linear relationship of two variables, is defined by $r = m(s_x/s_y)$, *where* s_x *and* s_y *are the standard deviations of the x-values and y-values, respectively. Due to its definition, the values of r lie in the range* $-1 \le r \le 1$. *If r is near 1, the correlation is considered good. For the values of r between* -0.5 *and* $+0.5$, *the correlation is poor. If r is near* -1, *the variables are said to be negatively correlated; that is, one increases as the other decreases. In Exercises 15–18, compute r for the given data.*

15. Exercise 1

16. Exercise 2

17. Exercise 4

18. Example 1

22.7 Nonlinear Regression

Nonlinear Regression • Least-Squares Curve • Types of Curves on Calculator

If the experimental points do not appear to be on a straight line, but we recognize them as being approximately on some other type of curve, then nonlinear regression must be used to fit a curve to the data points. For example, if the points are apparently on a parabola, we would want to fit a quadratic equation instead of a line.

Often, the method of linear least squares can be adapted to fit curves. The method is to create new variables from the available data. By using the appropriate transformation on the data, the curved function can be written as a linear function of the new variables. In particular, we consider nonlinear transformations $f(x)$ of the x variable, and extend the least-squares line to

$$y = m[f(x)] + b \tag{22.20}$$

Here, $f(x)$ must be calculated first, and then the problem can be treated as a least-squares line to find the values of m and b. Some of the functions $f(x)$ that may be considered for use are x^2, $1/x$, 10^x, and $\ln x$.

EXAMPLE 1 Fitting $y = mx^2 + b$ to a set of points

Find the least-squares curve $y = mx^2 + b$ for the following points:

x	0	1	2	3	4	5
y	1	5	12	24	53	76

In using Eq. (22.12), $f(x) = x^2$. Our first step is to calculate values of x^2, and then we use x^2 as we used x in finding the equation of the least-squares line.

x	$f(x) = x^2$	y	x^2y	$(x^2)^2$
0	0	1	0	0
1	1	5	5	1
2	4	12	48	16
3	9	24	216	81
4	16	53	848	256
5	25	76	1900	625
	55	171	3017	979

$$n = 6$$

$$m = \frac{6(3017) - 55(171)}{6(979) - 55^2} = 3.05$$

$$b = \frac{(979)(171) - (3017)(55)}{6(979) - 55^2} = 0.52$$

Therefore, the required equation is $y = 3.05x^2 + 0.52$. The graph of this equation and the data points are shown in Fig. 22.26. ■

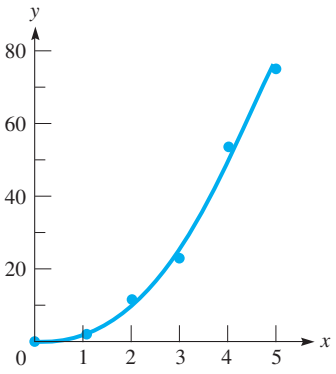

Fig. 22.26

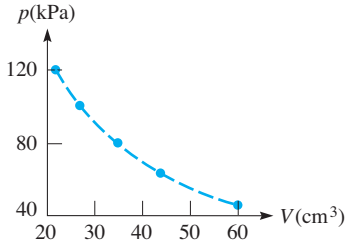

Fig. 22.27

EXAMPLE 2 Fitting $y = m(1/x) + b$ to a set of points

In a physics experiment, the pressure p and volume V of a gas were measured at constant temperature. When the points were plotted, they were seen to approximate the hyperbola $y = c/x$. Find the least-squares approximation to the hyperbola $y = m(1/x) + b$ for the given data. See Fig. 22.27.

p (kPa)	V (cm^3)
120.0	21.0
99.2	25.0
81.3	31.8
60.6	41.1
42.7	60.1

$x\ (=V)$	$f(x) = \frac{1}{x}$	$y\ (=p)$	$(\frac{1}{x})y$	$(\frac{1}{x})^2$
21.0	0.047 619 0	120.0	5.714 285 7	0.002 267 6
25.0	0.040 000 0	99.2	3.968 000 0	0.001 600 0
31.8	0.031 446 5	81.3	2.556 603 8	0.000 988 9
41.1	0.024 330 9	60.6	1.474 452 6	0.000 592 0
60.1	0.016 638 9	42.7	0.710 482 5	0.000 276 9
	0.160 035 3	403.8	14.423 824 6	0.005 725 4

