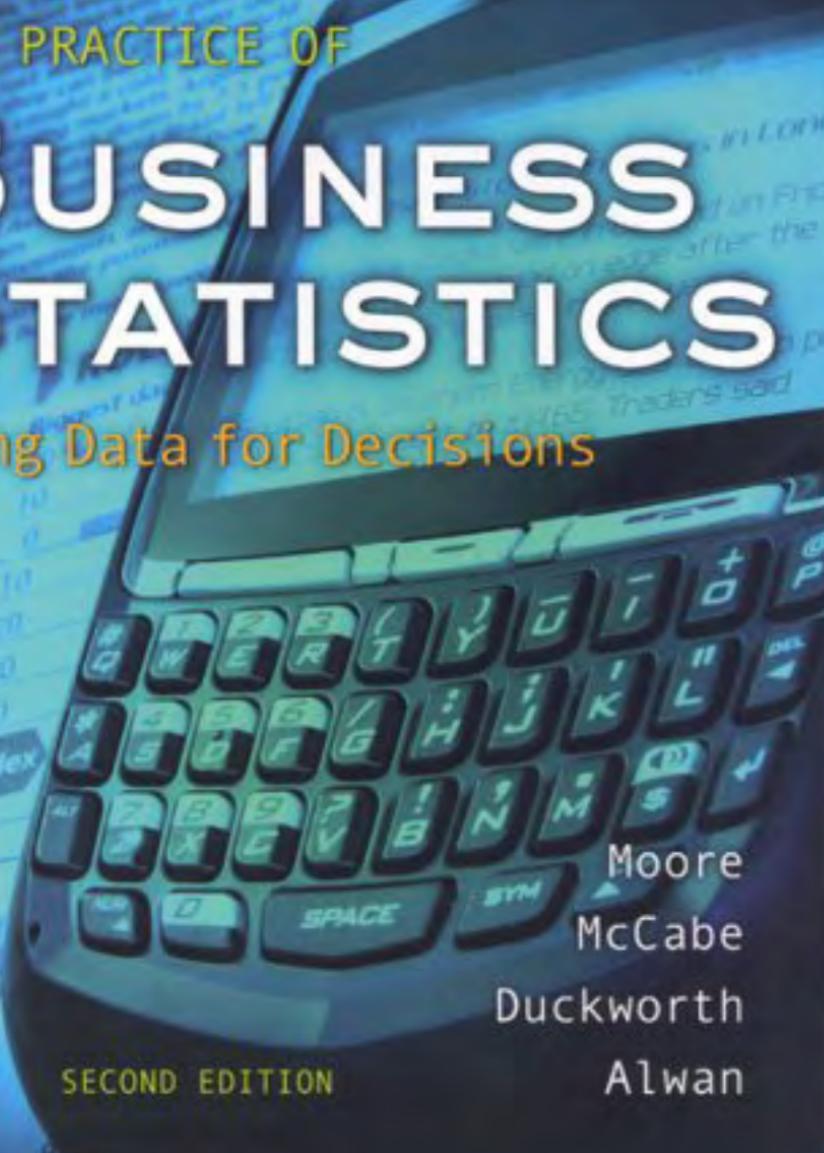


THE PRACTICE OF

BUSINESS STATISTICS

Using Data for Decisions



Moore

McCabe

Duckworth

Alwan

SECOND EDITION

Data

PART
1

CHAPTER 1
Examining Distributions

CHAPTER 2
Examining Relationships

CHAPTER 3
Producing Data

Prelude

The case of the missing vans

Auto manufacturers lend their dealers money to help them keep vehicles on their lots. The loans are repaid when the vehicles are sold. A Long Island auto dealer named John McNamara borrowed over \$6 billion from General Motors between 1985 and 1991. In December 1990 alone, Mr. McNamara borrowed \$425 million to buy 17,000 GM vans customized by an Indiana company for sale overseas. GM happily lent McNamara the money because he always repaid the loans.

Let's pause to consider the numbers, as GM should have done but didn't. The entire van-customizing industry produces only about 17,000 customized vans a month. So McNamara was claiming to buy an entire month's production. These large, luxurious, and gas-guzzling vehicles are designed for U.S. interstate highways. The recreational vehicle trade association says that only 1.35% were exported in 1990. It's not plausible to claim that 17,000 vans in a single month are being bought for export. McNamara's claimed purchases were large even when compared with total production of vans. Chevrolet, for example, produced 100,067 full-sized vans in all of 1990.

Having looked at the numbers, you can guess the rest. McNamara admitted in federal court in 1992 that he was defrauding GM on a massive scale. The Indiana company was a shell set up by McNamara, its invoices were phony, and the vans didn't exist. McNamara borrowed vastly from GM, used most of each loan to pay off the previous loan (thus establishing a record as a good credit risk), and skimmed off a bit for himself. The bit he skimmed amounted to over \$400 million. GM set aside \$275 million to cover its losses. Two executives, who should have looked at the numbers relevant to their business, were fired.¹

Examining Distributions

- 1.1 Displaying Distributions with Graphs**
- 1.2 Describing Distributions with Numbers**
- 1.3 The Normal Distributions**



Introduction

Statistics is the science of data. We therefore begin our study of statistics by mastering the art of examining data. Any set of data contains information about some group of *individuals*. The information is organized in *variables*.

INDIVIDUALS AND VARIABLES

Individuals are the objects described by a set of data. Individuals may be people, but they may also be business firms, common stocks, or other objects.

A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

A college's student data base, for example, includes data about every currently enrolled student. The students are the individuals described by the data set. For each individual, the data contain the values of variables such as date of birth, gender (female or male), choice of major, and grade point average. In practice, any set of data is accompanied by background information that helps us understand the data. When you plan a statistical study or explore data from someone else's work, ask yourself the following questions:

1. **Who?** What **individuals** do the data describe? How **many** individuals appear in the data?
2. **What?** How **many variables** do the data contain? What are the **exact definitions** of these variables? In what **units of measurement** is each variable recorded? The sales of business firms, for example, might be measured in dollars, in millions of dollars, or in euros.
3. **Why?** What **purpose** do the data have? Do we hope to answer some specific questions? Do we want to draw conclusions about individuals other than the ones we actually have data for? Are the variables recorded suitable for the intended purpose?

Some variables, like gender and college major, simply place individuals into categories. Others, like height and grade point average, take numerical values for which we can do arithmetic. It makes sense to give an average income for a company's employees, but it does not make sense to give an "average" gender. We can, however, count the numbers of female and male employees and do arithmetic with these counts.

CATEGORICAL AND QUANTITATIVE VARIABLES

A **categorical variable** places an individual into one of several groups or categories.

A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.

The **distribution** of a variable tells us what values it takes and how often it takes these values.

EXAMPLE 1.1**A corporate data set**

Here is a small part of the data set in which CyberStat Corporation records information about its employees:

	A	B	C	D	E	F
1	Name	Age	Gender	Race	Salary	Job Type
2	Fleetwood, Delores	39	Female	White	62,100	Management
3	Perez, Juan	27	Male	White	47,350	Technical
4	Wang, Lin	22	Female	Asian	18,250	Clerical
5	Johnson, LaVerne	48	Male	Black	77,600	Management
6						

Enter NUM

The *individuals* described are the employees. Each row records data on one *case*. You will often see each row of data called a *case*. Each column contains the values of one *variable* for all the individuals. In addition to the person's name, there are 5 variables. Gender, race, and job type are categorical variables. Age and salary are quantitative variables. You can see that age is measured in years and salary in dollars.

Most data tables follow this format—each row is an individual, and each column is a variable. This data set appears in an Excel *spreadsheet* display that has rows and columns ready for your use. Spreadsheets are commonly used to enter and transmit data.

APPLY YOUR KNOWLEDGE

- 1.1 Motor vehicle fuel economy. Here is a small part of a data set that describes the fuel economy (in miles per gallon) of 2001 model year motor vehicles:

Make and model	Vehicle type	Transmission type	Number of cylinders	City MPG	Highway MPG
:					
BMW 330CI	Subcompact	Automatic	6	19	27
BMW 330CI	Subcompact	Manual	6	21	30
Buick Century	Midsized	Automatic	6	20	29
Chevrolet Blazer	Four-wheel drive	Automatic	6	15	20
:					

- (a) What are the individuals in this data set?
 (b) For each individual, what variables are given? Which of these variables are categorical and which are quantitative?
- 1.2 Data from a pharmaceutical company medical study contain values of many variables for each of the people who were the subjects of the study. Which of the following variables are categorical and which are quantitative?
- (a) Gender (female or male)
 - (b) Age (years)
 - (c) Race (Asian, black, white, or other)
 - (d) Smoker (yes or no)
 - (e) Systolic blood pressure (millimeters of mercury)
 - (f) Level of calcium in the blood (micrograms per milliliter)

*exploratory
data analysis*

1.1 Displaying Distributions with Graphs

Statistical tools and ideas help us examine data in order to describe their main features. This examination is called **exploratory data analysis**. Like an explorer crossing unknown lands, we want first to simply describe what we see. Here are two basic strategies that help us organize our analysis exploration of a set of data:

- Begin by examining each variable by itself. Then move on to study the relationships among the variables.
- Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We will follow these principles in organizing our learning. This chapter presents methods for describing a single variable. We study relationships among several variables in Chapter 2. Within each chapter, we begin with graphical displays, then add numerical summaries for a more complete description.

Categorical variables: bar graphs and pie charts

The values of a categorical variable are labels for the categories, such as “male” and “female.” The distribution of a categorical variable lists the categories and gives either the **count** or the **percent** of individuals who fall in each category.

EXAMPLE 1.2

Firestone tires

In the summer of 2000, Firestone received much attention in the media due to a number of traffic accidents believed to be caused by tread separation in Firestone tires. By the end of the year, 148 fatalities were suspected to be due to defective Firestone tires. Here is the distribution of tire models for 2969 accidents reported to the government that involved Firestone tires. The table reflects the wording of individual accident reports. We suspect, for example, that “ATX” and “Firestone ATX” refer to the same tire model, but we cannot be sure.²

Tire Model	Count	Percent
ATX	554	18.7
Firehawk	38	1.3
Firestone	29	1.0
Firestone ATX	106	3.6
Firestone Wilderness	131	4.4
Radial ATX	48	1.6
Wilderness	1246	42.0
Wilderness AT	709	23.9
Wilderness HT	108	3.6
Total	2969	100.1

roundoff error

The percents add to 100.1 rather than to 100 because of **roundoff error**. Each entry in the percent column is rounded to the nearest tenth of a percent, and the rounded percents do not add to exactly 100.

bar graph

The graphs in Figure 1.1 display these data. The **bar graph** in Figure 1.1(a) quickly compares the sizes of the nine response groups. The heights of the nine bars show the counts for the nine tire model categories. The **pie chart** in Figure 1.1(b) helps us see what part of the whole each group forms. For example, the “Wilderness” slice makes up 42% of the pie because 42% of the accidents involved the Wilderness tire model. To make a pie chart, you must include all the categories that make up a whole. Bar graphs are more flexible. For example, you can use a bar graph to compare the numbers of students at your college majoring in biology, business, and political science. A pie chart cannot make this comparison because not all students fall into one of these three majors.

The categories in a bar graph can be put in any order. In Figure 1.1(a), we arranged the categories alphabetically. Let’s arrange the categories in order of their percents, from highest percent to lowest percent. Figure 1.1(c) displays our new bar graph with percents marked on the vertical axis. A bar graph whose categories are ordered from most frequent to least frequent is called a **Pareto chart**.³ Pareto charts identify the “vital few” categories that contain most of our observations. In Figure 1.1(c), we see that 84.6% of the accidents involved just three of the nine tire models (Wilderness,

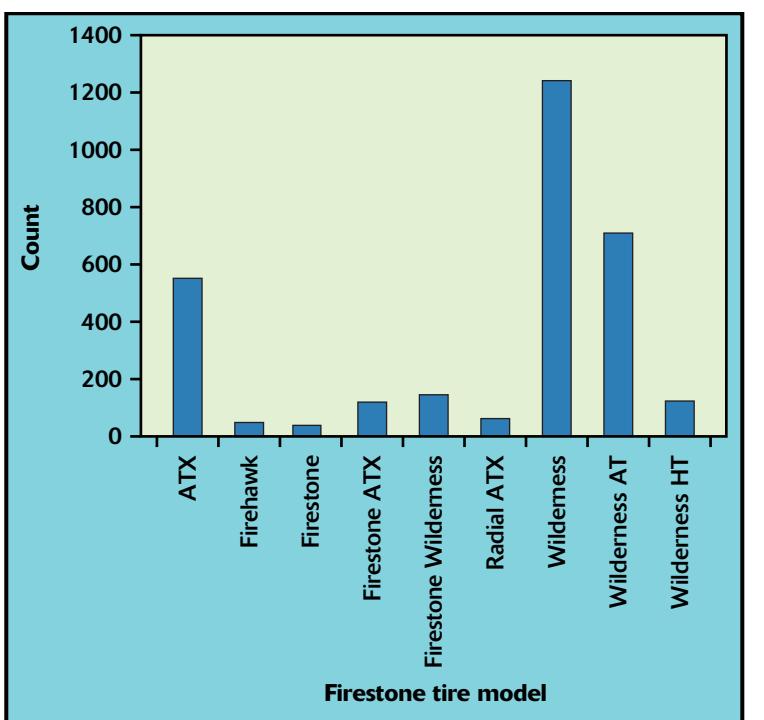
pie chart**Pareto chart**

FIGURE 1.1(a) Bar graph of accidents involving Firestone tire models.

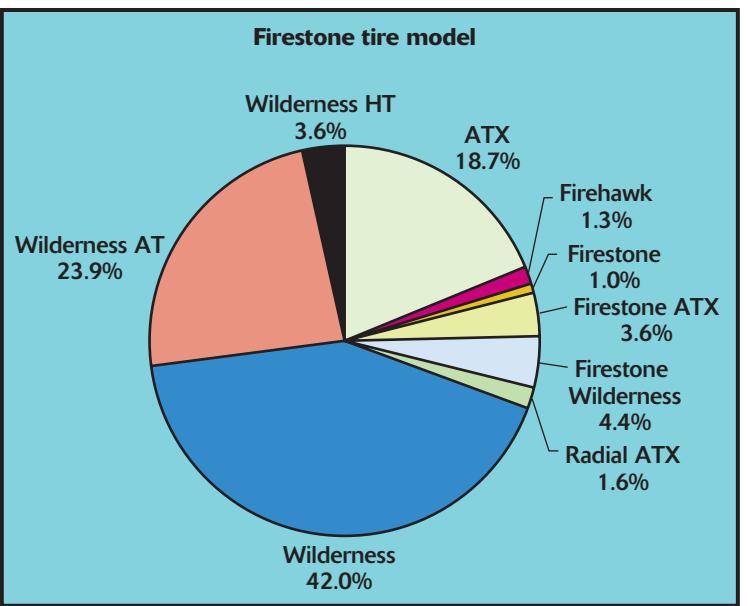


FIGURE 1.1(b) Pie chart of the Firestone tire data.

Wilderness AT, ATX). Perhaps these three tire models have a common flaw that contributes to the large number of accidents involving one of them. Perhaps, however, these are best-selling models that have more accidents simply because there are more of them on the road. As is often the case, we need more data to draw firm conclusions.

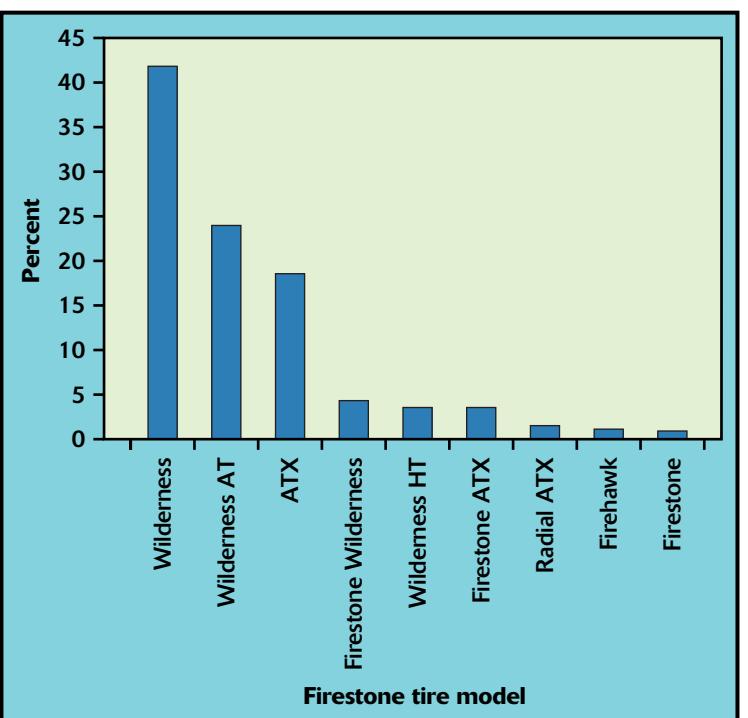


FIGURE 1.1(c) Pareto chart of the Firestone tire data.

Bar graphs, pie charts, and Pareto charts help an audience grasp a distribution quickly. They are, however, of limited use for data analysis because it is easy to understand categorical data on a single variable such as “college major” without a graph. We will move on to quantitative variables, where graphs are essential tools.

APPLY YOUR KNOWLEDGE

- 1.3 Undergraduate majors. Here are data on the percent of undergraduate majors for various colleges within a university:⁴

Agriculture	13.9%
Business	17.0%
Design	8.2%
Education	8.5%
Engineering	21.0%
Liberal arts and sciences	29.9%

- (a) Present these data in a well-labeled bar graph.
(b) Would it also be correct to use a pie chart to display these data? Explain your answer.
- 1.4 Occupational deaths. In 1999 there were 6023 job-related deaths in the United States. Among these were 807 deaths in agriculture-related jobs (including forestry and fishing), 121 in mining, 1190 in construction, 719 in manufacturing, 1006 in transportation and public utilities, 237 in wholesale trade, 507 in retail trade, 105 in finance-related jobs (including insurance and real estate), 732 in service-related jobs, and 562 in government jobs.⁵
- (a) Find the percent of occupational deaths for each of these job categories, rounded to the nearest whole percent. What percent of job-related deaths were in categories not listed here?
(b) Make a well-labeled bar graph of the distribution of occupational deaths. Be sure to include an “other occupations” bar.
(c) Make a well-labeled Pareto chart of these data. What percent of all occupational deaths are accounted for by the first three categories in your Pareto chart?
(d) Would it also be correct to use a pie chart to display these data? Explain your answer.

Quantitative variables: histograms

Quantitative variables often take many values. A graph of the distribution is clearer if nearby values are grouped together. The most common graph of the distribution of one quantitative variable is a **histogram**.



CASE 1.1 STATE UNEMPLOYMENT RATES

Each month the Bureau of Labor Statistics (BLS) announces the *unemployment rate* for the previous month. Unemployment rates are economically important and politically sensitive. Unemployment may of course differ greatly among the states because types of work are unevenly distributed

TABLE 1.1 Unemployment rates by state, December 2000

State	Percent	State	Percent	State	Percent
Alabama	4.0	Louisiana	5.3	Ohio	3.7
Alaska	6.1	Maine	2.6	Oklahoma	2.6
Arizona	3.3	Maryland	3.3	Oregon	4.0
Arkansas	3.9	Massachusetts	2.0	Pennsylvania	3.8
California	4.3	Michigan	3.4	Puerto Rico	8.9
Colorado	2.1	Minnesota	2.8	Rhode Island	3.2
Connecticut	1.5	Mississippi	4.3	South Carolina	3.3
Delaware	3.3	Missouri	3.2	South Dakota	2.3
Florida	3.2	Montana	4.9	Tennessee	3.8
Georgia	3.0	Nebraska	2.5	Texas	3.4
Hawaii	3.6	Nevada	4.0	Utah	2.7
Idaho	5.0	New Hampshire	2.2	Vermont	2.4
Illinois	4.5	New Jersey	3.5	Virginia	1.9
Indiana	2.7	New Mexico	4.9	Washington	4.9
Iowa	2.5	New York	4.2	West Virginia	5.5
Kansas	3.2	North Carolina	3.6	Wisconsin	3.0
Kentucky	3.7	North Dakota	2.7	Wyoming	3.7

Source: Bureau of Labor Statistics.

across the country. Table 1.1 presents the unemployment rates for each of the 50 states and Puerto Rico in December 2000. We will examine these data in detail, starting with the nature of the variable recorded.

What does it mean to be “unemployed” in the eyes of the government? People who are not available for work (retired people, for example, or students who do not want to work while in school) should not be counted as unemployed just because they don’t have a job. To be unemployed, a person must first be in the *labor force*. That is, she must be available for work and looking for work. The unemployment rate is

$$\text{unemployment rate} = \frac{\text{number of people unemployed}}{\text{number of people in the labor force}}$$

To complete the exact definition of the unemployment rate, the BLS has very detailed descriptions of what it means to be “in the labor force” and what it means to be “employed.” For example, if you are on strike but expect to return to the same job, you count as employed. If you are not working and did not look for work in the last two weeks, you are not in the labor force. So people who say they want to work but are too discouraged to keep looking for a job don’t count as unemployed. The details matter. The official unemployment rate would be different if the government used a different definition of unemployment.



To understand the unemployment situation in December 2000, we begin with a graph.

EXAMPLE 1.3**A histogram of state unemployment rates**

To make a histogram of the data in Table 1.1, proceed as follows.

Step 1. Divide the range of the data into classes of equal width. The unemployment rates range from 1.5% to 8.9%, so we choose as our classes

$$1.0 \leq \text{unemployment rate} < 2.0$$

$$2.0 \leq \text{unemployment rate} < 3.0$$

⋮

$$8.0 \leq \text{unemployment rate} < 9.0$$

Be sure to specify the classes precisely so that each individual falls into exactly one class. An unemployment rate of 1.9% would fall into the first class, but 2.0% falls into the second.

Step 2. Count the number of individuals in each class. Here are the counts:

Class	Count
1.0 to 1.9	2
2.0 to 2.9	13
3.0 to 3.9	21
4.0 to 4.9	10
5.0 to 5.9	3
6.0 to 6.9	1
7.0 to 7.9	0
8.0 to 8.9	1

Step 3. Draw the histogram. Mark on the horizontal axis the scale for the variable whose distribution you are displaying. The variable is “unemployment rate” in this example. The scale runs from 0 to 10 to span the data. The vertical axis contains the scale of counts. Each bar represents a class. The base of the bar covers the class, and the bar height is the class count. There is no horizontal space between the bars unless a class is empty, so that its bar has height zero. Figure 1.2(a) is our histogram.



The bars of a histogram should cover the entire range of values of a variable, with no space between bars unless a class is empty. When the possible values of a variable have gaps between them, extend the bases of the bars to meet halfway between two adjacent possible values. For example, in a histogram of the ages in years of university faculty, the bars representing 25 to 29 years and 30 to 34 years would meet at 29.5.

Our eyes respond to the *area* of the bars in a histogram.⁶ Because the classes are all the same width, area is determined by height and all classes are fairly represented. There is no one right choice of the classes in a histogram. Too few classes will give a “skyscraper” graph, with all values in a few

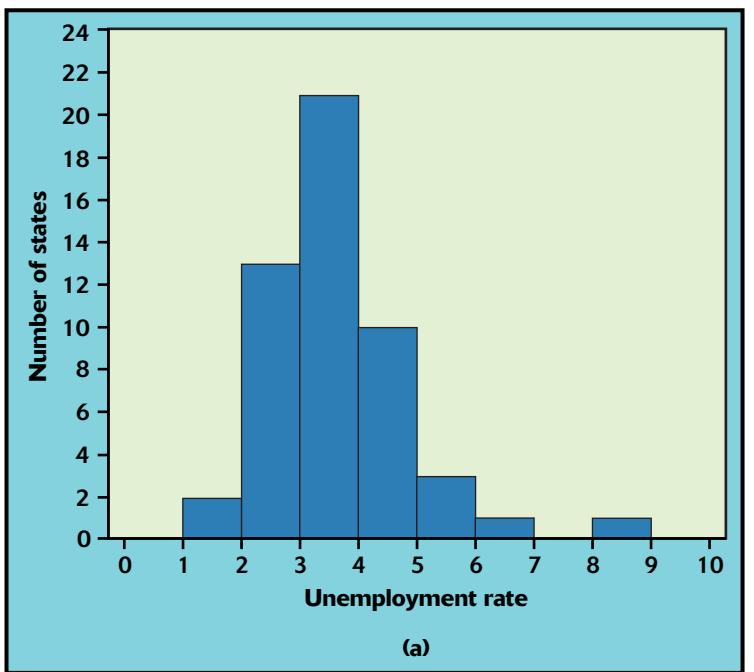


FIGURE 1.2 Histograms of the December 2000 unemployment rates in the 50 states and Puerto Rico, from Table 1.1. The two graphs differ only in the number of classes used.

TABLE 1.2 Highway gas mileage for 2001 model year midsize cars

Model	MPG	Model	MPG
Acura 3.5RL	24	Lexus GS300	24
Audi A6 Quattro	24	Lincoln-Mercury LS	25
BMW 740I	23	Lincoln-Mercury Sable	27
Buick Regal	30	Mazda 626	28
Cadillac Catera	24	Mercedes-Benz E320	28
Cadillac Eldorado	27	Mercedes-Benz E430	25
Chevrolet Lumina	29	Mercedes-Benz E55 AMG	24
Chrysler Sebring	30	Mitsubishi Diamante	25
Dodge Stratus	30	Mitsubishi Galant	28
Honda Accord	30	Nissan Maxima	26
Hyundai Sonata	28	Oldsmobile Intrigue	28
Infiniti I30	26	Saab 9-3	28
Infiniti Q45	23	Saturn L100	33
Jaguar S/C	22	Toyota Camry	32
Jaguar Vanden Plas	24	Volkswagen Passat	31
Jaguar XJ8L	24	Volvo S80	26

classes with tall bars. Too many will produce a “pancake” graph, with most classes having one or no observations. Neither choice will give a good picture of the shape of the distribution. You must use your judgment in choosing classes to display the shape. Statistics software will choose the classes for you. The computer’s choice is usually a good one, but you can change it if you want. Figure 1.2(b) is another histogram of the unemployment data, with classes half as wide as in Figure 1.2(a). We have kept the same scales in both figures for easy comparison.

APPLY YOUR KNOWLEDGE

- 1.5 **Automobile fuel economy.** Environmental Protection Agency regulations require automakers to give the city and highway gas mileages for each model of car. Table 1.2 gives the highway mileages (miles per gallon) for 32 midsize 2001 model year cars.⁷ Make a histogram of the highway mileages of these cars.
- 1.6 **Architectural firms.** Table 1.3 contains data describing firms engaged in commercial architecture in the Indianapolis, Indiana, area.⁸ One of the variables is the count of full-time staff members employed by each firm. Make a histogram of the staff counts.

Interpreting histograms

Making a statistical graph is not an end in itself. The purpose of the graph is to help us understand the data. After you make a graph, always ask, “What do I see?” Once you have displayed a distribution, you can see its important features as described in the following box.

EXAMINING A DISTRIBUTION

In any graph of data, look for the **overall pattern** and for striking deviations from that pattern.

You can describe the overall pattern of a histogram by its **shape**, **center**, and **spread**.

An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern.

We will learn how to describe center and spread numerically in Section 1.2. For now, we can describe the center of a distribution by its *midpoint*, the value with roughly half the observations taking smaller values and

TABLE 1.3 Indianapolis architectural firms, 1998

Name	1998 total billings (\$millions)	1998 arch. billings (\$millions)	1997 arch. billings (\$millions)	Architects	Engineers	Staff
Schmidt Associates	11.5	11.5	5.0	19	7	111
CSO Architects	12.6	9.3	8.8	29	12	126
BSA Design	13.8	9.0	7.5	31	21	155
InterDesign Group	6.4	6.4	4.7	19	3	57
Browning Day Mullins	8.5	6.2	3.9	24	0	70
Ratio Architects	7.8	6.2	5.6	21	0	68
Odle McGuire	8.3	4.2	4.1	9	2	62
Gibraltar	5.8	3.6	4.2	12	4	52
American Consulting	12.0	3.5	3.3	5	23	131
Fanning/Howey	5.3	3.5	3.8	12	4	61
HNTB Corporation	15.0	3.4	3.0	10	35	110
SchenkelSchultz	2.7	2.7	2.5	5	0	22
Simmons & Associates	2.2	2.2	2.4	2	1	13
Paul I. Cripe	7.5	2.1	1.7	5	13	115
Plus4 Architects	2.7	2.1	2.0	5	0	15
Architectural Alliance	2.0	2.0	1.2	4	0	14
Blackburn Architects	2.6	1.8	1.5	8	1	24
Snapp & Associates	1.8	1.8	1.4	3	1	7
Sebree & Associates	1.7	1.7	1.0	3	0	15
Armstrong & Associates	9.1	1.6	2.4	3	23	96
Lamson & Condon	1.6	1.6	2.0	4	0	17
RQAW	6.3	1.6	2.3	6	14	72
Woollen Molzan	1.6	1.6	1.3	5	0	15
United Consulting	7.0	1.3	0.7	2	12	70
URS Greiner Woodward	1.7	1.3	0.9	5	1	17

half taking larger values. We can describe the spread of a distribution by giving the *smallest and largest values*.

EXAMPLE 1.4**The distribution of unemployment rates**

Look again at the histogram in Figure 1.2(b). The first feature we notice is an *outlier*. Puerto Rico has much higher unemployment (8.9%) than any of the 50 states. Some outliers are due to mistakes, such as typing 5.9 as 8.9. Other outliers point to the special nature of some observations. Although Puerto Rico is included in the government report from which we took the unemployment data, it is not a state and we should omit it when thinking about domestic unemployment.

Shape: The distribution has a *single peak*. Even omitting the outlier, the histogram is somewhat *right-skewed*—that is, the right tail extends farther from the peak than does the left tail. The states with unemployment rates at or higher than 5% are Alaska, Idaho, Louisiana, and West Virginia. These states are often near the top in unemployment due to generally weak job opportunities. Note that winter weather does not explain high unemployment in Alaska and Idaho—the data are *seasonally adjusted* to compensate for regular changes such as job losses in the winter months in northern states. **Center:** The midpoint of the 51 observations is close to the single peak, at about 3.5%. **Spread:** The spread is 1.5% to 6.5% if we ignore Puerto Rico.



When you describe a distribution, concentrate on the main features. Look for major peaks, not for minor ups and downs in the bars of the histogram. Look for clear outliers, not just for the smallest and largest observations. Look for rough *symmetry* or clear *skewness*.

SYMMETRIC AND SKEWED DISTRIBUTIONS

A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.

A distribution is **skewed to the right** if the right side of the histogram (containing the half of the observations with larger values) extends much farther out than the left side. It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.

EXAMPLE 1.5**Major league baseball salaries**

Figure 1.3 displays a histogram of the average salaries (in millions of dollars) for players on the 30 major league baseball teams as of opening day of the 2000 season. The distribution has a single peak between 2.0 and 2.5 million dollars and falls off on either side of this peak. The two sides of the histogram are roughly the same shape, so we call the distribution *symmetric*.

In mathematics, symmetry means that the two sides of a figure like a histogram are exact mirror images of each other. Data are almost never exactly symmetric, but we are willing to call histograms like that in Figure 1.3 “approximately symmetric” as an overall description.

For comparison, Figure 1.4 shows the distribution of salaries for the players on one team, the Cincinnati Reds. This distribution also has a single peak but is

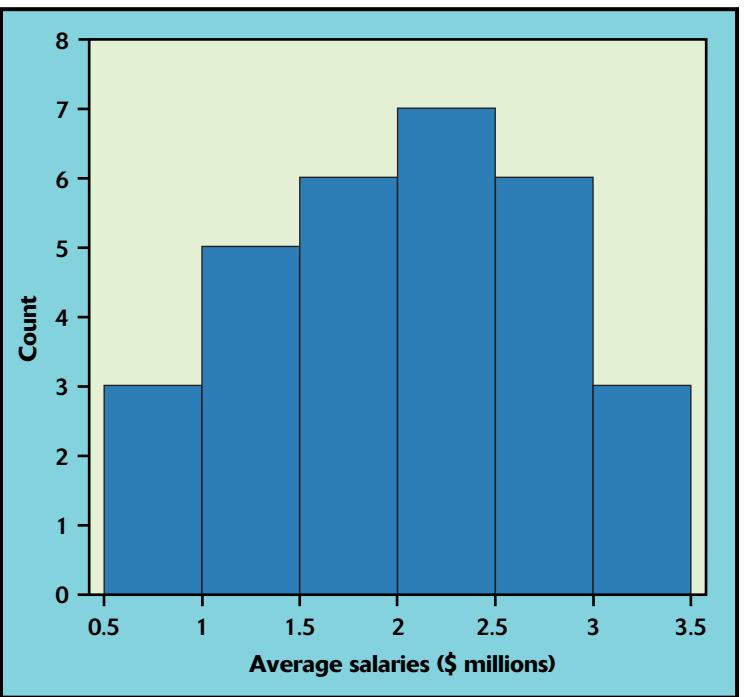


FIGURE 1.3 The distribution of the average salaries for the 30 major league baseball teams on opening day of the 2000 season, for Example 1.5.

skewed to the right. Many of the players receive relatively small salaries (less than one million dollars), but a few earn large salaries (more than three million dollars). These large salaries form the long right tail of the histogram.

Notice that the vertical scale in Figure 1.4 is not the *count* of salaries but the *percent* of salaries in each class. A histogram of percents rather than counts is

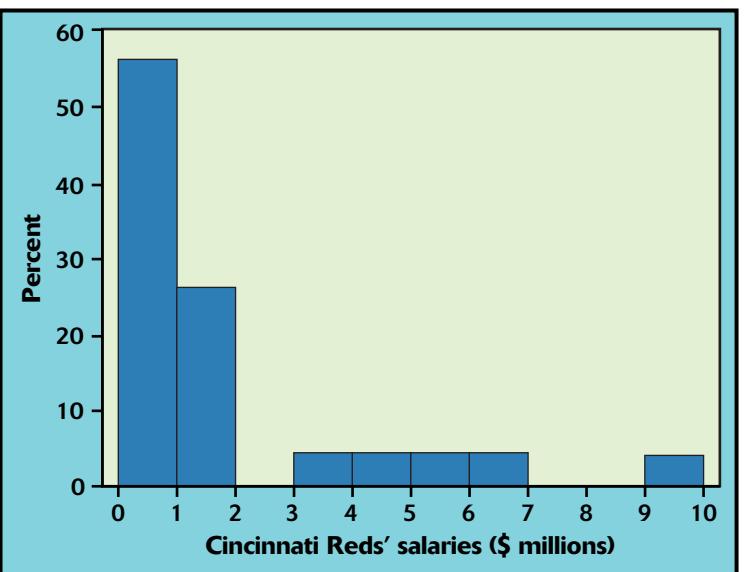


FIGURE 1.4 The distribution of the individual salaries of Cincinnati Reds players on opening day of the 2000 season, for Example 1.5.

convenient when the counts are very large or when we want to compare several distributions.



The overall shape of a distribution is important information about a variable. Some types of data regularly produce distributions that are symmetric or skewed. For example, data on the diameters of ball bearings produced by a manufacturing process tend to be symmetric. Data on incomes (whether of individuals, companies, or nations) are usually strongly skewed to the right. There are many moderate incomes, some large incomes, and a few very large incomes. Do remember that many distributions have shapes that are neither symmetric nor skewed. Some data show other patterns. Scores on an exam, for example, may have a cluster near the top of the scale if many students did well. Or they may show two distinct peaks if a tough problem divided the class into those who did and didn't solve it. Use your eyes and describe what you see.

APPLY YOUR KNOWLEDGE

- 1.7 **Automobile fuel economy.** Table 1.2 (page 13) gives data on the fuel economy of 2001 model midsize cars. Based on a histogram of these data:
 - (a) Describe the main features (shape, center, spread, outliers) of the distribution of highway mileage.
 - (b) The government imposes a “gas guzzler” tax on cars with low gas mileage. Do you think that any of these cars are subject to this tax?
- 1.8 How would you describe the center and spread of the distribution of average salaries in Figure 1.3? Of the distribution of Cincinnati Reds’ salaries in Figure 1.4?

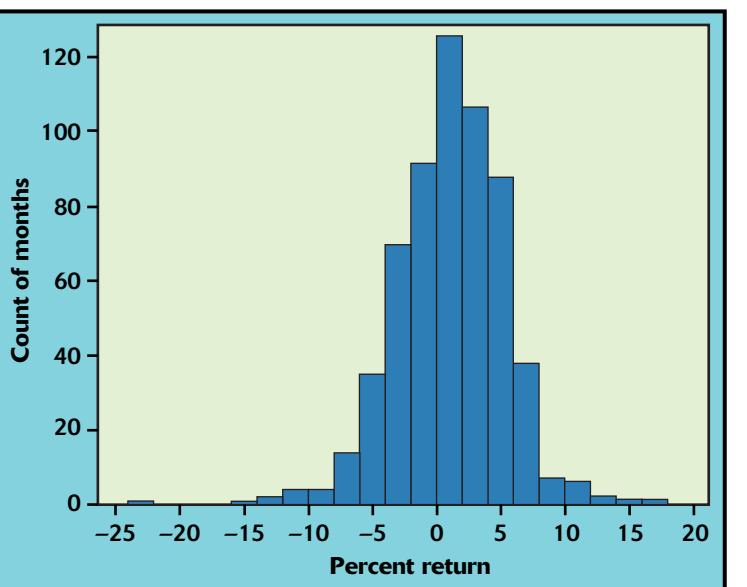


FIGURE 1.5 The distribution of monthly returns for all U.S. common stocks in the 600 months from January 1951 to December 2000, for Exercise 1.9.

- 1.9 **Returns on common stocks.** The total return on a stock is the change in its market price plus any dividend payments made. Total return is usually expressed as a percent of the beginning price. Figure 1.5 is a histogram of the distribution of the monthly returns for all stocks listed on U.S. markets for the years 1951 to 2000 (600 months).⁹ The low outlier is the market crash of October 1987, when stocks lost more than 22% of their value in one month.
- (a) Describe the overall shape of the distribution of monthly returns.
 - (b) What is the approximate center of this distribution? (For now, take the center to be the value with roughly half the months having lower returns and half having higher returns.)
 - (c) Approximately what were the smallest and largest total returns, leaving out the outlier? (This describes the spread of the distribution.)
 - (d) A return less than zero means that stocks lost value in that month. About what percent of all months had returns less than zero?

Quantitative variables: stemplots

Histograms are not the only graphical display of distributions of quantitative variables. For small data sets, a *stemplot* is quicker to make and presents more detailed information.

STEMPLOT

To make a *stemplot*:

1. Separate each observation into a **stem** consisting of all but the final (rightmost) digit and a **leaf**, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.
2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

EXAMPLE 1.6

A stemplot of state unemployment rates



The state unemployment rates in Table 1.1 (page 10) have only two digits, so the whole-number part of the observation is the stem and the second digit (tenths) is the leaf. For example, Alabama's 4.0% appears as a leaf 0 on the 4 stem. California's 4.3% adds a leaf 3 on the same stem. After placing each state as a leaf on the proper stem, rearrange the leaves in order from left to right. Figure 1.6(a) is our stemplot for the 50 states (omitting Puerto Rico).

A stemplot looks like a histogram turned on end. The stemplot in Figure 1.6(a) resembles the histogram in Figure 1.2(a). The stemplot, unlike the histogram, preserves the actual value of each observation. We interpret stemplots like histograms, looking for the overall pattern and for any outliers.