(*Calculator note:* The final digits for the values shown may vary depending on the calculator and how the values are used. Here, all individual values are shown with 8 digits (rounded off), although more digits were used. The value of $1/x$ was found from the value of x, with the 8 digits shown. However, the values of $(1/x)y$ and $(1/x)^2$ were found from the value of $1/x$, using the extra digits. The sums were found using the rounded-off values shown. However, since the data contain only 3 digits, any variation in the final digits for $1/x$, $(1/x)y$, or $(1/x)^2$, will not matter.)

$$m = \frac{5(14.423\,824\,6) - 0.160\,035\,3(403.8)}{5(0.005\,725\,4) - 0.160\,035\,3^2} = 2490$$

$$b = \frac{(0.005\,725\,4)(403.8) - (14.423\,824\,6)(0.160\,035\,3)}{5(0.005\,725\,4) - 0.160\,035\,3^2} = 1.2$$

The equation of the hyperbola $y = m(1/x) + b$ is

$$y = \frac{2490}{x} + 1.2$$

This hyperbola and data points are shown in Fig. 22.28. ■

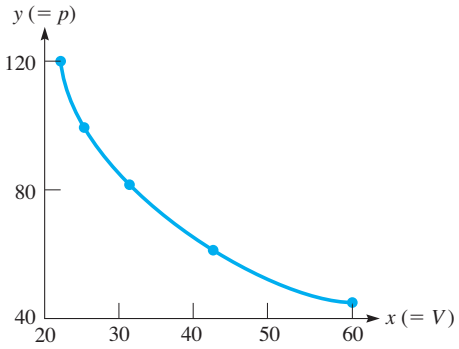

Fig. 22.28

EXAMPLE 3 Fitting $y = m(10^x) + b$ to a set of points

It has been found experimentally that the tensile strength of brass (a copper-zinc alloy) increases (within certain limits) with the percent of zinc. The following table shows the values that have been found. See Fig. 22.29.

Tensile Strength (GPa)	0.32	0.36	0.40	0.44	0.48
Percent of Zinc	0	5	13	22	34

Fit a curve of the form $y = m(10^x) + b$ to the data. Let x = tensile strength and y = percent of zinc.

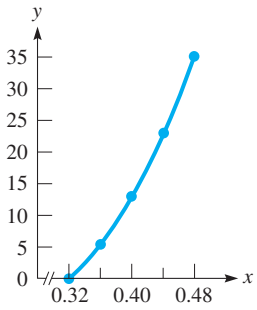

Fig. 22.29

x	$f(x) = 10^x$	y	$(10^x)y$	$(10^x)^2$
0.32	2.089 296 1	0	0.000 000	4.365 158 3
0.36	2.290 867 7	5	11.454 338	5.248 074 6
0.40	2.511 886 4	13	32.654 524	6.309 573 4
0.44	2.754 228 7	22	60.593 031	7.585 775 8
0.48	3.019 951 7	34	102.678 36	9.120 108 4
	12.666 230 6	74	207.380 25	32.628 690 5

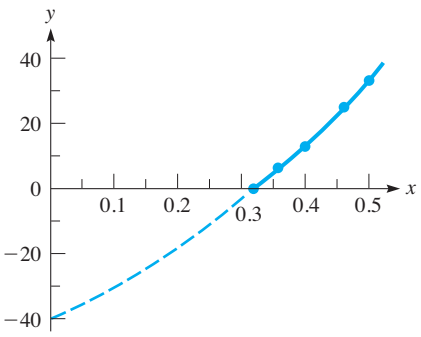

Fig. 22.30

(See the note on calculator use in Example 2.)

$$m = \frac{5(207.380\,25) - 12.666\,230\,6(74)}{5(32.628\,690\,5) - 12.666\,230\,6^2} = 36.8$$

$$b = \frac{32.628\,690\,5(74) - 207.380\,25(12.666\,230\,6)}{5(32.628\,690\,5) - 12.666\,230\,6^2} = -78.3$$