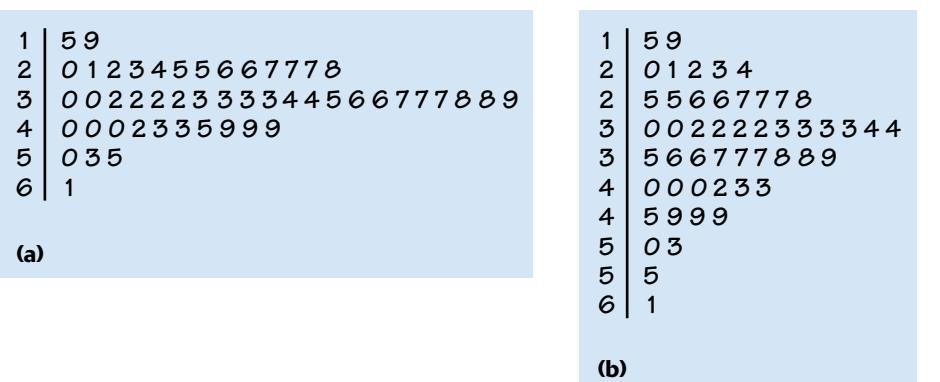


FIGURE 1.6 Stemplots of the December 2000 unemployment rates in the 50 states, for Example 1.6. Figure 1.6(b) uses split stems. Compare the histograms of these data in Figure 1.2.

You can choose the classes in a histogram. The classes (the stems) of a stemplot are given to you. When the observed values have many digits, it is often best to **round** the numbers to just a few digits before making a stemplot. For example, a stemplot of data like

3.468 2.137 2.981 1.095 ...

would have very many stems and no leaves or just one leaf on most stems. You can round these data to

3.5 2.1 3.0 1.1 ...

before making a stemplot.

splitting stems

You can also **split stems** to double the number of stems when all the leaves would otherwise fall on just a few stems. Each stem then appears twice. Leaves 0 to 4 go on the upper stem and leaves 5 to 9 go on the lower stem. Splitting the stems in Figure 1.6(a), for example, produces the new stemplot in Figure 1.6(b). The greater number of stems gives a clearer picture of the distribution. Rounding and splitting stems are matters for judgment, like choosing the classes in a histogram. Stemplots work well for small sets of data. When there are more than 100 observations, a histogram is almost always a better choice.

APPLY YOUR KNOWLEDGE

- 1.10 **Architectural firms.** Table 1.3 (page 14) gives the number of full-time staff employed by Indianapolis architectural firms. Make a stemplot of the staff counts. What are the main features of the shape of this distribution?
- 1.11 **Supermarket shoppers.** A marketing consultant observed 50 consecutive shoppers at a supermarket. One variable of interest was how much each shopper spent in the store. Here are the data (in dollars), arranged from smallest to largest:

3.11 8.88 9.26 10.81 12.69 13.78 15.23 15.62 17.00 17.39
 18.36 18.43 19.27 19.50 19.54 20.16 20.59 22.22 23.04 24.47
 24.58 25.13 26.24 26.26 27.65 28.06 28.08 28.38 32.03 34.98
 36.37 38.64 39.16 41.02 42.97 44.08 44.67 45.40 46.69 48.65
 50.39 52.75 54.80 59.07 61.22 70.32 82.70 85.76 86.37 93.34

Round these amounts to the nearest dollar and then make a stemplot of these data. About where is the center of the distribution? Are there any outliers? What is the spread of the values (ignoring any outliers)? Is the distribution symmetric, skewed left, or skewed right? Make a second stemplot of the data by splitting the stems as described in this section.

Time plots

Many variables are measured at intervals over time. We might, for example, measure the cost of raw materials for a manufacturing process each month or the price of a stock at the end of each day. In these examples, our main interest is change over time. To display change over time, make a *time plot*.

TIME PLOT

A time plot of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the variable you are measuring on the vertical scale. Connecting the data points by lines helps emphasize any change over time.

time series Measurements of a variable taken at regular intervals over time form a **time series**. Plots against time can reveal the main features of a time series. As with distributions, look first for overall patterns and then for striking deviations from those patterns. Here are some types of overall patterns to look for in a time series.

SEASONAL VARIATION AND TREND

A pattern in a time series that repeats itself at known regular intervals of time is called **seasonal variation**.
A **trend** in a time series is a persistent, long-term rise or fall.

EXAMPLE 1.7

The price of oranges

Figure 1.7 is a time plot of the average price of fresh oranges over the decade from 1991 to 2000.¹⁰ This information is collected each month as part of the government's reporting of retail prices. The monthly consumer price index is the most publicized product of this effort. The price is presented as an **index number**. That is, the price scale gives the price as a percent of the average price of oranges in the years 1982 to 1984. This is indicated by the legend "1982–84 = 100." The first value is 205.7 for January 1991, so at that time oranges cost about 206% of their 1982 to 1984 average price. The index number is based on the retail price of oranges at many stores in all parts of the country. The price of oranges at a single store will drop when the manager decides to advertise a sale on oranges, and it will rise when the sale ends. The index number is a nationwide average price that is less variable than the price at one store.

Figure 1.7 shows an upward *trend* in the price of oranges starting in 1993. Statistical software can draw a line on the time plot that represents the trend.

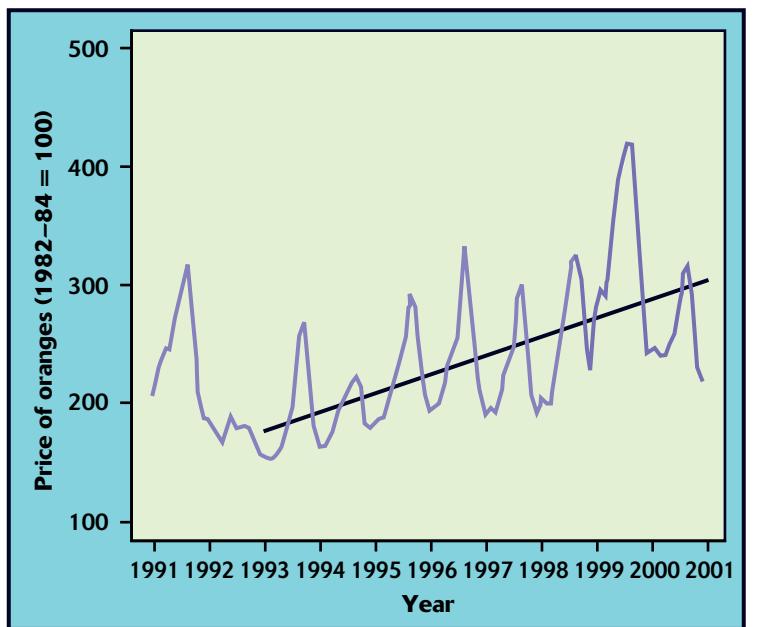


FIGURE 1.7 Time plot of an index number (1982–84 = 100) for the average retail price of oranges each month from 1991 through 2000, for Example 1.7.

Such a line appears in Figure 1.7. Superimposed on this trend is a strong *seasonal variation*, a regular rise and fall that recurs each year. Orange prices are usually highest in August or September, when the supply is lowest. Prices then fall in anticipation of the harvest and are lowest in January or February, when the harvest is complete and oranges are plentiful. The large spikes in 1991 and 1999 may be deviations from the overall trend and seasonal variation, perhaps due to poor harvests.



Because many economic time series show strong seasonal variation, government agencies often adjust for this variation before releasing economic data. The data are then said to be **seasonally adjusted**. Seasonal adjustment helps avoid misinterpretation. A rise in the unemployment rate from December to January, for example, does not mean that the economy is slipping. Unemployment almost always rises in January as temporary holiday help is laid off and outdoor employment in the north drops because of bad weather. The seasonally adjusted unemployment rate reports an increase only if unemployment rises more than normal from December to January.

EXAMPLE 1.8**Unemployment rates**

We have looked at the distribution of state unemployment rates for December 2000. These are called **cross-sectional data** because they concern a group of individuals (the states) at one time. Time series data can put cross-sectional data



CASE 1.1



FIGURE 1.8 Time plot of the monthly national unemployment rate from January 1990 to August 2001, for Example 1.8.

cross-sectional data

such as the December 2000 unemployment rates into a context. Figure 1.8 is a time plot of the monthly national unemployment rate from 1990 to August 2001. The data are seasonally adjusted, so that long-term trends are easy to see.

We now see that December 2000 (marked by the broken line) was a time of low unemployment, near the bottom of a decline from a peak of 7.8% in June 1992. As the economic boom of the 1990s ended, unemployment rates began to turn up again.

APPLY YOUR KNOWLEDGE

- 1.12 Yields of Treasury bills. Treasury bills are short-term borrowing by the U.S. government. They are important in financial theory because the interest rate for Treasury bills is a “risk-free rate” that says what return investors can get while taking (almost) no risk. More risky investments should in theory offer higher returns in the long run. Here are the annual returns on Treasury bills from 1970 to 2000.¹¹

Year	Rate	Year	Rate	Year	Rate	Year	Rate
1970	6.52	1978	7.19	1986	6.16	1994	3.91
1971	4.39	1979	10.38	1987	5.47	1995	5.60
1972	3.84	1980	11.26	1988	6.36	1996	5.20
1973	6.93	1981	14.72	1989	8.38	1997	5.25
1974	8.01	1982	10.53	1990	7.84	1998	4.85
1975	5.80	1983	8.80	1991	5.60	1999	4.69
1976	5.08	1984	9.84	1992	3.50	2000	5.69
1977	5.13	1985	7.72	1993	2.90		

- cycles*
- (a) Make a time plot of the returns paid by Treasury bills in these years.
 - (b) Interest rates, like many economic variables, show **cycles**, clear but irregular up-and-down movements. In which years did the interest rate cycle reach temporary peaks?
 - (c) A time plot may show a consistent trend underneath cycles. When did interest rates reach their overall peak during these years? Has there been a general trend downward since that year?

SECTION 1.1 SUMMARY

- A data set contains information on a number of **individuals**. Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables**. A variable describes some characteristic of an individual, such as a person's height, gender, or salary.
- Some variables are **categorical** and others are **quantitative**. A categorical variable places each individual into a category, like male or female. A quantitative variable has numerical values that measure some characteristic of each individual, like height in centimeters or salary in dollars per year.
- **Exploratory data analysis** uses graphs and numerical summaries to describe the variables in a data set and the relations among them.
- The **distribution** of a variable describes what values the variable takes and how often it takes these values.
- To describe a distribution, begin with a graph. **Bar graphs** and **pie charts** describe the distribution of a categorical variable, and **Pareto charts** identify the most important few categories for a categorical variable. **Histograms** and **stemplots** graph the distributions of quantitative variables.
- When examining any graph, look for an **overall pattern** and for notable **deviations** from the pattern.
- **Shape**, **center**, and **spread** describe the overall pattern of a distribution. Some distributions have simple shapes, such as **symmetric** and **skewed**. Not all distributions have a simple overall shape, especially when there are few observations.
- **Outliers** are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.
- When observations on a variable are taken over time, make a **time plot** that graphs time horizontally and the values of the **time series** vertically. A time plot can reveal **trends**, **seasonal variation**, or other changes over time.

SECTION 1.1 EXERCISES

- 1.13 Mutual funds. The following is a small part of a data set that describes mutual funds available to the public.

Fund	Category	Net assets (\$millions)	Year-to-date return	Largest holding
:				
Fidelity Low-Priced Stock	Small value	6,189	4.56%	Dallas Semi-conductor
Price Inter-national Stock	International stock	9,745	-0.45%	Vodafone
Vanguard 500 Index	Large blend	89,394	3.45%	General Electric
:				

- (a) What individuals does this data set describe?
- (b) In addition to the fund's name, how many variables does the data set contain? Which of these variables are categorical and which are quantitative?
- (c) What are the units of measurement for each of the quantitative variables?
- 1.14 **Insurance: cause of death.** Cause of death is an important issue for life insurance companies—especially identifying the most likely causes of death for various demographic subgroups of the population. The number of deaths among persons aged 15 to 24 years in the United States in 1997 due to the seven leading causes of death for this age group were accidents, 12,958; homicide, 5793; suicide, 4146; cancer, 1583; heart disease, 1013; congenital defects, 383; AIDS, 276.¹²
- (a) Make a bar graph to display these data.
- (b) What additional information do you need to make a pie chart?
- 1.15 **Location of a new facility.** A company deciding where to build a new facility will rank potential locations in terms of how desirable it is to operate a business in each location. Describe five variables that you would measure for each location if you were designing a study to help your company rank locations. Give reasons for each of your choices.
- 1.16 **Hurricanes.** Hurricanes cause a great deal of property damage and personal injury when they hit populated landmasses. A substantial portion of hurricane loss is paid for by insurance companies in the form of paying for lost property and medical bills. For this reason, insurance companies study the number and severity of the hurricanes that can potentially cause them loss. The histogram in Figure 1.9 shows the number of hurricanes reaching the east coast of the United States each year over a 70-year period.¹³ Give a brief description of the overall shape of this distribution. About where does the center of the distribution lie?
- 1.17 **Left skew.** Sketch a histogram for a distribution that is skewed to the left. Suppose that you and your friends emptied your pockets of coins and recorded the year marked on each coin. The distribution of dates would be skewed to the left. Explain why.
- 1.18 **The changing age distribution of the United States.** The distribution of the ages of a nation's population has a strong influence on economic and social

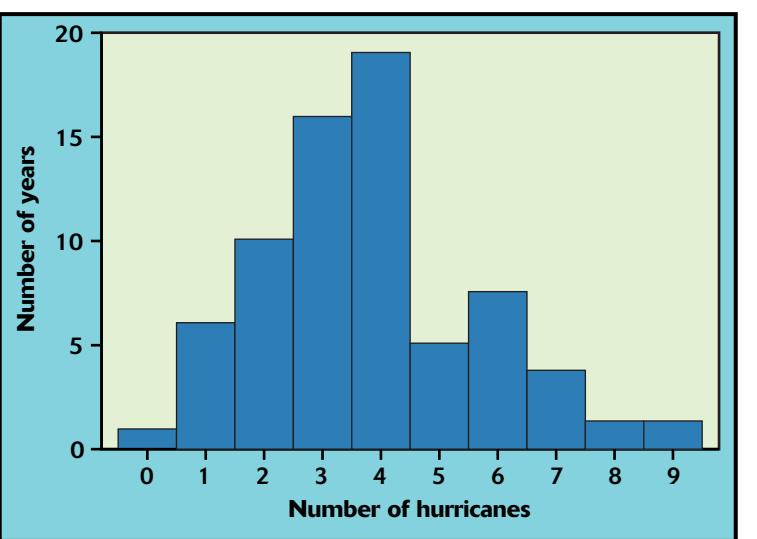


FIGURE 1.9 The distribution of the annual number of hurricanes on the U.S. east coast over a 70-year period, for Exercise 1.16.

conditions. Table 1.4 shows the age distribution of U.S. residents in 1950 and 2075, in millions of people. The 1950 data come from that year's census, while the 2075 data are projections made by the Census Bureau.

- (a) Because the total population in 2075 is much larger than the 1950 population, comparing percents in each age group is clearer than comparing counts. Make a table of the percent of the total population in each age group for both 1950 and 2075.
- (b) Make a histogram with vertical scale in percents of the 1950 age distribution. Describe the main features of the distribution. In particular, look at the percent of children relative to the rest of the population.

TABLE 1.4 Age distribution in the United States, 1950 and 2075 (in millions of persons)

Age group	1950	2075
Under 10 years	29.3	53.3
10–19 years	21.8	53.2
20–29 years	24.0	51.2
30–39 years	22.8	50.5
40–49 years	19.3	47.5
50–59 years	15.5	44.8
60–69 years	11.0	40.7
70–79 years	5.5	30.9
80–89 years	1.6	21.7
90–99 years	0.1	8.8
100–109 years	—	1.1
Total	151.1	403.7

- (c) Make a histogram with vertical scale in percents of the projected age distribution for the year 2075. Use the same scales as in (b) for easy comparison. What are the most important changes in the U.S. age distribution projected for the years between 1950 and 2075?
- 1.19 Reliability of household appliances.** Always ask whether a particular variable is really a suitable measure for your purpose. You are writing an article for a consumer magazine based on a survey of the magazine's readers on the reliability of their household appliances. Of 13,376 readers who reported owning Brand A dishwashers, 2942 required a service call during the past year. Only 192 service calls were reported by the 480 readers who owned Brand B dishwashers.
- Why is the count of service calls (2942 versus 192) not a good measure of the reliability of these two brands of dishwashers?
 - Use the information given to calculate a suitable measure of reliability. What do you conclude about the reliability of Brand A and Brand B?
- 1.20 Bear markets.** Investors speak of a “bear market” when stock prices drop substantially. Table 1.5 gives data on all declines of at least 10% in the Standard & Poor’s 500-stock index between 1940 and 1999. The data show how far the index fell from its peak and how long the decline in stock prices lasted.
- Make a stemplot of the percent declines in stock prices during these bear markets. Make a second stemplot, splitting the stems. Which graph do you prefer? Why?
 - The shape of this distribution is irregular, but we could describe it as somewhat skewed. Is the distribution skewed to the right or to the left?
 - Describe the center and spread of the data. What would you tell an investor about how far stocks fall in a bear market?
- 1.21 “The Fortune 500.”** Each year *Fortune* magazine lists the top 500 companies in the United States, ranked according to their total annual sales in dollars. Describe three other variables that could reasonably be used to measure the “size” of a company.

TABLE 1.5 Size and length of bear markets

Year	Decline (percent)	Duration (months)	Year	Decline (percent)	Duration (months)
1940–1942	42	28	1966	22	8
1946	27	5	1968–1970	36	18
1950	14	1	1973–1974	48	21
1953	15	8	1981–1982	26	19
1955	10	1	1983–1984	14	10
1956–1957	22	15	1987	34	3
1959–1960	14	15	1990	20	3
1962	26	6			

- 1.22 **Salary distributions in a factory.** A manufacturing company is reviewing the salaries of its full-time employees below the executive level at a large plant. The clerical staff is almost entirely female, while a majority of the production workers and technical staff is male. As a result, the distributions of salaries for male and female employees may be quite different. Table 1.6 gives the counts and percents of women and men in each salary class. Make histograms from these data, choosing the type that is most appropriate for comparing the two distributions. Then describe the overall shape of each salary distribution and the chief differences between them.
- 1.23 **The cost of Internet access.** How much do users pay for Internet service? Here are the monthly fees (in dollars) paid by a random sample of 50 users of commercial Internet service providers in August 2000.¹⁴

20	40	22	22	21	21	20	10	20	20
20	13	18	50	20	18	15	8	22	25
22	10	20	22	22	21	15	23	30	12
9	20	40	22	29	19	15	20	20	20
20	15	19	21	14	22	21	35	20	22

Make a stemplot of these data. Briefly describe the pattern you see. About how much do you think America Online and its larger competitors were charging in August 2000? Which members of the sample may have been early adopters of fast access via cable modems or DSL lines?

TABLE 1.6**Salary distributions of female and male workers in a large factory**

Salary (\$1000)	Women		Men	
	Number	%	Number	%
10–15	89	11.8	26	1.1
15–20	192	25.4	221	9.0
20–25	236	31.2	677	27.9
25–30	111	14.7	823	33.6
30–35	86	11.4	365	14.9
35–40	25	3.3	182	7.4
40–45	11	1.5	91	3.7
45–50	3	0.4	33	1.4
50–55	2	0.3	19	0.8
55–60	0	0.0	11	0.4
60–65	0	0.0	0	0.0
65–70	1	0.1	3	0.1
Total	756	100.1	2451	100.0

- 1.24 Architects and engineers. Table 1.3 (page 14) gives the numbers of architects and engineers employed by Indianapolis architectural firms. A **back-to-back stemplot**

*back-to-back
stemplot*

helps us compare these two distributions. Write the stems as usual, but with a vertical line both to their left and to their right. On the right, put leaves for architects. On the left, put the leaves for engineers. Arrange the leaves on each stem in increasing order out from the stem. Now write a brief comparison of the distributions.

- 1.25 Watch those scales! The impression that a time plot gives depends on the scales you use on the two axes. If you stretch the vertical axis and compress the time axis, change appears to be more rapid. Compressing the vertical axis and stretching the time axis make change appear slower. Make two time plots of the 1970–1981 data in Exercise 1.12 (page 22), one that makes rates appear to increase very rapidly and one that suggests only a modest increase. The moral of this exercise is: pay close attention to the scales when you look at a time plot.

Table 1.7 presents government data that describe some aspects of the prosperity and quality of life in the 50 states and the District of Columbia. Study of a data set with many variables begins by examining each variable by itself. Exercises 1.26 to 1.28 concern the data in Table 1.7.

- 1.26 Population of the states. Make a stemplot of the population of the states. Briefly describe the shape, center, and spread of the distribution of population. Explain why the shape of the distribution is not surprising. Are there any states that you consider outliers?

- 1.27 Income in the states. Table 1.7 contains two measures of “average income.” Mean personal income is the total income of all people living in a state divided by the state’s population. Median household income is the midpoint of the distribution of incomes of all households in the state. (A household consists of all people living at the same address, whether or not they are related to each other.)

- Explain why you expect household income to be larger than personal income.
- Round the income data to the nearest hundred dollars and write them in units of thousands of dollars. For example, Alabama’s mean personal income \$22,104 becomes 22.1 and Alaska’s mean personal income becomes 26.5. Make a back-to-back stemplot (see Exercise 1.24) of personal and household income.
- Write a brief description of the shape of each distribution and also a brief comparison of the two distributions.

- 1.28 Where are the doctors? Table 1.7 gives the number of medical doctors per 100,000 people in each state.

- Why is the number of doctors per 100,000 people a better measure of the availability of health care than a simple count of the number of doctors in a state?
- Make a graph that displays the distribution of M.D.’s per 100,000 people. Write a brief description of the distribution. Are there any outliers? If so, can you explain them?

TABLE 1.7 Data for the states

State	2000 Census population	Mean personal income	Median household income	Percent poverty	M.D.'s per 100,000	Violent crimes per 100,000
Alabama	4,447,100	22,104	36,266	14.5	194	565
Alaska	626,932	26,468	50,692	9.4	160	701
Arizona	5,130,632	23,772	37,090	16.6	200	624
Arkansas	2,673,400	20,974	27,665	14.8	185	527
California	33,871,648	28,352	40,934	15.4	244	798
Colorado	4,301,261	29,542	46,599	9.2	234	363
Connecticut	3,405,565	38,759	46,508	9.5	344	391
Delaware	783,600	30,734	41,458	10.3	230	678
District of Columbia	572,059	38,429	33,433	22.3	702	2024
Florida	15,982,378	26,650	34,909	13.1	232	1024
Georgia	8,186,453	25,793	38,665	13.6	204	607
Hawaii	1,211,537	26,944	40,827	10.9	252	278
Idaho	1,293,953	21,731	36,680	13.0	150	257
Illinois	12,419,293	29,764	43,178	10.1	253	861
Indiana	6,080,485	24,967	39,731	9.4	192	515
Iowa	2,926,324	24,664	37,019	9.1	171	310
Kansas	2,688,418	25,752	36,711	9.6	202	409
Kentucky	4,041,769	22,171	36,252	13.5	205	317
Louisiana	4,468,976	22,006	31,735	19.1	239	856
Maine	1,274,923	23,661	35,640	10.4	214	121
Maryland	5,296,486	30,868	50,016	7.2	362	847
Massachusetts	6,349,097	33,809	42,345	8.7	402	644
Michigan	9,938,444	26,655	41,821	11.0	218	590
Minnesota	4,919,479	28,359	47,926	10.4	247	338
Mississippi	2,844,658	19,544	29,120	17.6	156	469
Missouri	5,595,211	25,181	40,201	9.8	225	577
Montana	902,195	20,795	31,577	16.6	188	132
Nebraska	1,711,263	25,519	36,413	12.3	213	438
Nevada	1,998,257	28,040	39,756	10.6	169	799
New Hampshire	1,235,786	29,919	44,958	9.8	230	113
New Jersey	8,414,350	34,985	49,826	8.6	287	493
New Mexico	1,819,046	20,551	31,543	20.4	209	853
New York	18,976,457	32,714	37,394	16.7	375	689
North Carolina	8,049,313	24,778	35,838	14.0	225	607
North Dakota	642,200	22,344	30,304	15.1	219	87
Ohio	11,353,140	25,910	38,925	11.2	230	435

(continued)

TABLE 1.7 Data for the states (continued)

State	2000 Census population	Mean personal income	Median household income	Percent poverty	M.D.'s per 100,000	Violent crimes per 100,000
Oklahoma	3,450,654	21,722	33,727	14.1	166	560
Oregon	3,421,399	25,530	39,067	15.0	221	444
Pennsylvania	12,281,054	27,619	39,015	11.2	282	442
Rhode Island	1,048,319	27,624	40,686	11.6	324	334
South Carolina	4,012,012	21,967	33,267	13.7	201	990
South Dakota	754,844	22,797	32,786	10.8	177	197
Tennessee	5,689,283	24,286	34,091	13.4	242	790
Texas	20,851,820	25,728	35,783	15.0	196	603
Utah	2,233,169	21,667	44,299	9.0	197	334
Vermont	608,827	24,922	39,372	9.9	288	120
Virginia	7,078,515	28,230	43,354	8.8	233	345
Washington	5,894,121	28,824	47,421	8.9	229	441
West Virginia	1,808,344	19,960	26,704	17.8	210	219
Wisconsin	5,363,675	25,853	41,327	8.8	224	271
Wyoming	493,782	23,882	35,250	10.6	167	255

1.2 Describing Distributions with Numbers



EARNINGS OF HOURLY BANK WORKERS

Banks employ many workers paid by the hour, such as tellers and data clerks. A large bank, which we will call simply National Bank to preserve confidentiality, has been accused of discrimination in paying its hourly workers. In the course of an internal investigation into the charges, the bank collected data on all 1745 hourly workers listed in its personnel records. The data set is *hourly.dat*, described in detail in the Data Appendix at the back of the book. Because the data set is so large, we will use randomly selected samples of full-time hourly workers in four race-and-gender groups to illustrate statistical analysis before applying the methods we develop to the complete data set. Table 1.8 presents the annual earnings for the workers in our samples.

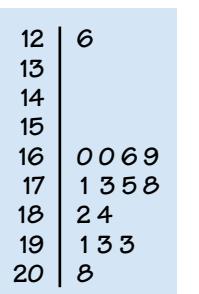


The stemplot in Figure 1.10 shows us the *shape*, *center*, and *spread* of the earnings of the 15 black females in Table 1.8. The stems are thousands of dollars and the leaves are hundreds, rounded to the nearest hundred. As is often the case when there are few observations, the shape of the distribution is irregular. The earnings are quite tightly clustered except

TABLE 1.8 Annual earnings of hourly workers at National Bank

Black females	Black males	White females	White males
\$16,015	\$18,365	\$25,249	\$15,100
17,516	17,755	19,029	22,346
17,274	16,890	17,233	22,049
16,555	17,147	26,606	26,970
20,788	18,402	28,346	16,411
19,312	20,972	31,176	19,268
17,124	24,750	18,863	28,336
18,405	16,576	15,904	19,007
19,090	16,853	22,477	22,078
12,641	21,565	19,102	19,977
17,813	29,347	18,002	17,194
18,206	19,028	21,596	30,383
19,338		26,885	18,364
15,953		24,780	18,245
16,904		14,698	23,531
		19,308	
		17,576	
		24,497	
		20,612	
		17,757	

for one low outlier. The midpoint of the earnings is roughly \$17,500. The spread, omitting the outlier, is from about \$16,000 to about \$20,800. Shape, center, and spread provide a good description of the overall pattern of any distribution for a quantitative variable. In this section, we will learn specific ways to use numbers to measure the center and the spread of a distribution. The numbers, like the graphs of Section 1.1, are aids to understanding the data, not “the answer” in themselves.

**FIGURE 1.10** Stemplot of the annual earnings of 15 black female hourly workers at National Bank.

Measuring center: the mean

A description of a distribution almost always includes a measure of its center or average. The most common measure of center is the ordinary arithmetic average, or *mean*.

THE MEAN \bar{x}

To find the **mean** of a set of observations, add their values and divide by the number of observations. If the n observations are x_1, x_2, \dots, x_n , their mean is

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

or in more compact notation,

$$\bar{x} = \frac{1}{n} \sum x_i$$

The \sum (capital Greek sigma) in the formula for the mean is short for “add them all up.” The subscripts on the observations x_i are just a way of keeping the n observations distinct. They do not necessarily indicate order or any other special facts about the data. The bar over the x indicates the mean of all the x -values. Pronounce the mean \bar{x} as “x-bar.” This notation is very common. When writers who are discussing data use \bar{x} or \bar{y} , they are talking about a mean.

EXAMPLE 1.9

Mean earnings of black female workers

The mean amount earned by the 15 black female bank workers in Table 1.8 is



$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{16,015 + 17,516 + \cdots + 16,904}{15} \\ &= \frac{262,934}{15} = \$17,528.93\end{aligned}$$

In practice, you can key the data into your calculator and hit the mean key. You don't have to actually add and divide. But you should know that this is what the calculator is doing.

The lowest-paid worker earned \$12,641. Use your calculator to check that the mean income of the other 14 workers *excluding this outlier* is $\bar{x} = \$17,878.07$. The single outlier reduces the mean by almost \$350.



Example 1.9 illustrates an important fact about the mean as a measure of center: it is sensitive to the influence of one or more extreme observations. These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. Because the mean cannot resist the

resistant measure

influence of extreme observations, we say that it is *not* a **resistant measure** of center.

APPLY YOUR KNOWLEDGE

- 1.29 **Bank workers.** Find the mean earnings of the remaining three groups in Table 1.8. Does comparing the four means suggest that National Bank pays male hourly workers more than females or white workers more than blacks? (Of course, detailed investigation of such things as job type and seniority is needed before we claim discrimination.)



- 1.30 **Supermarket shoppers.** Here are the amounts spent (in dollars) by 50 consecutive shoppers at a supermarket, arranged in increasing order:

3.11	8.88	9.26	10.81	12.69	13.78	15.23	15.62	17.00
17.39	18.36	18.43	19.27	19.50	19.54	20.16	20.59	22.22
23.04	24.47	24.58	25.13	26.24	26.26	27.65	28.06	28.08
28.38	32.03	34.98	36.37	38.64	39.16	41.02	42.97	44.08
44.67	45.40	46.69	48.65	50.39	52.75	54.80	59.07	61.22
70.32	82.70	85.76	86.37	93.34				

- (a) Find the mean amount spent from the formula for the mean. Then enter the data into your calculator or software and use the mean function to obtain the mean. Verify that you get the same result.
- (b) A stemplot (Exercise 1.11) suggests that the largest four values may be a cluster of outliers. Find the mean for the 46 observations that remain when you drop these outliers. How do the outliers change the mean?

Measuring center: the median

In Section 1.1, we used the midpoint of a distribution as an informal measure of center. The *median* is the formal version of the midpoint, with a specific rule for calculation.

THE MEDIAN M

The **median M** is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $(n + 1)/2$ observations up from the bottom of the list.
3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. The location of the median is again $(n + 1)/2$ from the bottom of the list.

Note that the formula $(n + 1)/2$ does *not* give the median, just the location of the median in the ordered list. Medians require little arithmetic, so they are easy to find by hand for small sets of data. Arranging even a moderate number of observations in order is very tedious, however, so that finding the

median by hand for larger sets of data is unpleasant. Even simple calculators have an \bar{x} button, but you will need software or a graphing calculator to automate finding the median.

EXAMPLE 1.10

Median earnings of black female workers

To find the median earnings of our 15 black female bank workers, first arrange the data in order from smallest to largest:

12641 15953 16015 16555 16904 17124 17274 **17516** 17813
18206 18405 19090 19312 19338 20788

The count of observations $n = 15$ is odd. The median, then, is the center observation in the ordered list. The center observation is the 8th observation as given by our formula for locating the median,

$$\text{location of } M = \frac{n+1}{2} = \frac{15}{2} = 8$$

The median is the bold \$17,516 in the list. This number has 7 observations to its left and 7 observations to its right.

How much does the outlier \$12,641 affect the median? Drop it from the list and find the median for the remaining $n = 14$ black females. The rule for locating the median in the list gives

$$\text{location of } M = \frac{n+1}{2} = \frac{15}{2} = 7.5$$

The location 7.5 means “halfway between the 7th and 8th observations in the ordered list.”

15953 16015 16555 16904 17124 17274 **17516** 17813 18206
18405 19090 19312 19338 20788

With an even number of observations, there is no center observation. The bold 7th and 8th observations are the middle pair in the list. There are 6 observations on either side of this pair. The median is halfway between \$17,516 and \$17,813. So

$$M = \frac{17,516 + 17,813}{2} = 17,664.50$$


Comparing the mean and the median

Examples 1.9 and 1.10 illustrate an important difference between the mean and the median. The low outlier pulls the mean earnings down by \$350. The median drops by only \$148 when we add the low outlier to our list. The median is more *resistant* than the mean. If the lowest-paid worker had earned nothing, the median for all 15 workers would remain \$17,516. The smallest observation just counts as one observation below the center, no matter how far below the center it lies. The mean uses the actual value of each observation and so will chase a single small observation downward.

The best way to compare the response of the mean and median to extreme observations is to use an interactive applet that allows you to place points on a line and then drag them with your computer's mouse.



Exercises 1.42 to 1.44 use the *Mean and Median* applet on the Web site for this book, www.whfreeman.com/pbs, to compare mean and median.

The mean and median of a symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is farther out in the long tail than is the median. For example, the distribution of house prices is strongly skewed to the right. There are many moderately priced houses and a few very expensive mansions. The few expensive houses pull the mean up but do not affect the median. The mean price of existing houses sold in 2000 was \$176,200, but the median price for these same houses was only \$139,000. Reports about house prices, incomes, and other strongly skewed distributions usually give the median (“midpoint”) rather than the mean (“arithmetic average”). However, if you are a tax assessor interested in the total value of houses in your area, use the mean. The total is the mean times the number of houses, but it has no connection with the median. The mean and median measure center in different ways, and both are useful.

APPLY YOUR KNOWLEDGE

- 1.31 **Bank workers.** Find the median earnings of the remaining three groups of workers in Table 1.8. Do your preliminary conclusions from comparing the medians differ from the results of comparing the mean earnings in Exercise 1.29? (Because the median is not distorted by outliers, we might prefer to base an initial look at possible inequity on the median rather than the mean.)
- 1.32 **Private consumption.** The success of companies expanding to developing regions of the world depends in part on the increase in private consumption in those regions. Here are World Bank data on the growth of per capita private consumption (percent per year) for the period 1990 to 1997 in countries in Asia (not including Japan):



Country	Growth
Bangladesh	2.3
China	8.8
Hong Kong, China	3.9
India	4.1
Indonesia	6.4
Korea (South)	5.9
Malaysia	4.2
Pakistan	2.9
Philippines	1.3
Singapore	5.1
Thailand	5.6
Vietnam	6.2

- Make a stemplot of the data. Note the high outlier.
- Find the mean and median growth rates. How does the outlier explain the difference between your two results?
- Find the mean and median growth rates without the outlier. How does comparing your results in (b) and (c) illustrate the resistance of the median and the lack of resistance of the mean?

- 1.33 The **richest 1%**. The distribution of individual incomes in the United States is strongly skewed to the right. In 1997, the mean and median incomes of the top 1% of Americans were \$330,000 and \$675,000. Which of these numbers is the mean and which is the median? Explain your reasoning.

Measuring spread: the quartiles

A measure of center alone can be misleading. Two nations with the same median household income are very different if one has extremes of wealth and poverty and the other has little variation among households. A drug with the correct mean concentration of active ingredient is dangerous if some batches are much too high and others much too low. We are interested in the *spread* or *variability* of incomes and drug potencies as well as their centers. **The simplest useful numerical description of a distribution consists of both a measure of center and a measure of spread.**

One way to measure spread is to give the smallest and largest observations. For example, the earnings of the black female bank workers in Table 1.8 range from \$12,641 to \$20,788. These single observations show the full spread of the data, but they may be outliers. We can improve our description of spread by also giving several percentiles. The *p*th percentile of a distribution is the value such that *p* percent of the observations fall at or below it. The median is just the 50th percentile, so the use of percentiles to report spread is particularly appropriate when the median is our measure of center. The most commonly used percentiles other than the median are the *quartiles*. The first quartile is the 25th percentile, and the third quartile is the 75th percentile. That is, the first and third quartiles show the spread of the middle half of the data. (The second quartile is the median itself.) To calculate a percentile, arrange the observations in increasing order and count up the required percent from the bottom of the list. Our definition of percentiles is a bit inexact because there is not always a value with exactly *p* percent of the data at or below it. We will be content to take the nearest observation for most percentiles, but the quartiles are important enough to require an exact recipe. The rule for calculating the quartiles uses the rule for the median.

*p*th percentile

THE QUARTILES Q_1 and Q_3

To calculate the quartiles:

1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.
2. The **first quartile Q_1** is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
3. The **third quartile Q_3** is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Here is an example that shows how the rules for the quartiles work for both odd and even numbers of observations.

EXAMPLE 1.11**Finding the quartiles**

The earnings of the 15 black female bank workers in our sample of National Bank hourly employees (arranged from smallest to largest) are

12641 15953 16015 16555 16904 17124 17274 **17516**
17813 18206 18405 19090 19312 19338 20788

The count of observations $n = 15$ is odd, so the median is the 8th observation in the list, the bold 17,516. This number has 7 observations to its left and 7 observations to its right. The first quartile is the median of the first 7 observations, and the third quartile is the median of the last 7 observations. Check that $Q_1 = 16,555$ and $Q_3 = 19,090$.

Notice that the quartiles are resistant. For example, Q_1 would have the same value if the low outlier 12,641 were 1000.

Table 1.8 also reports the annual earnings of 12 black male hourly workers. Here they are, arranged from smallest to largest:

16576 16853 16890 17147 17755 18365 | 18402
19028 20972 21565 24750 29347

The median is halfway between the 6th and 7th entries in the ordered list. We have marked its location by |. The value of the median is

$$M = \frac{18,365 + 18,402}{2} = 18,383.5$$

The first quartile is the median of the 6 observations to the left of the | in the list, and the third quartile is the median of the 6 observations to the right. Check that $Q_1 = 17,018.5$ and $Q_3 = 21,268.5$.

We find other percentiles more informally. For example, we take the 90th percentile of the earnings of the 12 black male workers to be the 11th in the ordered list, because $0.90 \times 12 = 10.8$, which we round to 11. The 90th percentile is therefore \$24,750.



Be careful when several observations take the same numerical value. Write down all of the observations and apply the rules just as if they all had distinct values.

Some software packages use a slightly different rule to find the quartiles, so computer results may differ a bit from your own work. Don't worry about this. The differences will always be too small to be important.

The five-number summary and boxplots

The smallest and largest observations tell us little about the distribution as a whole, but they give information about the tails of the distribution that is missing if we know only Q_1 , M , and Q_3 . To get a quick summary of both center and spread, combine all five numbers. The result is the *five-number summary* and a graph based on it.

THE FIVE-NUMBER SUMMARY AND BOXPLOTS

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

Minimum Q_1 M Q_3 Maximum

A **boxplot** is a graph of the five-number summary.

- A central box spans the quartiles.
- A line in the box marks the median.
- Lines extend from the box out to the smallest and largest observations.

Boxplots are most useful for side-by-side comparison of several distributions.