The equation of the curve is $y = 36.8(10^x) - 78.3$. It must be remembered that for practical purposes, y must be positive. The graph of the equation is shown in Fig. 22.30, with the solid portion denoting the meaningful part of the curve. The points of the data are also shown. ■

As we noted in the previous section, a graphing calculator, a spreadsheet, or computer software can be used to determine the equation of the regression curve, and to display its graph. The different models that can be considered include

Linear:	$y = ax + b$
Quadratic:	$y = ax^2 + bx + c$
Cubic:	$y = ax^3 + bx^2 + cx + d$
Quartic:	$y = ax^4 + bx^3 + cx^2 + dx + e$
Logarithmic:	$y = a + b \ln x$
Exponential:	$y = ab^x$
Power:	$y = ax^b$
Logistic:	$y = \dfrac{c}{1 - ae^{-bx}}$
Sinusoidal:	$y = a \sin(bx + c) + d$

EXERCISES 22.7

In the following exercises, find the equation of the indicated least-squares curve. Sketch the curve and plot the data points on the same graph.

1. In Example 1, replace the y-values with 2, 3, 10, 25, 44, and 65. Then follow the instructions above.

2. For the points in the following table, find the least-squares curve $y = m\sqrt{x} + b$.

x	0	4	8	12	16
y	1	9	11	14	15

3. In Example 2, change the V (volume of the gas) values to 19.9, 24.5, 29.4, 39.4, and 56.0. Then find $y\ (= p)$ as a function of $x\ (= V)$, as in Example 2.

4. In Example 3, change the y (percent of zinc) values to 2, 8, 15, 23, and 32. Then find y as a function of x (tensile strength), as in Example 3.

5. The following data were found for the distance y that an object rolled down an inclined plane in time t. Determine the least-squares curve $y = mt^2 + b$. Compare the equation with that using the *quadratic regression* feature on a graphing calculator.

t (s)	1.0	2.0	3.0	4.0	5.0
y (cm)	6.0	23	55	98	148

6. The increase in length y of a certain metallic rod was measured in relation to particular increases x in temperature. Find the least-squares curve $y = mx^2 + b$.

x (°C)	50.0	100	150	200	250
y (cm)	1.00	4.40	9.40	16.4	24.0

7. The pressure p at which Freon, a refrigerant, vapourizes for temperature T is given in the following table. Find the least-squares curve $p = mT^2 + b$.

T (°C)	0	10	20	30	40
p (kPa)	480	600	830	1040	1400

8. A fraction f of annual hot-water loads at a certain facility are heated by solar energy. The fractions f for certain values of the collector area A are given in the following table. Find the least-squares curve $f = m\sqrt{A} + b$.

A (m^2)	0	12	27	56	90
f	0.0	0.2	0.4	0.6	0.8

9. The makers of a special blend of coffee found that the demand for the coffee depended on the price charged. The price P per pound and the monthly sales S are shown in the following table. Find the least-squares curve $P = m(1/S) + b$.

S (thousands)	240	305	420	480	560
P (dollars)	5.60	4.40	3.20	2.80	2.40

10. The resonant frequency f of an electric circuit containing a 4-μF capacitor was measured as a function of an inductance L in the circuit. The following data were found. Find the least-squares curve $f = m(1/\sqrt{L}) + b$.

L (H)	1.0	2.0	4.0	6.0	9.0
f (Hz)	490	360	250	200	170

11. The displacement y of an object at the end of a spring at given times t is shown in the following table. Find the least-squares curve $y = me^{-t} + b$.

t (s)	0.0	0.5	1.0	1.5	2.0	3.0
y (cm)	6.1	3.8	2.3	1.3	0.7	0.3

12. The average daily temperatures T (in °C) for each month in Montreal (Environment Canada Archives) are given in the following table:

t	J	F	M	A	M	J	J	A	S	O	N	D
T(°C)	−9	−7	−1	7	14	19	22	21	16	9	2	−6

Find the least-squares curve $T = m\cos[\frac{\pi}{6}(t - 0.5)] + b$. Assume the average temperature is for the 15th of each month. Then the values of t (in months) are 0.5, 1.5, . . . , 11.5.

CHAPTER 22 EQUATIONS

Arithmetic mean
$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} \tag{22.1}$$

Standard deviation
$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \tag{22.2}$$

$$s = \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n(n - 1)}} \tag{22.3}$$

Normal distribution
$$y = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \tag{22.4}$$

Standard normal distribution
$$y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{22.5}$$

Standard (z) score
$$z = \frac{x - \mu}{\sigma} \tag{22.6}$$

Standard error of $\bar{x}$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{22.7}$$

Standard error of $\hat{p}$
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}} \tag{22.8}$$

$(1 - \alpha)100\%$ confidence interval for μ
$$\bar{x} \pm E, \text{ where } E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad \text{if } \sigma \text{ is known}$$
$$E = z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \quad \text{if } \sigma \text{ is unknown} \tag{22.12}$$

Sample size required (μ)
$$n = \left[\frac{z_{\alpha/2}\sigma}{E}\right]^2 \tag{22.13}$$

$(1 - \alpha)100\%$ confidence interval for p
$$\hat{p} \pm E, \text{ where } E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \tag{22.14}$$

Sample sized required (p)

An estimate $\hat{p}$ is available

$$n = \hat{p}(1 - \hat{p})\left[\frac{z_{\alpha/2}}{E}\right]^2 \tag{22.15}$$

No estimate $\hat{p}$ is available

$$n = \frac{1}{4}\left[\frac{z_{\alpha/2}}{E}\right]^2 \tag{22.16}$$

Least-squares lines

$$y = mx + b \tag{22.17}$$

$$m = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} \tag{22.18}$$

$$b = \frac{\left(\sum x^2\right)\left(\sum y\right) - \left(\sum xy\right)\left(\sum x\right)}{n\sum x^2 - \left(\sum x\right)^2} \tag{22.19}$$

Nonlinear curves

$$y = m[f(x)] + b \tag{22.20}$$

CHAPTER 22 REVIEW EXERCISES

In Exercises 1–10, use the following data. An airline's records showed that the percent of on-time flights each day for a 20-day period was as follows:

72, 75, 76, 70, 77, 73, 80, 75, 82, 85,
77, 78, 74, 86, 72, 77, 67, 78, 69, 80

1. Determine the median.

2. Determine the mode.

3. Determine the mean.

4. Determine the standard deviation.

5. Construct a frequency distribution table with five classes and a lowest class limit of 67.

6. Draw a frequency polygon for the data in Exercise 5.

7. Draw a histogram for the data in Exercise 5.

8. Construct a relative frequency table for the data in Exercise 5.

9. Construct a cumulative frequency table for the data of Exercise 5.

10. Draw an ogive for the data of Exercise 5.

In Exercises 11–16, use the following data: An important property of oil is its coefficient of viscosity, which gives a measure of how well it flows. In order to determine the viscosity of a certain motor oil, a refinery took samples from 12 different storage tanks and tested them at 50°C. *The results (in pascal-seconds) were 0.24, 0.28, 0.29, 0.26, 0.27, 0.26, 0.25, 0.27, 0.28, 0.26, 0.26, 0.25.*

11. Find the mean.

12. Find the median.

13. Find the standard deviation.

14. Draw a histogram.

15. Draw a frequency polygon.

16. Determine the range.

In Exercises 17–26, use the following data: A sample of wind generators was tested for power output when the wind speed was 30 km/h. *The following table gives the class marks of the powers produced and the number of generators in each class.*

Power (W)	650	660	670	680	690
No. Generators	3	2	7	12	27

Power (W)	700	710	720	730
No. Generators	34	15	16	5

17. Find the median.

18. Find the mean.

19. Find the mode.

20. Draw a histogram.

21. Find the standard deviation.

22. Draw a frequency polygon.

23. Make a cumulative frequency table.

24. Draw an ogive.

25. The mean and the standard deviation obtained from the raw data are $\bar{x} = 696$ and $s = 17.7$. Find a 95% confidence interval for the true mean power output produced by all generators of this kind when wind is 30 km/h.

26. How many more observations should be taken so that a 95% confidence interval for the true mean power output will have a maximum margin of error $E = 2.5$ W?