You can draw boxplots either horizontally or vertically. Be sure to include a numerical scale in the graph. When you look at a boxplot, first locate the median, which marks the center of the distribution. Then look at the spread. The quartiles show the spread of the middle half of the data, and the extremes (the smallest and largest observations) show the spread of the entire data set. We now have the tools for a preliminary examination of the National Bank earnings data.

EXAMPLE 1.12

Adverse impact at National Bank



The complete data set *hourly.dat* contains information about 1276 full-time hourly workers at National Bank. Among these workers are 315 black females, 697 white females, 81 black males, and 162 white males. Statistical software gives us the five-number summaries for the earnings of each group:

	A	B	C	D	E	F	G	H
1	Earnings	N	Min	25%	Median	75%	Max	
2	Black female	315	12641	17145.5	18354	20363	33703	
3	White female	697	12641	18492	21492	25583	37513	
4	Black male	81	14731	17155	18400	20972	32678	
5	White male	162	12605	18343.25	22029.5	25302	36608	
6								

The boxplots in Figure 1.11 compare the four distributions of earnings.

Concentrate on the box spanning the quartiles, because the extremes in a large data set are often outliers. We see at once that whites, both female and male, tend to earn more than blacks. The median income of white hourly workers is greater than the third quartile of black income. There are only minor differences between females and males within each race. In legal terms, the data demonstrate that National Bank's pay policies have an *adverse impact* on blacks. The bank must now investigate whether there is a legitimate reason for the black/white earnings difference, such as hours worked, seniority, or job classification.

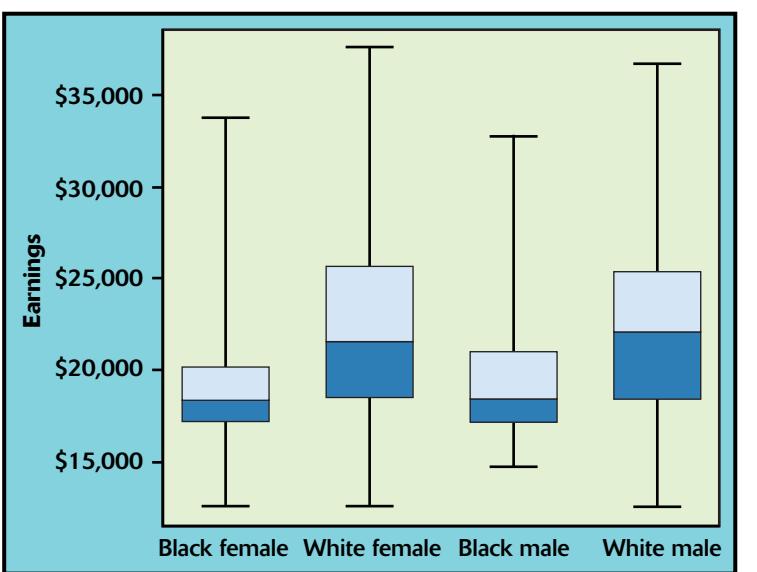


FIGURE 1.11 Side-by-side boxplots comparing the earnings of four groups of hourly workers at National Bank, for Example 1.12.

Figure 1.11 illustrates the power of simple statistical descriptions to make sense of a large amount of information. Although of less interest in this case, boxplots also suggest the symmetry or skewness of a distribution. In a symmetric distribution, the first and third quartiles are equally distant from the median. In most distributions that are skewed to the right, on the other hand, the third quartile will be farther above the median than the first quartile is below it. The extremes behave the same way, but remember that they are just single observations and may say little about the distribution as a whole, especially if either is an outlier. Figure 1.11 suggests that the earnings distributions are somewhat right-skewed.

APPLY YOUR KNOWLEDGE

- 1.34 **Bank workers.** Complete the calculation of the five-number summaries of earnings for the four sample groups of workers in Table 1.8 (page 31). Make side-by-side boxplots to compare the groups. Do these small samples faithfully reflect the nature of the complete data set as shown in Figure 1.11?
- 1.35 **Private consumption.** Exercise 1.32 (page 35) gives the growth of per capita private consumption for 12 Asian countries. The World Bank data also looked at 13 Eastern European countries. Here is the growth of per capita private consumption for each of these countries:



CASE 1.12

Country	Growth	Country	Growth
Albania	6.0	Latvia	1.0
Belarus	-5.2	Poland	4.9
Bulgaria	-1.0	Romania	1.4
Croatia	3.5	Russian Federation	7.0
Czech Republic	3.2	Slovenia	3.7
Estonia	-1.7	Ukraine	-12.1
Hungary	-1.5		

TABLE 1.9 Average annual wages for managerial and administrative occupations, 1998

OCC Code	OCC Title	Wage (\$1000)
15032	Lawn Service Managers	28
15026	Food Service and Lodging Managers	29
15031	Nursery and Greenhouse Managers	30
19002	Public Admin. Chief Execs., Legislators, and Gen. Admin.	32
15011	Property and Real Estate Managers and Administrators	36
15002	Postmasters and Mail Superintendents	46
13008	Purchasing Managers	47
13014	Administrative Services Managers	49
15017	Construction Managers	52
13005	Personnel, Training, and Labor Relations Managers	52
15008	Medicine and Health Services Managers	52
19999	All Other Managers and Administrators	53
15023	Communications, Transportation, and Utilities Operations	54
15005	Education Administrators	58
15014	Industrial Production Managers	58
13002	Financial Managers	59
13011	Marketing, Advertising, and Public Relations Managers	60
19005	General Managers and Top Executives	63
15021	Mining, Quarrying, and Oil and Gas Well Drilling Managers	63
13017	Engineering, Mathematical, and Natural Sciences Managers	72

- (a) Find the five-number summary for each group of countries (Asian and Eastern European).
- (b) Make side-by-side boxplots to compare the growth of per capita private consumption for the two groups of countries. What do you conclude?
- 1.36 **Managerial and administrative employment.** The Bureau of Labor Statistics collects data on employment and wages for various occupations in the United States. One major division of occupations is Managerial and Administrative Occupations (OCC Code 10000). Table 1.9 displays the 1998 annual wage averages (rounded to the nearest thousand) for this occupational category.¹⁵
- (a) Make a stemplot of the distribution of wages for these managerial and administrative occupations.
- (b) From the shape of your stemplot, do you expect the median to be much less than the mean, about the same as the mean, or much greater than the mean?
- (c) Find the mean and the five-number summary of the data set. Verify your expectation about the median compared to the mean.
- (d) What is the range of the middle half of the average annual wages of people employed in managerial and administrative occupations?

Measuring spread: the standard deviation

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the mean to

measure center and the *standard deviation* to measure spread. The standard deviation measures spread by looking at how far the observations are from their mean.

THE STANDARD DEVIATION s

The **variance** s^2 of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations x_1, x_2, \dots, x_n is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

or, more compactly,

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

The **standard deviation** s is the square root of the variance s^2 :

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

In practice, use software or your calculator to obtain the standard deviation from keyed-in data. Doing an example step-by-step will help you understand how the variance and standard deviation work, however.

EXAMPLE 1.13

Calculating the standard deviation

Planning to be a lawyer or other legal professional? The Bureau of Labor Statistics lists average hourly wages for 8 categories of law-related occupations (OCC Code 28000) (the units are dollars per hour).¹⁶

30 17 36 14 17 12 15 17

First find the mean:

$$\begin{aligned}\bar{x} &= \frac{30 + 17 + 36 + 14 + 17 + 12 + 15 + 17}{8} \\ &= \frac{158}{8} = 19.75 \text{ \$ per hour}\end{aligned}$$

Figure 1.12 displays the data as points above the number line, with their mean marked by an asterisk (*). The arrows mark two of the deviations from the mean.

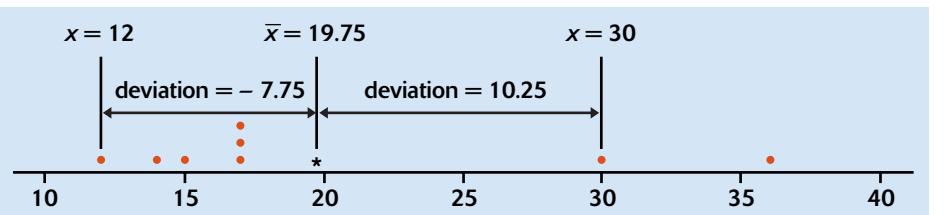


FIGURE 1.12 Average hourly wages for eight categories of legal professionals, with their mean (*) and the deviations of two observations from the mean, for Example 1.13.

The deviations show how spread out the data are about their mean. They are the starting point for calculating the variance and the standard deviation.

Observations x_i	Deviations $x_i - \bar{x}$	Squared deviations $(x_i - \bar{x})^2$
30	$30 - 19.75 = 10.25$	$10.25^2 = 105.0625$
17	$17 - 19.75 = -2.75$	$(-2.75)^2 = 7.5625$
36	$36 - 19.75 = 16.25$	$16.25^2 = 264.0625$
14	$14 - 19.75 = -5.75$	$(-5.75)^2 = 33.0625$
17	$17 - 19.75 = -2.75$	$(-2.75)^2 = 7.5625$
12	$12 - 19.75 = -7.75$	$(-7.75)^2 = 60.0625$
15	$15 - 19.75 = -4.75$	$(-4.75)^2 = 22.5625$
17	$17 - 19.75 = -2.75$	$(-2.75)^2 = 7.5625$
sum = 0		sum = 507.5000

The variance is the sum of the squared deviations divided by 1 less than the number of observations:

$$s^2 = \frac{507.5}{7} = 72.5$$

The standard deviation is the square root of the variance:

$$s = \sqrt{72.5} = 8.5147 \text{ \$ per hour}$$



Notice that the “average” in the variance s^2 divides the sum by 1 less than the number of observations, that is, $n - 1$ rather than n . The reason is that the deviations $x_i - \bar{x}$ always sum to exactly 0, so that knowing $n - 1$ of them determines the last one. Only $n - 1$ of the squared deviations can vary freely, and we average by dividing the total by $n - 1$. The number $n - 1$ is called the **degrees of freedom** of the variance or standard deviation. Many calculators offer a choice between dividing by n and dividing by $n - 1$, so be sure to use $n - 1$.

More important than the details of hand calculation are the properties that determine the usefulness of the standard deviation:

- s measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- $s = 0$ only when there is *no spread*. This happens only when all observations have the same value. Otherwise $s > 0$. As the observations become more spread out about their mean, s gets larger.
- s has the same units of measurement as the original observations. For example, if you measure wages in dollars per hour, s is also in dollars per hour. This is one reason to prefer s to the variance s^2 , which is in “dollars-per-hour squared.”
- Like the mean \bar{x} , s is not resistant. Strong skewness or a few outliers can greatly increase s . For example, the standard deviation of the earnings of

degrees of freedom

the 15 black female workers in Table 1.8 is \$1909.61. (Use software or a calculator to verify this.) If we omit the low outlier \$12,641, the standard deviation drops to \$1339.29.

You may rightly feel that the importance of the standard deviation is not yet clear. We will see in the next section that the standard deviation is the natural measure of spread for an important class of symmetric distributions, the Normal distributions. The usefulness of many statistical procedures is tied to distributions with particular shapes. This is certainly true of the standard deviation.

Choosing measures of center and spread

How do we choose between the five-number summary and \bar{x} and s to describe the center and spread of a distribution? Because the two sides of a strongly skewed distribution have different spreads, no single number such as s describes the spread well. The five-number summary, with its two quartiles and two extremes, does a better job.

CHOOSING A SUMMARY

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with extreme outliers. Use \bar{x} and s only for reasonably symmetric distributions that are free of outliers.

EXAMPLE 1.14

Risk and return

A central principle in the study of investments is that taking bigger risks is rewarded by higher returns, at least on the average over long periods of time. It is usual in finance to measure risk by the standard deviation of returns, on the grounds that investments whose returns show a large spread from year to year are less predictable and therefore more risky than those whose returns have a small spread. Compare, for example, the approximate mean and standard deviation of the annual percent returns on American common stocks and U.S. Treasury bills over the period from 1950 to 2000:

Investment	Mean return	Standard deviation
Common stocks	13.3%	17.1%
Treasury bills	5.2%	2.9%

Stocks are risky. They went up more than 13% per year on the average during this period, but they dropped almost 28% in the worst year. The large standard deviation reflects the fact that stocks have produced both large gains and large losses. When you buy a Treasury bill, on the other hand, you are lending money to the government for one year. You know that the government will pay you back with interest. That is much less risky than buying stocks, so (on the average) you get a smaller return.

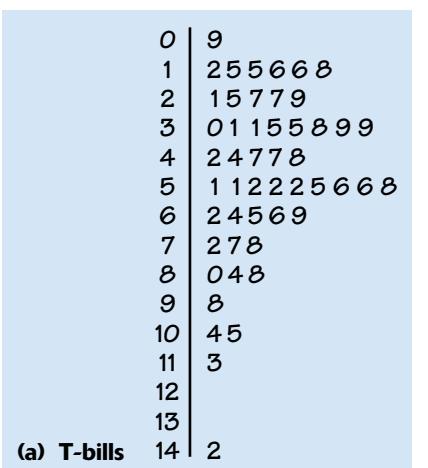


FIGURE 1.13(a) Stemplots of the annual returns on Treasury bills from 1950 to 2000, for Example 1.14. The stems are percents.

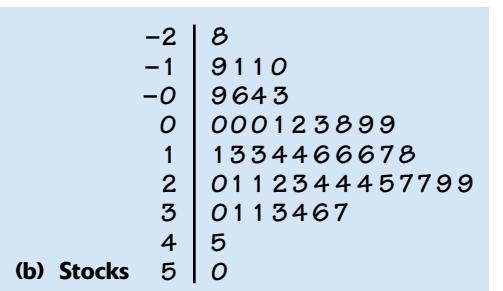


FIGURE 1.13(b) Stemplots of the annual returns on common stocks from 1950 to 2000, for Example 1.14. The stems are tens of percents.

Are \bar{x} and s good summaries for distributions of investment returns? Figure 1.13 displays stemplots of the annual returns for both investments. Because stock returns are so much more spread out than Treasury bill returns, a back-to-back stemplot does not work well. The stems in the stock stemplot are tens of percents; the stems for bills are percents. The lowest returns are -28% for stocks and 0.9% for bills. You see that returns on Treasury bills have a right-skewed distribution. Convention in the financial world calls for \bar{x} and s because some parts of investment theory use them. For describing this right-skewed distribution, however, the five-number summary would be more informative.

Remember that a graph gives the best overall picture of a distribution. Numerical measures of center and spread report specific facts about a distribution, but they do not describe its entire shape. Numerical summaries do not disclose the presence of multiple peaks or gaps, for example. **Always plot your data.**

APPLY YOUR KNOWLEDGE

- 1.37 **Privately held restaurant companies.** The Forbes 500 list of the largest privately held companies includes six restaurant-industry companies. Here they are, with annual revenues in millions of dollars (2000):

Company name	Revenue (\$millions)
Metromedia	1600
Domino's Pizza	1157
Buffets	937
Ilitch Ventures	800
AFC Enterprises	707
RTM Restaurant Group	700

(Only Domino's Pizza has a name familiar to consumers. You can check www.forbes.com/private500/ to learn the brand names under which the other companies operate restaurants.) A graph of only 6 observations gives little information, so we proceed to compute the mean and standard deviation.

- (a) Find the mean from its definition. That is, find the sum of the 6 observations and divide by 6.
 - (b) Find the standard deviation from its definition. That is, find the deviations of each observation from the mean, square the deviations, then obtain the variance and the standard deviation. Example 1.13 shows the method.
 - (c) Now enter the data into your calculator and use the mean and standard deviation buttons to obtain \bar{x} and s . Do the results agree with your hand calculations?
- 1.38 Where are the doctors? Table 1.7 (page 29) gives the number of medical doctors per 100,000 people in each state. A graph of the distribution (Exercise 1.28) shows that the District of Columbia is a high outlier. Because D.C. is a city rather than a state, we will omit it here.
- (a) Make a graph of the data if you did not do so earlier.
 - (b) Calculate both the five-number summary and \bar{x} and s for the number of doctors per 100,000 people in the 50 states. Based on your graph, which description do you prefer?
 - (c) What facts about the distribution can you see in the graph that the numerical summaries don't reveal? Remember that measures of center and spread are not complete descriptions of a distribution.

SECTION 1.2 SUMMARY

- A numerical summary of a distribution should report its **center** and its **spread or variability**.
- The **mean** \bar{x} and the **median** M describe the center of a distribution in different ways. The mean is the arithmetic average of the observations, and the median is the midpoint of the values.
- When you use the median to indicate the center of the distribution, describe its spread by giving the **quartiles**. The **first quartile** Q_1 has one-fourth of the observations below it, and the **third quartile** Q_3 has three-fourths of the observations below it.
- The **five-number summary** consisting of the median, the quartiles, and the high and low extremes provides a quick overall description of a

distribution. The median describes the center, and the quartiles and extremes show the spread.

- **Boxplots** based on the five-number summary are useful for comparing several distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the extremes and show the full spread of the data.
- The variance s^2 and especially its square root, the standard deviation s , are common measures of spread about the mean as center. The standard deviation s is zero when there is no spread and gets larger as the spread increases.
- A **resistant measure** of any aspect of a distribution is relatively unaffected by changes in the numerical value of a small proportion of the total number of observations, no matter how large these changes are. The median and quartiles are resistant, but the mean and the standard deviation are not.
- The mean and standard deviation are good descriptions for symmetric distributions without outliers. They are most useful for the Normal distributions, introduced in the next section. The five-number summary is a better exploratory summary for skewed distributions.

SECTION 1.2 EXERCISES

- 1.39 **Supermarket shoppers.** In Exercise 1.30 you found the mean amount spent by 50 consecutive shoppers at a supermarket. Now find the median of these amounts. Is the median smaller or larger than the mean? Explain why this is so.
- 1.40 **Unemployment in the states.** Table 1.1 (page 10) records the unemployment rate for each of the states. Figure 1.2 (page 12) is a histogram of these data. Do you prefer the five-number summary or \bar{x} and s as a brief numerical description? Why? Calculate your preferred description.
- 1.41 **The Platinum Gasaver.** National Fuelsaver Corporation manufactures the Platinum Gasaver, a device they claim “may increase gas mileage by 22%.” In an advertisement published in the *Des Moines Register*, the gas mileages with and without the device were presented for 15 “identical” 5-liter vehicles. The percent changes in gas mileage for the vehicles were calculated and are presented here:

48.3 46.9 46.8 44.6 40.2 38.5 34.6 33.7
28.7 28.7 24.8 10.8 10.4 6.9 -12.4

The 12.4% *decrease* in gas mileage is an outlier in this data set.

- (a) Find the mean \bar{x} and the standard deviation s .
- (b) Find \bar{x} and s for the 14 observations that remain when you ignore the outlier. How does the outlier affect the values of \bar{x} and s ?
- (c) What do you think the advertisement means when it calls these vehicles “identical”?

The following three exercises use the Mean and Median applet available at www.whfreeman.com/pbs to explore the behavior of the mean and median.



- 1.42 **Mean = median?** Place two observations on the line by clicking below it. Why does only one arrow appear?



- 1.43 **Extreme observations.** Place three observations on the line by clicking below it, two close together near the center of the line, and one somewhat to the right of these two.

- Pull the single rightmost observation out to the right. (Place the cursor on the point, hold down a mouse button, and drag the point.) How does the mean behave? How does the median behave? Explain briefly why each measure acts as it does.
- Now drag the single point to the left as far as you can. What happens to the mean? What happens to the median as you drag this point past the other two (watch carefully)?



- 1.44 **Don't change the median.** Place 5 observations on the line by clicking below it.

- Add one additional observation *without changing the median*. Where is your new point?
- Use the applet to convince yourself that when you add yet another observation (there are now 7 in all), the median does not change no matter where you put the 7th point. Explain why this must be true.