In Exercises 27–30, use the following data: A Geiger counter records the presence of high-energy nuclear particles. Even though no apparent radioactive source is present, some particles will be recorded. These are primarily cosmic rays, which are caused by very high-energy

particles from outer space. In an experiment to measure the amount of cosmic radiation, the number of counts were recorded during 200 5-s intervals. The following table gives the number of counts and the number of 5-s intervals having this number of counts. Draw a frequency curve for these data.

Counts	0	1	2	3	4	5	6	7	8	9	10
Intervals	3	10	25	45	29	39	26	11	7	2	3

27. Find the median.

28. Find the mean.

29. Draw a histogram.

30. Make a relative frequency table.

In Exercises 31–36, use the following data: Police radar on a city street recorded the speeds of 110 cars in a 65 km/h *zone. The following table shows the class marks of the speeds recorded and the number of cars in each class.*

Speed (km/h)	40	45	50	55	60	65	70	75	80	85
No. cars	3	4	4	5	8	22	48	10	4	2

31. Find the mean.

32. Find the median

33. Find the standard deviation.

34. Draw an ogive.

35. The mean and the standard deviation obtained from the raw data are $\bar{x} = 66$ and $s = 9.1$. Find a 90% confidence interval for the true mean speed in this zone.

36. How many more observations should be taken so that a 90% confidence interval for the true mean speed will have a maximum margin of error $E = 1.0$ km/h?

In Exercises 37–38, solve the given problems.

37. Use Chebychev's theorem to find the percentage of values that are between 27.8 and 36.2 in a data set with mean 32 and standard deviation 2.1.

38. Use Chebychev's theorem to find the percentage of values that are between 174.2 and 189.8 in a data set with mean 182 and standard deviation 2.6.

In Exercises 39–42, use the following data. A random sample of size $n = 185$ is taken from a large population, and 36 are found to be defective.

39. Construct a 95% confidence interval for the population proportion of defectives p.

40. Construct a 90% confidence interval for the population proportion of defectives p.

41. How large a sample must be taken so that a 90% confidence interval for p will have a maximum margin of error of 3.2%? Assume that the information from the sample is used.

42. How large a sample must be taken so that a 90% confidence interval for p will have a maximum margin of error of 3.2%? Assume that no prior information is used.

In Exercises 43 and 44, use the following information: A company that makes electric light bulbs tests 500 bulbs each day for defects. The number of defective bulbs, along with the proportion of defective bulbs for 20 days, is shown in the following table.

Day	*Number Defective*	*Proportion Defective*
1	23	0.046
2	31	0.062
3	19	0.038
4	27	0.054
5	29	0.058
6	39	0.078
7	26	0.052
8	17	0.034
9	28	0.056
10	33	0.066
11	22	0.044
12	29	0.058
13	20	0.040
14	35	0.070
15	21	0.042
16	32	0.064
17	25	0.050
18	23	0.046
19	29	0.058
20	32	0.064
Total	540	

43. For a p chart, find the values of the central line, UCL, and LCL.

44. Plot a p chart.

In Exercises 45 and 46, use the following information: Five ball bearings are taken from the production line every 15 min and their diameters are measured. The diameters of the sample ball bearings for 16 successive subgroups are given in the following table.

Subgroup	*Diameters* (mm) *of Five Ball Bearings*				
1	4.98	4.92	5.02	4.91	4.93
2	5.03	5.01	4.94	5.06	5.07
3	5.05	5.03	5.00	5.02	4.96
4	5.01	4.92	4.91	4.99	5.03
5	4.92	4.97	5.02	4.95	4.94
6	5.02	4.95	5.01	5.07	5.15
7	4.93	5.03	5.02	4.96	4.99
8	4.85	4.91	4.88	4.92	4.90
9	5.02	4.95	5.06	5.04	5.06
10	4.98	4.98	4.93	5.01	5.00
11	4.90	4.97	4.93	5.05	5.02
12	5.03	5.05	4.92	5.03	4.98
13	4.90	4.96	5.00	5.02	4.97
14	5.09	5.04	5.05	5.02	4.97
15	4.88	5.00	5.02	4.97	4.94
16	5.02	5.09	5.03	4.99	5.03

45. Plot an $\bar{x}$ chart.

46. Plot an R chart.

In Exercises 47–50, use the following data: After analysing data for a long period of time, it was determined that samples of 500 readings of an organic pollutant for an area are distributed normally. For this pollutant, $\mu = 2.20\ \mu g/m^3$ *and* $\sigma = 0.50\ \mu g/m^3$.

47. In a sample, how many readings are expected to be between $1.50\ \mu g/m^3$ and $2.50\ \mu g/m^3$?

48. In a sample, how many readings are expected to be between $2.50\ \mu g/m^3$ and $3.50\ \mu g/m^3$?

49. In a sample, how many readings are expected to be above $1.00\ \mu g/m^3$?

50. In a sample, how many readings are expected to be below $2.00\ \mu g/m^3$?

In Exercises 51–60, find the indicated least-squares curve. Sketch the curve and data points on the same graph.

51. In a certain experiment, the resistance R of a certain resistor was measured as a function of the temperature T. The data found are shown in the following table. Find the least-squares line, expressing R as a function of T.

T (°C)	0.0	20.0	40.0	60.0	80.0	100
R (Ω)	25.0	26.8	28.9	31.2	32.8	34.7

52. An air-pollution monitoring station took samples of air each hour during the later morning hours and tested each sample for the number n of parts per million (ppm) of carbon monoxide. The results are shown in the table, where t is the number of hours after 6 A.M. Find the least-squares line for n as a function of t.

t (h)	0.0	1.0	2.0	3.0	4.0	5.0	6.0
n (ppm)	8.0	8.2	8.8	9.5	9.7	10.0	10.7

53. The *Mach number* of a moving object is the ratio of its speed to the speed of sound (1200 km/h). The following table shows the speed s of a jet aircraft, in terms of Mach numbers, and the time t after it starts to accelerate. Find the least-squares line of s as a function of t.

t (min)	0.00	0.60	1.20	1.80	2.40	3.00
s (Mach number)	0.88	0.97	1.03	1.11	1.19	1.25

54. In an experiment to determine the relation between the load x on a spring and the length y of the spring, the following data were found. Find the least-squares line that expresses y as a function of x.

Load (kg)	0.0	1.0	2.0	3.0	4.0	5.0
Length (cm)	10.0	11.2	12.3	13.4	14.6	15.9

55. The distance s of a missile above the ground at time t after being released from a plane is given by the following table. Find the least-squares curve of the form $s = mt^2 + b$ for these data.

t (s)	0.0	3.0	6.0	9.0	12.0	15.0	18.0
s (m)	3000	2960	2820	2600	2290	1900	1410

56. In an elementary experiment that measured the wavelength L of sound as a function of the frequency f, the following results were obtained.

Frequency (Hz)	240	320	400	480	560
Wavelength (cm)	140	107	81.0	70.0	60.0

Find the least-squares curve of the form $L = m(1/f) + b$ for these data.

57. Measurements were made on the current i (in A) in an electric circuit as a function of the time t (in s). The circuit contained a resistance of 5.00 Ω and an inductance of 10.0 H. The following data were found.

t (s)	0.00	2.00	4.00	6.00	8.00
i (A)	0.00	2.52	3.45	3.80	3.92

Find the least-squares curve $i = m(e^{-0.500t}) + b$ for these data. From the equation, determine the value of the current as t approaches infinity.

58. The power P (in W) generated by a wind turbine was measured for various wind velocities v (in km/h), as shown in the following table.

v (km/h)	10	15	20	25	30	40
p (W)	75	250	600	1200	2100	4800

Find the least-squares curve of the form $P = mv^3 + b$ for these data.

59. The vertical distance y of the cable of a suspension bridge above the surface of the bridge is measured at a horizontal distance x along the bridge from its centre. See Fig. 22.31. The results are as follows:

x (m)	0	100	200	300	400	500
y (m)	15	17	23	33	47	65

Plot these points and choose an appropriate function $f(x)$ for $y = m[f(x)] + b$. Then find the equation of the least-squares curve.