- 1.45 **Education and income.** Each March, the Bureau of Labor Statistics records the incomes of all adults in a sample of 50,000 American households. We are interested in how income varies with the highest education level a person has reached. Computer software applied to the data from the March 2000 survey gives the following results for people aged 25 or over:

	A	B	C	D	E	F	G	H
1	Education	N	5%	25%	Median	75%	95%	
2	High school diploma	31970	0	7800	17000	29600	56294	
3	Some college, no degree	18797	0	8083	19600	34800	70026	
4	Bachelor's degree	14705	500	17501	34150	55307	110086	
5	Master's degree	4918	3300	27043	45069	68500	132560	
6	Professional degree	1229	3000	33922	65850	118992	236967	
7								

It is common to make boxplots of large data sets using the 5% and 95% points in place of the minimum and maximum. The highest income among the 31,970 people with only a high school education, for example, is \$425,510. It is more informative to see that 95% of this group earned less than \$56,294. The 5% and 95% points contain between them the middle 90% of the observations.

- Use this output to make boxplots that compare the income distributions for the five education groups.
- Write a brief summary of the relationship between education and income. For example, do people who start college but don't get a degree do much better than people with only a high school education?

- 1.46 **Education and income.** The output in the previous exercise shows that the data set contains information about 31,970 people with only a high school

diploma and 1229 people with a professional degree. What are the positions of the median, the two quartiles, and the 5th and 95th percentiles in the ordered list of incomes for each of these groups?

- 1.47 **Crime in the states.** “Quality-of-life” issues increasingly influence companies that wish to locate their operations to attract skilled employees. Table 1.7 (page 29) gives data for the states on a quality-of-life issue, the rate of violent crime. Give a graphical and numerical description of the distribution of crime rates, with a brief summary of your findings.
- 1.48 **Crime in the states, continued.** The Census Bureau groups the states into nine geographic regions. By combining some adjacent regions, we can group states as follows:

Northeast	Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont
South	Alabama, Delaware, Florida, Georgia, Kentucky, Maryland, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, West Virginia
Midwest	Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin

Use a graph and numerical descriptions to compare the distributions of violent crime rates in these three parts of the United States. Briefly summarize your findings.

- 1.49 **\bar{x} and s are not enough.** The mean \bar{x} and standard deviation s measure center and spread but are not a complete description of a distribution. Data sets with different shapes can have the same mean and standard deviation. To demonstrate this fact, find \bar{x} and s for these two small data sets. Then make a stemplot of each and comment on the shape of each distribution.

Data A	9.14	8.14	8.74	8.77	9.26	8.10	6.13
	3.10	9.13	7.26	4.74			
Data B	6.58	5.76	7.71	8.84	8.47	7.04	5.25
	5.56	7.91	6.89	12.50			

- 1.50 **Wealth of *Forbes* readers.** The business magazine *Forbes* estimates that the “average” household wealth of its readers is either about \$800,000 or about \$2.2 million, depending on which “average” it reports. Which of these numbers is the mean wealth and which is the median wealth? Explain your answer.
- 1.51 **Returns on Treasury bills.** Figure 1.13(a) on page 44 is a stemplot of the annual returns on U.S. Treasury bills for the years 1950 to 2000. (The entries are rounded to the nearest tenth of a percent.)
- Use the stemplot to find the five-number summary of T-bill returns.
 - The mean of these returns is about 5.19%. Explain from the shape of the distribution why the mean return is larger than the median return.
- 1.52 **Salary increase for the owner.** Last year a small accounting firm paid each of its five clerks \$25,000, two junior accountants \$60,000 each, and the firm’s owner \$255,000.

- (a) What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary?
- (b) This year the firm gives no raises to the clerks and junior accountants, while the owner's take increases to \$455,000. How does this change affect the mean? How does it affect the median?
- 1.53 A hot stock?** We saw in Example 1.14 that it is usual in the study of investments to use the mean and standard deviation to summarize and compare investment returns. Table 1.10 gives the monthly returns on Philip Morris stock for the period from June 1990 to July 2001. (The return on an investment consists of the change in its price plus any cash payments made, given here as a percent of its price at the start of each month.)
- (a) Make either a histogram or a stemplot of these data. How did you decide which graph to make?
- (b) There are two clear outliers. What are the values of these observations? (The most extreme observation is explained by news of action against smoking, which depressed this tobacco company stock.) Describe the shape, center, and spread of the data after you omit the two outliers.*
- (c) Find the mean monthly return and the standard deviation of the returns (include the outliers). If you invested \$100 in this stock at the beginning of a month and got the mean return, how much would you have at the end of the month?
- (d) The distribution can be described as “symmetric and single-peaked, with two low outliers.” If you invested \$100 in this stock at the beginning of the worst month in the data (the more extreme outlier), how much would you have at the end of the month? Find the mean and standard deviation

TABLE 1.10 Monthly percent returns on Philip Morris stock from June 1990 to July 2001

3.0	-5.7	1.2	4.1	3.2	7.3	7.5	18.7	3.7	-1.8
2.4	-6.5	6.7	9.4	-2.0	-2.8	-3.4	19.2	-4.8	0.5
-0.6	2.8	-0.5	-4.5	8.7	2.7	4.1	-10.3	4.8	-2.3
-3.1	-10.2	-3.7	-26.6	7.2	-2.4	-2.8	3.4	-4.6	17.2
4.2	0.5	8.3	-7.1	-8.4	7.7	-9.6	6.0	6.8	10.9
1.6	0.2	-2.4	-2.4	3.9	1.7	9.0	3.6	7.6	3.2
-3.7	4.2	13.2	0.9	4.2	4.0	2.8	6.7	-10.4	2.7
10.3	5.7	0.6	-14.2	1.3	2.9	11.8	10.6	5.2	13.8
-14.7	3.5	11.7	1.5	2.0	-3.2	-3.9	-4.7	9.8	4.9
-8.3	4.8	-3.2	-10.9	0.7	6.4	11.3	-5.1	12.3	10.5
9.4	-3.6	-12.4	-16.5	-8.9	-0.4	10.0	5.4	-7.3	0.5
-7.4	-22.9	-0.5	-10.6	-9.2	-3.3	5.2	5.4	19.4	3.5
-4.9	17.8	0.7	24.4	4.3	16.6	0.0	9.5	-0.4	5.6
2.6	-2.7	-8.1	4.2						

*Both outliers are negative rates of return; however, there is one positive rate of return that is almost as separated from the neighboring rates of return in the positive direction as the two outliers on the negative side. For the sake of this exercise, we will not consider this highest positive rate of return to be an outlier.

again, this time leaving out the two low outliers. How much did these two observations affect the summary measures? Would leaving out these two observations substantially change the median? The quartiles? How do you know, without actual calculation? (Returns over longer periods of time, or returns on portfolios containing several investments, tend to follow a Normal distribution more closely than these monthly returns. So use of the mean and standard deviation is better justified for such data.)

- 1.54 **Initial public offerings.** During the stock market boom of the 1990s, initial public offerings (IPOs) of the stock of new companies often produced enormous gains for people who bought the stocks when they first became available. At least that's what legend says. A study of all 4567 companies that went public in the years 1990 to 2000 (excluding very small IPOs) found that on the average their stock prices had either *risen* 111% or *declined* 31% by the end of the year 2000.¹⁷ One of these numbers is the mean change in price and one is the median change. Which is which, and how can you tell?

- 1.55 **Highly paid athletes.** A news article reports that of the 411 players on National Basketball Association rosters in February 1998, only 139 “made more than the league average salary” of \$2.36 million. Is \$2.36 million the mean or median salary for NBA players? How do you know?

- 1.56 **Mean or median?** Which measure of center, the mean or the median, should you use in each of the following situations?
- Middletown is considering imposing an income tax on citizens. The city government wants to know the average income of citizens so that it can estimate the total tax base.
 - In a study of the standard of living of typical families in Middletown, a sociologist estimates the average family income in that city.

- 1.57 **A standard deviation contest.** You must choose four numbers from the whole numbers 0 to 10, with repeats allowed.
- Choose four numbers that have the smallest possible standard deviation.
 - Choose four numbers that have the largest possible standard deviation.
 - Is more than one choice possible in either (a) or (b)? Explain.

- 1.58 **Discovering outliers.** Whether an observation is an outlier is a matter of judgment. When large numbers of data are scanned automatically, it is convenient to have a rule for identifying suspected outliers. The $1.5 \times IQR$ rule is in common use:

1.5 × IQR rule

interquartile range

- The **interquartile range** IQR is the distance between the first and third quartiles, $IQR = Q_3 - Q_1$. This is the spread of the middle half of the data.
 - An observation is a suspected outlier if it lies more than $1.5 \times IQR$ below the first quartile Q_1 or above the third quartile Q_3 .
- Confirm that the low outlier in the earnings of black female bank workers (Table 1.8, page 31) lies more than $1.5 \times IQR$ below Q_1 . That is, the $1.5 \times IQR$ rule fingers this observation as a suspected outlier.
 - Does the rule point out any suspected outliers among the black male workers in Table 1.8?

1.3 The Normal Distributions

We now have a kit of graphical and numerical tools for describing distributions. What is more, we have a clear strategy for exploring data on a single quantitative variable:

1. Always plot your data: make a graph, usually a histogram or a stemplot.
2. Look for the overall pattern (shape, center, spread) and for striking deviations such as outliers.
3. Calculate a numerical summary to briefly describe center and spread.

Here is one more step to add to this strategy:

4. Sometimes the overall pattern of a large number of observations is so regular that we can describe it by a smooth curve.

Density curves

Figure 1.14 is a histogram of the city gas mileage achieved by all 856 2001 model year motor vehicles listed in the government's annual fuel economy report.¹⁸ The distribution is quite regular. With the exception of a few high outliers, the histogram is roughly symmetric, and both tails fall off quite smoothly from a single center peak located at approximately 20 miles per gallon. The smooth curve overlying the histogram bars in Figure 1.14 is an approximate description of the overall pattern of the data. The curve

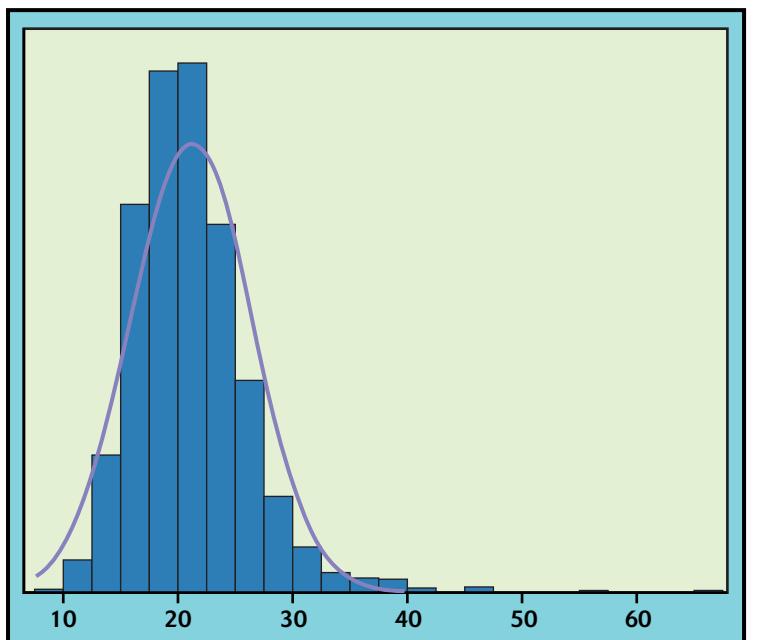


FIGURE 1.14 Histogram of the city gas mileage (miles per gallon) of 856 2001 model year motor vehicles. The smooth curve shows the overall shape of the distribution.

mathematical model

is a **mathematical model** for the distribution. A mathematical model is an idealized description. It gives a compact picture of the overall pattern of the data but ignores minor irregularities as well as any outliers.

We will see that it is easier to work with the smooth curve in Figure 1.14 than with the histogram. The reason is that the histogram depends on our choice of classes, while with a little care we can use a curve that does not depend on any choices we make. Here's how we do it.

EXAMPLE 1.15**From histogram to density curve**

Our eyes respond to the *areas* of the bars in a histogram. The bar areas represent proportions of the observations. Figure 1.15(a) is a copy of Figure 1.14 with the leftmost bars shaded. The area of the shaded bars in Figure 1.15(a) represents the proportion of 2001 model year vehicles with miles per gallon ratings of less than 20. There are 384 such vehicles, which represent $384/856 = 0.449 = 44.9\%$ of all 2001 model year vehicles.

Now concentrate on the curve drawn through the bars. In Figure 1.15(b), the area under the curve to the left of 20 is shaded. Adjust the scale of the graph so that *the total area under the curve is exactly 1*. This area represents the proportion 1, that is, all the observations. Areas under the curve then represent proportions of the observations. The curve is now a *density curve*. The shaded area under the density curve in Figure 1.15(b) represents the proportion of 2001 model year vehicles with miles per gallon ratings of less than 20. This area is 0.410, only 0.039 away from the histogram result. You can see that areas under the density curve give quite good approximations of areas given by the histogram.

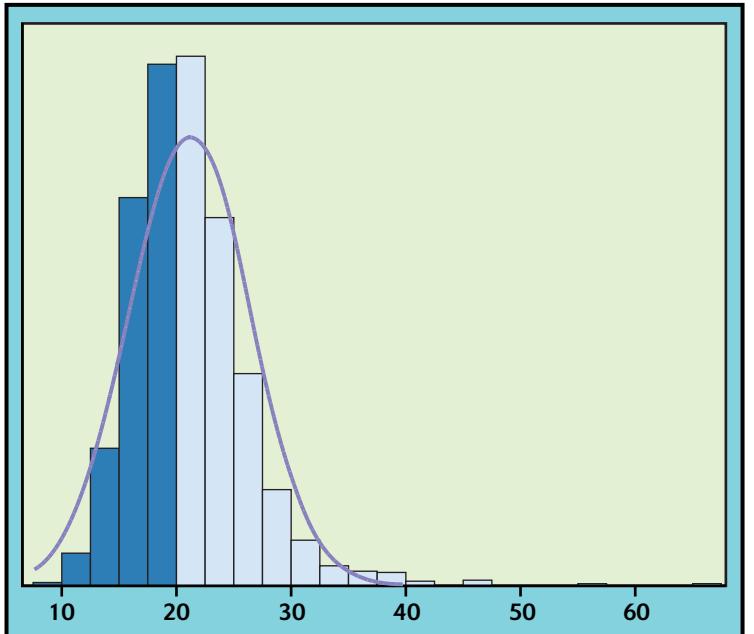


FIGURE 1.15(a) The proportion of gas mileages less than 20.0 from the histogram is 0.449.

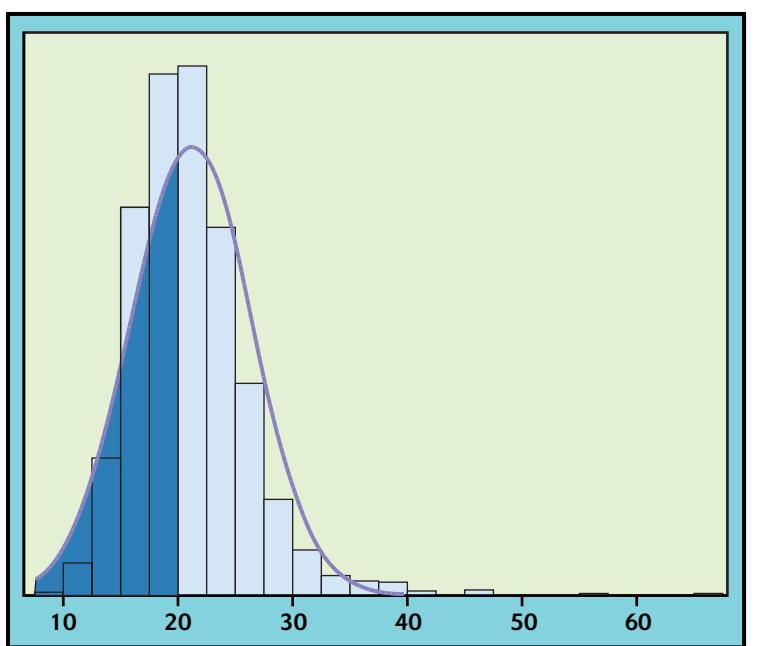


FIGURE 1.15(b) The proportion of gas mileages less than 20.0 from the density curve is 0.410.

DENSITY CURVE

- A density curve is a curve that
- is always on or above the horizontal axis, and
 - has area exactly 1 underneath it.

A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values is the proportion of all observations that fall in that range.

Normal curve

The density curve in Figures 1.14 and 1.15 is a **Normal curve**. Density curves, like distributions, come in many shapes. Figure 1.16 shows two density curves: a symmetric Normal density curve and a right-skewed curve. A density curve of the appropriate shape is often an adequate description of the overall pattern of a distribution. Outliers, which are deviations from the overall pattern, are not described by the curve. Of course, no set of real data is exactly described by a density curve. The curve is an approximation that is easy to use and accurate enough for practical use.

The median and mean of a density curve

Our measures of center and spread apply to density curves as well as to actual sets of observations. The median and quartiles are easy. Areas under a density curve represent proportions of the total number of observations. The median is the point with half the observations on either side.

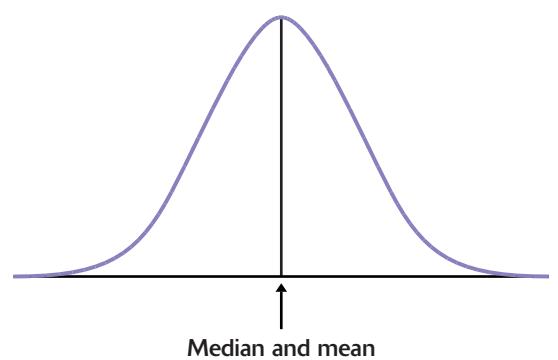


FIGURE 1.16(a) The median and mean of a symmetric density curve.

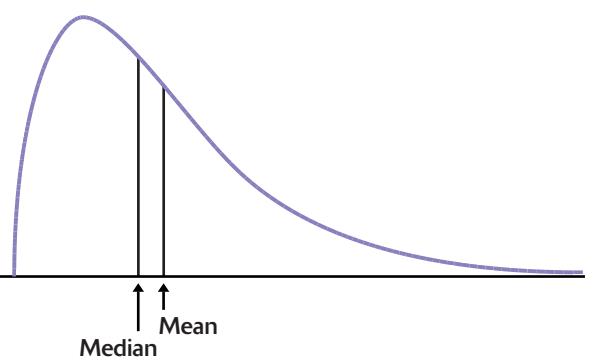


FIGURE 1.16(b) The median and mean of a right-skewed density curve.

So the **median of a density curve** is the equal-areas point, the point with half the area under the curve to its left and the remaining half of the area to its right. The quartiles divide the area under the curve into quarters. One-fourth of the area under the curve is to the left of the first quartile, and three-fourths of the area is to the left of the third quartile. You can roughly locate the median and quartiles of any density curve by eye by dividing the area under the curve into four equal parts.

Because density curves are idealized patterns, a symmetric density curve is exactly symmetric. The median of a symmetric density curve is therefore at its center. Figure 1.16(a) shows the median of a symmetric curve. It isn't so easy to spot the equal-areas point on a skewed curve. There are mathematical ways of finding the median for any density curve. We did that to mark the median on the skewed curve in Figure 1.16(b).

What about the mean? The mean of a set of observations is their arithmetic average. If we think of the observations as weights strung out along a thin rod, the mean is the point at which the rod would balance. This fact is also true of density curves. The **mean is the point at which the curve would balance if made of solid material**. Figure 1.17 illustrates this fact about the mean. A symmetric curve balances at its center because the two sides are identical. The **mean and median of a symmetric density curve are equal**, as in Figure 1.16(a). We know that the mean of a skewed distribution is pulled toward the long tail. Figure 1.16(b) shows how the mean of a skewed density curve is pulled toward the long tail more than is the median. It's hard to locate the balance point by eye on a skewed curve. There are mathematical ways of calculating the mean for any density curve, so we are able to mark the mean as well as the median in Figure 1.16(b).

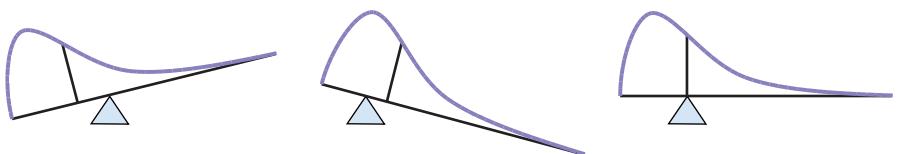


FIGURE 1.17 The mean is the balance point of a density curve.

MEDIAN AND MEAN OF A DENSITY CURVE

The **median** of a density curve is the equal-areas point, the point that divides the area under the curve in half.

The **mean** of a density curve is the balance point, at which the curve would balance if made of solid material.

The median and mean are the same for a symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.

We can roughly locate the mean, median, and quartiles of any density curve by eye. This is not true of the standard deviation. When necessary, we can once again call on more advanced mathematics to learn the value of the standard deviation. The study of mathematical methods for doing calculations with density curves is part of theoretical statistics. Though we are concentrating on statistical practice, we often make use of the results of mathematical study.

Because a density curve is an idealized description of the distribution of data, we need to distinguish between the mean and standard deviation of the density curve and the mean \bar{x} and standard deviation s computed from the actual observations. The usual notation for the mean of an idealized distribution is μ (the Greek letter mu). We write the standard deviation of a density curve as σ (the Greek letter sigma).

*mean μ
standard deviation σ*

APPLY YOUR KNOWLEDGE

- 1.59 (a) Sketch a density curve that is symmetric but has a shape different from that of the curve in Figure 1.16(a).
(b) Sketch a density curve that is strongly skewed to the left.
- 1.60 Figure 1.18 displays the density curve of a *Uniform distribution*. The curve takes the constant value 1 over the interval from 0 to 1 and is 0 outside that range of values. This means that data described by this distribution take values that are uniformly spread between 0 and 1. Use areas under this density curve to answer the following questions.
 - (a) Why is the total area under this curve equal to 1?
 - (b) What percent of the observations lie above 0.8?
 - (c) What percent of the observations lie below 0.6?
 - (d) What percent of the observations lie between 0.25 and 0.75?
 - (e) What is the mean μ of this distribution?

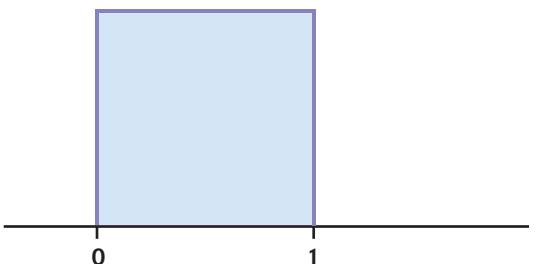


FIGURE 1.18 The density curve of a Uniform distribution, for Exercise 1.60.

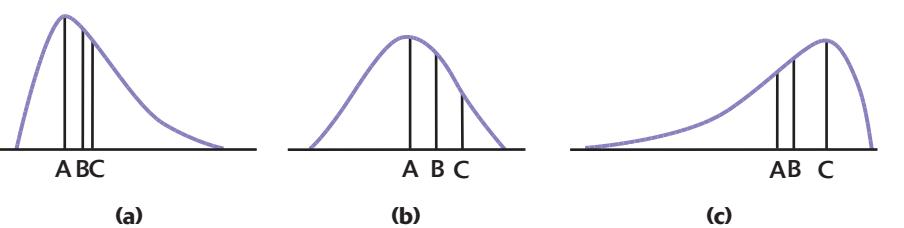


FIGURE 1.19 Three density curves, for Exercise 1.61.

- 1.61 Figure 1.19 displays three density curves, each with three points marked on them. At which of these points on each curve do the mean and the median fall?

Normal distributions

One particularly important class of density curves has already appeared in Figures 1.14 and 1.16(a). These density curves are symmetric, single-peaked, and bell-shaped. They are called *Normal curves*, and they describe **Normal distributions**. All Normal distributions have the same overall shape. The exact density curve for a particular Normal distribution is described by giving its mean μ and its standard deviation σ . The mean is located at the center of the symmetric curve and is the same as the median. Changing μ without changing σ moves the Normal curve along the horizontal axis without changing its spread. The standard deviation σ controls the spread of a Normal curve. Figure 1.20 shows two Normal curves with different values of σ . The curve with the larger standard deviation is more spread out.

The standard deviation σ is the natural measure of spread for Normal distributions. Not only do μ and σ completely determine the shape of a Normal curve, but we can locate σ by eye on the curve. Here's how. Imagine that you are skiing down a mountain that has the shape of a Normal curve. At first, you descend at an ever-steepener angle as you go out from the peak:



Fortunately, before you find yourself going straight down, the slope begins to grow flatter rather than steeper as you go out and down:



The points at which this change of curvature takes place are located at distance σ on either side of the mean μ . You can feel the change as you run a pencil along a Normal curve, and so find the standard deviation. Remember that μ and σ alone do not specify the shape of most distributions, and that the shape of density curves in general does not reveal σ . These are special properties of Normal distributions.

Why are the Normal distributions important in statistics? Here are three reasons. First, Normal distributions are good descriptions for some distribu-

Normal distributions

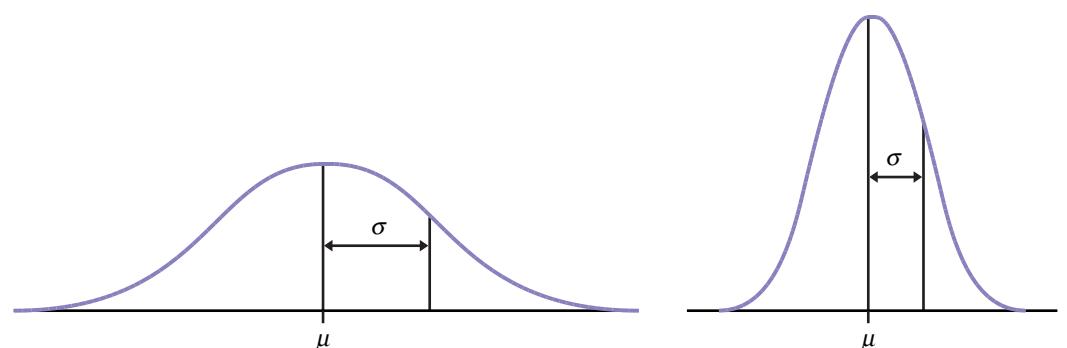


FIGURE 1.20 Two Normal curves, showing the mean μ (left) and standard deviation σ (right).

tions of *real data*. Distributions that are often close to Normal include scores on tests taken by many people (such as GMAT exams), repeated careful measurements of the same quantity (such as measurements taken from a production process), and characteristics of biological populations (such as yields of corn). Second, Normal distributions are good approximations to the results of many kinds of *chance outcomes*, such as tossing a coin many times. Third, and most important, we will see that many *statistical inference* procedures based on Normal distributions work well for other roughly symmetric distributions. However, even though many sets of data follow a Normal distribution, many do not. Most company salary distributions, for example, are skewed to the right and so are not Normal. Non-Normal data, like nonnormal people, not only are common but are sometimes more interesting than their normal counterparts.

The 68–95–99.7 rule

Although there are many Normal curves, they all have common properties. In particular, all Normal distributions obey the following rule.

THE 68–95–99.7 RULE

In the Normal distribution with mean μ and standard deviation σ :

- 68% of the observations fall within σ of the mean μ .
- 95% of the observations fall within 2σ of μ .
- 99.7% of the observations fall within 3σ of μ .

Figure 1.21 illustrates the 68–95–99.7 rule. By remembering these three numbers, you can think about Normal distributions without constantly making detailed calculations.

EXAMPLE 1.16

Using the 68–95–99.7 rule

The distribution of weights of 9-ounce bags of a particular brand of potato chips is approximately Normal with mean $\mu = 9.12$ ounces and standard deviation

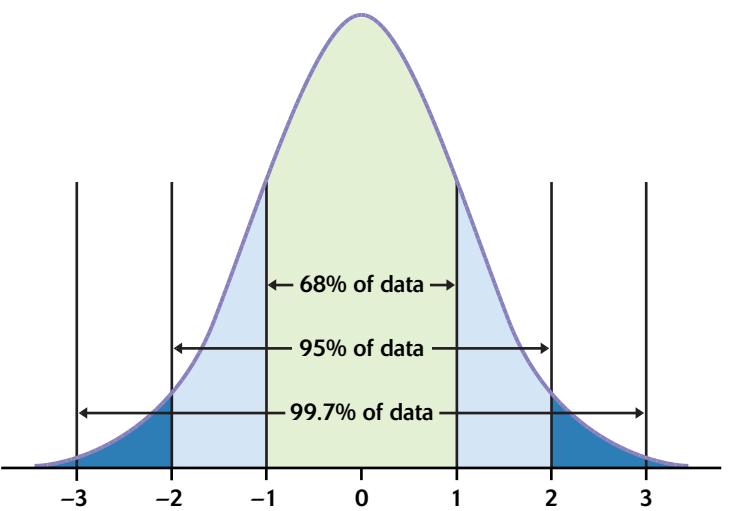


FIGURE 1.21 The 68–95–99.7 rule for Normal distributions.

$\sigma = 0.15$ ounce. Figure 1.22 shows what the 68–95–99.7 rule says about this distribution.

Two standard deviations is 0.3 ounces for this distribution. The 95 part of the 68–95–99.7 rule says that the middle 95% of 9-ounce bags weigh between $9.12 - 0.3$ and $9.12 + 0.3$ ounces, that is, between 8.82 ounces and 9.42 ounces. This fact is exactly true for an exactly Normal distribution. It is approximately true for the weights of 9-ounce bags of chips because the distribution of these weights is approximately Normal.

The other 5% of bags have weights outside the range from 8.82 to 9.42 ounces. Because the Normal distributions are symmetric, half of these bags are on the heavy side. So the heaviest 2.5% of 9-ounce bags are heavier than 9.42 ounces.

The 99.7 part of the 68–95–99.7 rule says that almost all bags (99.7% of them) have weights between $\mu - 3\sigma$ and $\mu + 3\sigma$. This range of weights is 8.67 to 9.57 ounces.

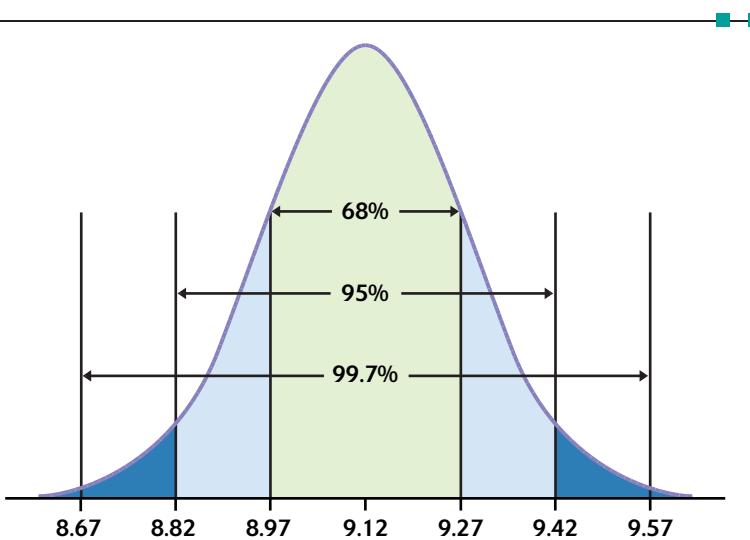


FIGURE 1.22 The 68–95–99.7 rule applied to the distribution of the weights of bags of potato chips, for Example 1.16. Here $\mu = 9.12$ ounces and $\sigma = 0.15$ ounce.

Because we will mention Normal distributions often, a short notation is helpful. We abbreviate the Normal distribution with mean μ and standard deviation σ as $N(\mu, \sigma)$. For example, the distribution of weights in the previous example is $N(9.12, 0.15)$.

APPLY YOUR KNOWLEDGE

- 1.62 **Heights of young men.** Product designers often must consider physical characteristics of their target population. For example, the distribution of heights of men of ages 20 to 29 years is approximately Normal with mean 69 inches and standard deviation 2.5 inches. Draw a Normal curve on which this mean and standard deviation are correctly located. (Hint: Draw the curve first, locate the points where the curvature changes, then mark the horizontal axis.)
- 1.63 **More on young men's heights.** The distribution of heights of young men is approximately Normal with mean 69 inches and standard deviation 2.5 inches. Use the 68–95–99.7 rule to answer the following questions.
 - (a) What percent of these men are taller than 74 inches?
 - (b) Between what heights do the middle 95% of young men fall?
 - (c) What percent of young men are shorter than 66.5 inches?
- 1.64 **IQ test scores.** Employers may ask job applicants to take an IQ test if the test has been shown to predict performance on the job. One such test is the Wechsler Adult Intelligence Scale (WAIS). Scores on the WAIS for the 20-to-34 age group are approximately Normally distributed with $\mu = 110$ and $\sigma = 25$. Use the 68–95–99.7 rule to answer these questions.
 - (a) About what percent of people in this age group have scores above 110?
 - (b) About what percent have scores above 160?
 - (c) In what range do the middle 95% of all scores lie?

The standard Normal distribution

As the 68–95–99.7 rule suggests, all Normal distributions share many common properties. In fact, all Normal distributions are the same if we measure in units of size σ about the mean μ as center. Changing to these units is called *standardizing*. To standardize a value, subtract the mean of the distribution and then divide by the standard deviation.

STANDARDIZING AND z -SCORES

If x is an observation from a distribution that has mean μ and standard deviation σ , the **standardized value** of x is

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

A standardized value is often called a **z -score**.

A z -score tells us how many standard deviations the original observation falls away from the mean, and in which direction. Observations larger than the mean are positive when standardized, and observations smaller than the mean are negative when standardized.

EXAMPLE 1.17**Standardizing potato chip bag weights**

The weights of 9-ounce potato chip bags are approximately Normal with $\mu = 9.12$ ounces and $\sigma = 0.15$ ounces. The standardized weight is

$$z = \frac{\text{weight} - 9.12}{0.15}$$

A bag's standardized weight is the number of standard deviations by which its weight differs from the mean weight of all bags. A bag weighing 9.3 ounces, for example, has *standardized weight*

$$z = \frac{9.3 - 9.12}{0.15} = 1.2$$

or 1.2 standard deviations above the mean. Similarly, a bag weighing 8.7 ounces has standardized weight

$$z = \frac{8.7 - 9.12}{0.15} = -2.8$$

or 2.8 standard deviations below the mean bag weight.



If the variable we standardize has a Normal distribution, standardizing does more than give a common scale. It makes all Normal distributions into a single distribution, and this distribution is still Normal. Standardizing a variable that has any Normal distribution produces a new variable that has the *standard Normal distribution*.

STANDARD NORMAL DISTRIBUTION

The **standard Normal distribution** is the Normal distribution $N(0, 1)$ with mean 0 and standard deviation 1.

If a variable x has any Normal distribution $N(\mu, \sigma)$ with mean μ and standard deviation σ , then the standardized variable

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

has the standard Normal distribution.

APPLY YOUR KNOWLEDGE

- 1.65 SAT versus ACT.** Eleanor scores 680 on the mathematics part of the SAT. The distribution of SAT scores in a reference population is Normal, with mean 500 and standard deviation 100. Gerald takes the American College Testing (ACT) mathematics test and scores 27. ACT scores are Normally distributed with mean 18 and standard deviation 6. Find the standardized scores for both students. Assuming that both tests measure the same kind of ability, who has the higher score?

Normal distribution calculations

An area under a density curve is a proportion of the observations in a distribution. Any question about what proportion of observations lies in

some range of values can be answered by finding an area under the curve. Because all Normal distributions are the same when we standardize, we can find areas under any Normal curve from a single table, a table that gives areas under the curve for the standard Normal distribution.

EXAMPLE 1.18**Using the standard Normal distribution**

What proportion of all 9-ounce bags of potato chips weighs less than 9.3 ounces? This proportion is the area under the $N(9.12, 0.15)$ curve to the left of the point 9.3. Because the standardized weight corresponding to 9.3 ounces is

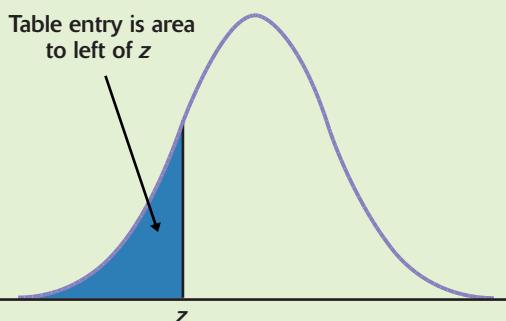
$$z = \frac{x - \mu}{\sigma} = \frac{9.3 - 9.12}{0.15} = 1.2$$

this area is the same as the area under the standard Normal curve to the left of the point $z = 1.2$. Figure 1.23(a) shows this area.

Many calculators will give you areas under the standard Normal curve. In case your calculator or software does not, Table A in the back of the book gives some of these areas.

THE STANDARD NORMAL TABLE

Table A is a table of areas under the standard Normal curve. The table entry for each value z is the area under the curve to the left of z .

**EXAMPLE 1.19****Using the standard Normal table**

Problem: Find the proportion of observations from the standard Normal distribution that are less than 1.2.

Solution: To find the area to the left of 1.20, locate 1.2 in the left-hand column of Table A, then locate the remaining digit 0 as .00 in the top row. The entry opposite 1.2 and under .00 is 0.8849. This is the area we seek. Figure 1.23(a) illustrates the relationship between the value $z = 1.20$ and the area 0.8849. Because $z = 1.20$ is the standardized value of weight 9.3 ounces, the proportion of 9-ounce bags that weigh less than 9.3 ounces is 0.8849 (about 88.5%).

Problem: Find the proportion of observations from the standard Normal distribution that are greater than -2.15.

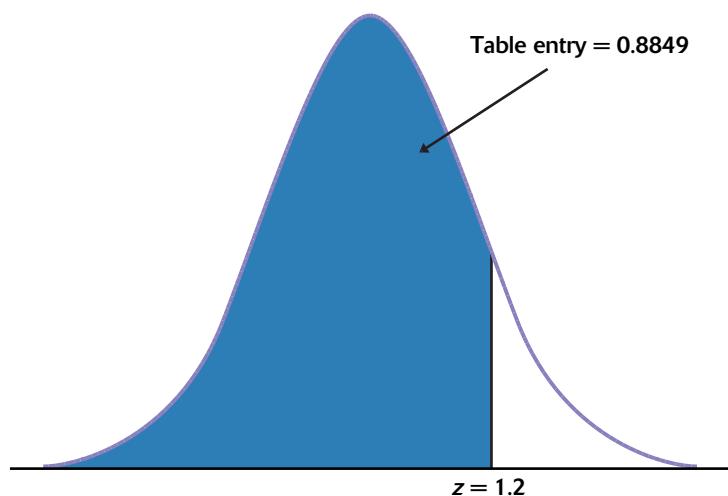


FIGURE 1.23(a) The area under a standard Normal curve to the left of the point $z = 1.2$ is 0.8849. Table A gives areas under the standard Normal curve.