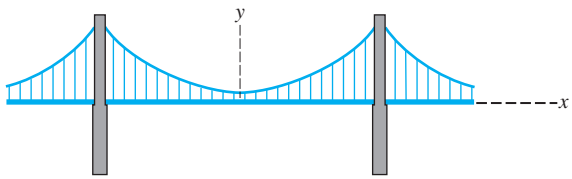

Fig. 22.31

60. After being heated, the temperature T of an insulated liquid is measured at times t as follows:

t (h)	0	2	4	6	8	10
T (°C)	100	85	72	63	54	48

Plot these points and choose an appropriate function $f(x)$ for $y = m[f(x)] + b$. Then find the equation of the least-squares curve.

In Exercises 61–64, use a graphing calculator, a spreadsheet, or software to solve the given problems related to the following data. Using aerial photography, the area A (in km^2) of an oil spill as a function of the time t (in h) after the spill was found to be as follows:

t (h)	1.0	2.0	4.0	6.0	8.0	10.0
A (km^2)	1.4	2.5	4.7	6.8	8.8	10.2

61. Find the linear equation $y = ax + b$ to fit these data.

62. Find the quadratic equation $y = ax^2 + bx + c$ to fit these data.

63. Find the power equation $y = ax^b$ to fit these data.

64. Find the value of the linear coefficient of correlation r for these data.

In Exercises 65–70, solve the given problems.

65. With 30.5% of the area under the normal curve between $z_1 = 0.5$ and z_2, to the right of z_1, find z_2.

66. With 79.8% of the area under the normal curve between z_1 and $z_2 = 2.1$, find z_1.

67. The nth root of the product of n positive numbers is the *geometric mean* of the numbers. Find the geometric mean of the carbon monoxide readings in Exercise 52.

68. One use of the geometric mean (see Exercise 67) is to find an average ratio. By finding the geometric mean, find the average Mach number for the jet in Exercise 53.

69. Show that Eqs. (22.18) and (22.19) satisfy the equation $\bar{y} = m\bar{x} + b$.

70. Given that $\sum(x - \bar{x})^2 = \sum x^2 - n\bar{x}^2$, derive Eq. (22.3) from Eq. (22.2).

Writing Exercise

71. A research institute is planning a study of the effect of education on the income of workers. Write two or three paragraphs explaining what data should be collected and which of the measures discussed in this chapter would be useful in analysing the data.

CHAPTER 22 PRACTICE TEST

In Problems 1–3, use the following set of numbers.

5, 6, 1, 4, 9, 5, 7, 3, 8, 10, 5, 8, 4, 9, 6

1. Find the median.

2. Find the mode.

3. Draw a histogram with five classes and the lowest class limit at 1.

In Problems 4–8, use the following data: Two machine parts are considered satisfactorily assembled if their total thickness (to the nearest 0.01 cm*) is between or equal to 0.92* cm *and 0.94* cm*. One hundred assemblies are tested, and the class mark of the thicknesses and the number of assemblies in each class are given in the following table.*

Total Thickness (cm)	0.90	0.91	0.92	0.93	0.94	0.95	0.96
Number	3	9	31	38	12	5	2

4. Find the mean.

5. Find the standard deviation.

6. Draw a frequency polygon.

7. Make a relative frequency table.

8. Draw an ogive (less than).

9. A random sample of size $n = 175$ is taken from a large population so that $\bar{x} = 143$ and $s = 5.2$. Construct a 95% confidence interval for the population mean μ.

10. A random sample of size $n = 120$ is taken from a large population and 33 are found defective. Construct a 99% confidence interval for the proportion p of defectives in the population.

11. For a set of values that are normally distributed, what percent of them is below the value (greater than the mean) for which the z-score is 0.2257?

12. The machine-part assemblies in Problems 4–8 were tested in groups of five each hour for 20 hours. Explain, in general, how to use the data from the test subgroups to plot an R chart.

13. Find the equation of the least-squares line for the points indicated in the following table. Graph the line and data points on the same graph.

x	1	3	5	7	9
y	5	11	17	20	27

14. The velocity (in m/s) of an object moving down an inclined plane was measured as a function of the distance (in m) it moved, with the following results:

Distance (m)	1.00	3.00	5.00	7.00	9.00
Velocity (m/s)	1.10	1.90	2.50	2.90	3.30

Find the equation of the least-squares curve of the form $y = m\sqrt{x} + b$, which expresses the velocity as a function of the distance.