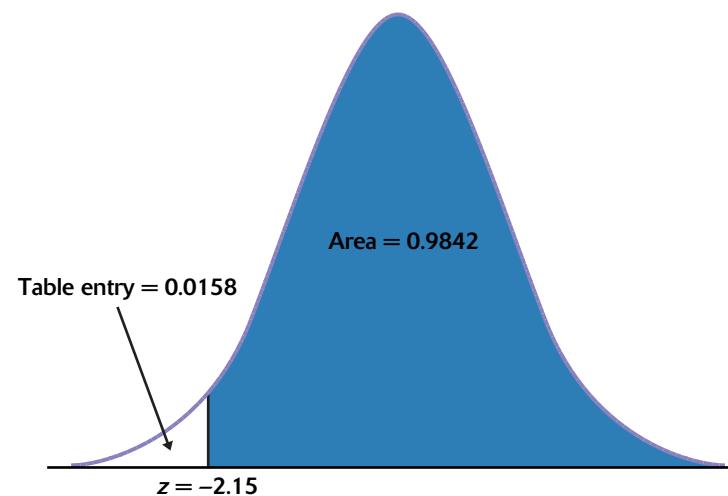


FIGURE 1.23(b) Areas under the standard Normal curve to the right and left of $z = -2.15$. Table A gives only areas to the left.

Solution: Enter Table A under $z = -2.15$. That is, find -2.1 in the left-hand column and $.05$ in the top row. The table entry is 0.0158 . This is the area to the *left* of -2.15 . Because the total area under the curve is 1 , the area lying to the *right* of -2.15 is $1 - 0.0158 = 0.9842$. Figure 1.23(b) illustrates these areas.

We can answer any question about proportions of observations in a Normal distribution by standardizing and then using the standard Normal table. Here is an outline of the method for finding the proportion of the distribution in any region.

FINDING NORMAL PROPORTIONS

1. State the problem in terms of the observed variable x .
2. Standardize x to restate the problem in terms of a standard Normal variable z . Draw a picture to show the area under the standard Normal curve.
3. Find the required area under the standard Normal curve using Table A and the fact that the total area under the curve is 1.

EXAMPLE 1.20**Normal distribution calculations**

The annual rate of return on stock indexes (which combine many individual stocks) is approximately Normal. Since 1945, the Standard & Poor's 500-stock index has had a mean yearly return of about 12%, with a standard deviation of 16.5%. Take this Normal distribution to be the distribution of yearly returns over a long period. The market is down for the year if the return on the index is less than zero. In what proportion of years is the market down?

1. *State the problem.* Call the annual rate of return for Standard & Poor's 500-stock Index x . The variable x has the $N(12, 16.5)$ distribution. We want the proportion of years with $x < 0$.
2. *Standardize.* Subtract the mean, then divide by the standard deviation, to turn x into a standard Normal z :

$$\begin{aligned} x &< 0 \\ \frac{x - 12}{16.5} &< \frac{0 - 12}{16.5} \\ z &< -0.73 \end{aligned}$$

Figure 1.24 shows the standard Normal curve with the area of interest shaded.

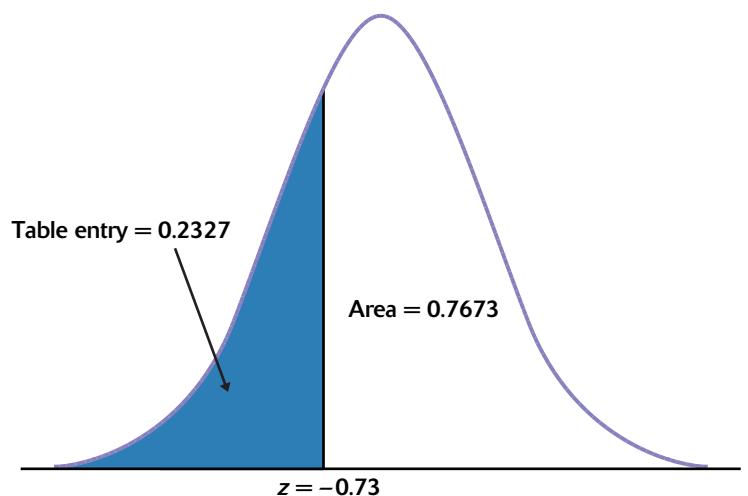


FIGURE 1.24 Areas under the standard Normal curve for Example 1.20.

3. *Use the table.* From Table A, we see that the proportion of observations less than -0.73 is 0.2327 . The market is down on an annual basis about 23.27% of the time. Notice that the area to the right of -0.73 is $1 - 0.2327 = 0.7673$. This indicates that the market is up about 76.73% of the time.

In a Normal distribution, the proportion of observations with $x < 0$ is the same as the proportion with $x \leq 0$. There is no area under the curve and exactly over 0, so the areas under the curve with $x < 0$ and $x \leq 0$ are the same. This isn't true of the actual data. There may be a year with exactly a 0% return. The Normal distribution is an easy-to-use approximation, not a description of every detail in the actual data.



The key to using either software or Table A to do a Normal calculation is to sketch the area you want, then match that area with the areas that the table or software gives you. The interactive *Normal Curve* applet available at www.whfreeman.com/pbs allows you to work directly with any Normal curve, finding areas by changing a shaded area with your computer's mouse. You can use this applet to do any Normal distribution problem, though the accuracy of its output is limited. Here is another example of the use of Table A.

EXAMPLE 1.21

More Normal distribution calculations

What percent of years have annual rates of return between 12% and 50%?

1. *State the problem.* We want the proportion of years with $12 \leq x \leq 50$.
2. *Standardize:*

$$\begin{aligned} 12 &\leq x \leq 50 \\ \frac{12 - 12}{16.5} &\leq \frac{x - 12}{16.5} \leq \frac{50 - 12}{16.5} \\ 0 &\leq z \leq 2.30 \end{aligned}$$

Figure 1.25 shows the area under the standard Normal curve.

3. *Use the table.* The area between 0 and 2.30 is the area below 2.30 minus the area below 0. Look at Figure 1.25 to check this. From Table A,

$$\begin{aligned} \text{area between } 0 \text{ and } 2.30 &= \text{area below } 2.30 - \text{area below } 0.00 \\ &= 0.9893 - 0.5000 = 0.4893 \end{aligned}$$

About 49% of years have annual rates of return between 12% and 50%.

Sometimes we encounter a value of z more extreme than those appearing in Table A. For example, the area to the left of $z = -4$ is not given directly in the table. The z -values in Table A leave only area 0.0002 in each tail unaccounted for. For practical purposes, we can act as if there is zero area outside the range of Table A.

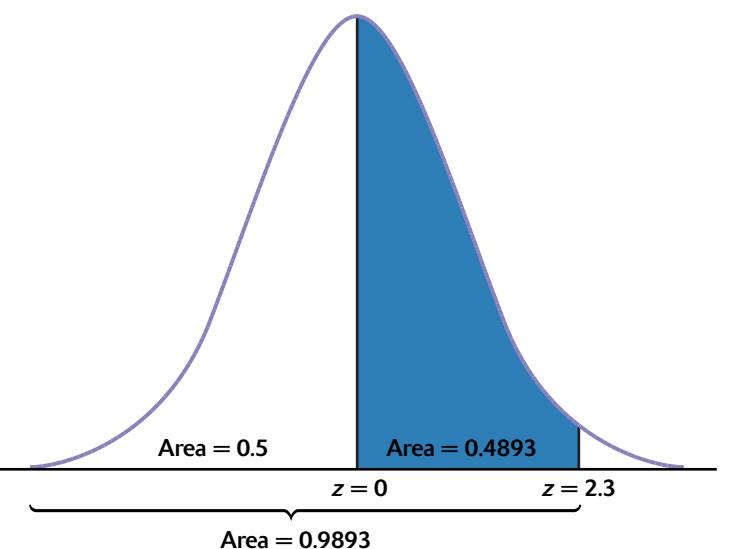


FIGURE 1.25 Areas under the standard Normal curve for Example 1.21.

APPLY YOUR KNOWLEDGE

- 1.66 Use Table A to find the proportion of observations from a standard Normal distribution that satisfies each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that is the answer to the question.
- $z < 2.85$
 - $z > 2.85$
 - $z > -1.66$
 - $-1.66 < z < 2.85$
- 1.67 **MPG for 2001 model year vehicles.** The miles per gallon ratings for 2001 model year vehicles vary according to an approximately Normal distribution with mean $\mu = 21.22$ miles per gallon and standard deviation $\sigma = 5.36$ miles per gallon.
- What percent of vehicles have miles per gallon ratings greater than 30?
 - What percent of vehicles have miles per gallon ratings between 30 and 35?
 - What percent of vehicles have miles per gallon ratings less than 12.45?

Finding a value when given a proportion

Examples 1.19 through 1.21 illustrate the use of Table A to find what proportion of the observations satisfies some condition, such as “annual rate of return between 12% and 50%.” We may instead want to find the observed value with a given proportion of the observations above or below it. To do this, use Table A backward. Find the given proportion in the body of the table, read the corresponding z from the left column and top row, then “unstandardize” to get the observed value. Here is an example.

EXAMPLE 1.22**“Backward” Normal calculations**

Miles per gallon ratings of compact cars (2001 model year) follow approximately the $N(25.7, 5.88)$ distribution. How many miles per gallon must a vehicle get to place in the top 10% of all 2001 model year compact cars?

- State the problem.* We want to find the miles per gallon rating x with area 0.1 to its right under the Normal curve with mean $\mu = 25.7$ and standard deviation $\sigma = 5.88$. That's the same as finding the miles per gallon rating x with area 0.9 to its left. Figure 1.26 poses the question in graphical form. Because Table A gives the areas to the left of z -values, always state the problem in terms of the area to the left of x .
- Use the table.* Look in the body of Table A for the entry closest to 0.9. It is 0.8997. This is the entry corresponding to $z = 1.28$. So $z = 1.28$ is the standardized value with area 0.9 to its left.
- Unstandardize* to transform the solution from the z back to the original x scale. We know that the standardized value of the unknown x is $z = 1.28$. So x itself satisfies

$$\frac{x - 25.7}{5.88} = 1.28$$

Solving this equation for x gives

$$x = 25.7 + (1.28)(5.88) = 33.2$$

This equation should make sense: it says that x lies 1.28 standard deviations above the mean on this particular Normal curve. That is the “unstandardized” meaning of $z = 1.28$. We see that a compact car must receive a rating of at least 33.2 miles per gallon to place in the highest 10%.

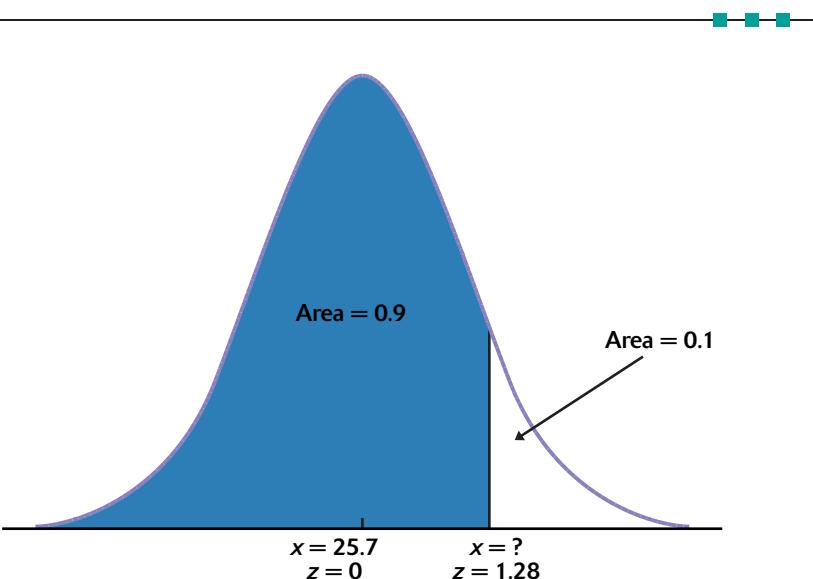


FIGURE 1.26 Locating the point on a Normal curve with area 0.10 to its right, for Example 1.22.

Here is the general formula for unstandardizing a z -score. To find the value x from the Normal distribution with mean μ and standard deviation σ corresponding to a given standard Normal value z , use

$$x = \mu + z\sigma$$

APPLY YOUR KNOWLEDGE

- 1.68 Use Table A to find the value z of a standard Normal variable that satisfies each of the following conditions. (Use the value of z from Table A that comes closest to satisfying the condition.) In each case, sketch a standard Normal curve with your value of z marked on the axis.
- (a) The point z with 25% of the observations falling below it.
 - (b) The point z with 40% of the observations falling above it.
- 1.69 **GMAT scores.** Most graduate schools of business require applicants for admission to take the Graduate Management Admission Council's GMAT examination.¹⁹ Total scores on the GMAT for the more than 500,000 people who took the exam between April 1997 and March 2000 are roughly Normally distributed with mean $\mu = 527$ and standard deviation $\sigma = 112$.
- (a) What percent of test takers have scores above 500?
 - (b) What GMAT scores fall in the lowest 25% of the distribution?
 - (c) How high a GMAT score is needed to be in the highest 5%?

Assessing the Normality of data

The Normal distributions provide good models for some distributions of real data. Examples include the miles per gallon ratings of 2001 model year vehicles, average payrolls of major league baseball teams, and statewide unemployment rates. The distributions of some other common variables are usually skewed and therefore distinctly nonnormal. Examples include variables such as personal income, gross sales of business firms, and the service lifetime of mechanical or electronic components. While experience can suggest whether or not a Normal model is plausible in a particular case, it is risky to assume that a distribution is Normal without actually inspecting the data.

The decision to describe a distribution by a Normal model may determine the later steps in our analysis of the data. Both calculations of proportions and statistical inference based on such calculations follow from the choice of a model. How can we judge whether data are approximately Normal?

A histogram or stemplot can reveal distinctly nonnormal features of a distribution, such as outliers, pronounced skewness, or gaps and clusters. If the stemplot or histogram appears roughly symmetric and single-peaked, however, we need a more sensitive way to judge the adequacy of a Normal model. The most useful tool for assessing Normality is another graph, the **Normal quantile plot**.*

Here is the idea of a simple version of a Normal quantile plot. It is not feasible to make Normal quantile plots by hand, but software makes them for us, using more sophisticated versions of this basic idea.

1. Arrange the observed data values from smallest to largest. Record what percentile of the data each value occupies. For example, the smallest

Normal quantile plot

*Some software calls these graphs *Normal probability plots*. There is a technical distinction between the two types of graphs, but the terms are often used loosely.

observation in a set of 20 is at the 5% point, the second smallest is at the 10% point, and so on.

2. Do Normal distribution calculations to find the z -scores at these same percentiles. For example, $z = -1.645$ is the 5% point of the standard Normal distribution, and $z = -1.282$ is the 10% point.
3. Plot each data point x against the corresponding z . If the data distribution is close to standard Normal, the plotted points will lie close to the 45-degree line $x = z$. If the data distribution is close to any Normal distribution, the plotted points will lie close to some straight line.

Any Normal distribution produces a straight line on the plot because standardizing turns any Normal distribution into a standard Normal. Standardizing is a linear transformation that can change the slope and intercept of the line in our plot but cannot turn a line into a curved pattern.

USE OF NORMAL QUANTILE PLOTS

If the points on a Normal quantile plot lie close to a straight line, the plot indicates that the data are Normal. Systematic deviations from a straight line indicate a non-Normal distribution. Outliers appear as points that are far away from the overall pattern of the plot.

Figures 1.27 to 1.29 are Normal quantile plots for data we have met earlier. The data x are plotted vertically against the corresponding standard Normal z -score plotted horizontally. The z -score scales extend from -3 to 3 because almost all of a standard Normal curve lies between these values. These figures show how Normal quantile plots behave.

EXAMPLE 1.23



Interpreting Normal quantile plots

Figure 1.27 is a Normal quantile plot of the earnings of 15 black female hourly workers at National Bank. Most of the points lie close to a straight line, indicating that a Normal model fits well. The low outlier lies below the line formed by the other observations. That is, this observation is farther out in the low direction than we would expect in a Normal distribution. We can describe this distribution as “roughly Normal, with one low outlier.” Compare the stemplot of these data in Figure 1.10 on page 31.

EXAMPLE 1.24

Salary data aren't Normal

Figure 1.28 is a Normal quantile plot of the Cincinnati Reds salary data. The histogram in Figure 1.4 (page 16) shows that the distribution is right-skewed. To see the right skewness in the Normal quantile plot, draw a line through the points at the left of the plot, which correspond to the smaller observations. The larger observations fall systematically above this line. That is, the right-of-center observations have larger values than in a Normal distribution. *In a right-skewed distribution, the largest observations fall distinctly above a line drawn through the main body of points.* Similarly, left skewness is evident when the smallest observations fall below the line. Unlike Figure 1.27, there are no individual outliers.

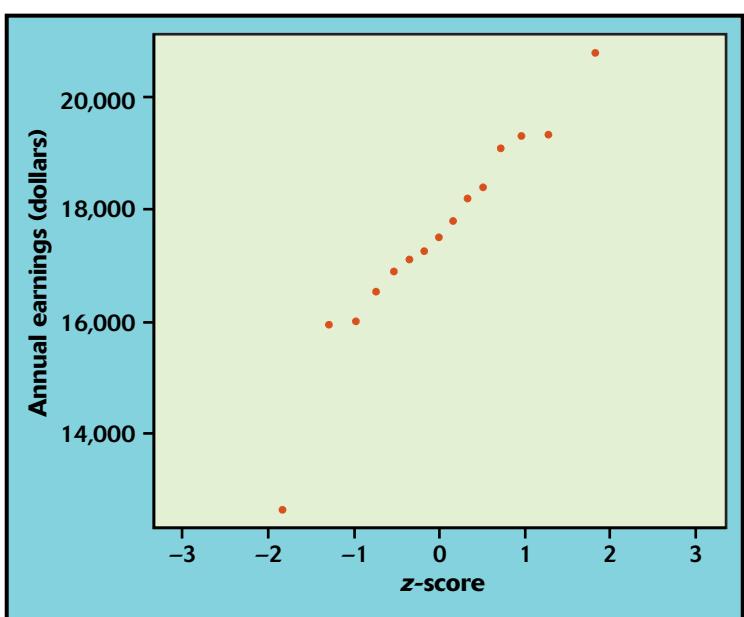


FIGURE 1.27 Normal quantile plot of the earnings of 15 black female hourly workers at National Bank, for Example 1.23. This distribution is roughly Normal except for one low outlier.

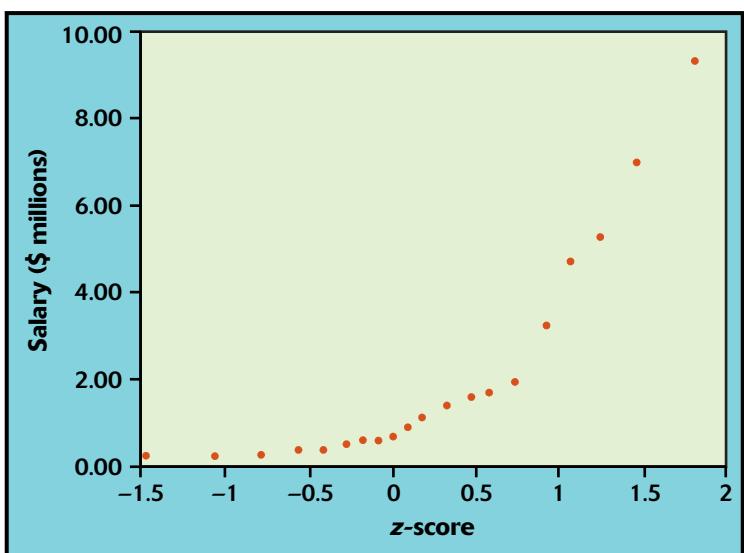


FIGURE 1.28 Normal quantile plot of the salaries of Cincinnati Reds players on opening day of the 2000 season, for Example 1.24. This distribution is skewed to the right.

EXAMPLE 1.25

Unemployment rates are roughly Normal

Figure 1.29 is a Normal quantile plot of the December 2000 unemployment rates in the 50 states, from Table 1.1. The plot is quite straight, showing good fit to a Normal model.

Normal quantile plots are designed specifically to assess Normality. The histograms in Figure 1.2 (ignoring Puerto Rico) (page 12) and the stemplots



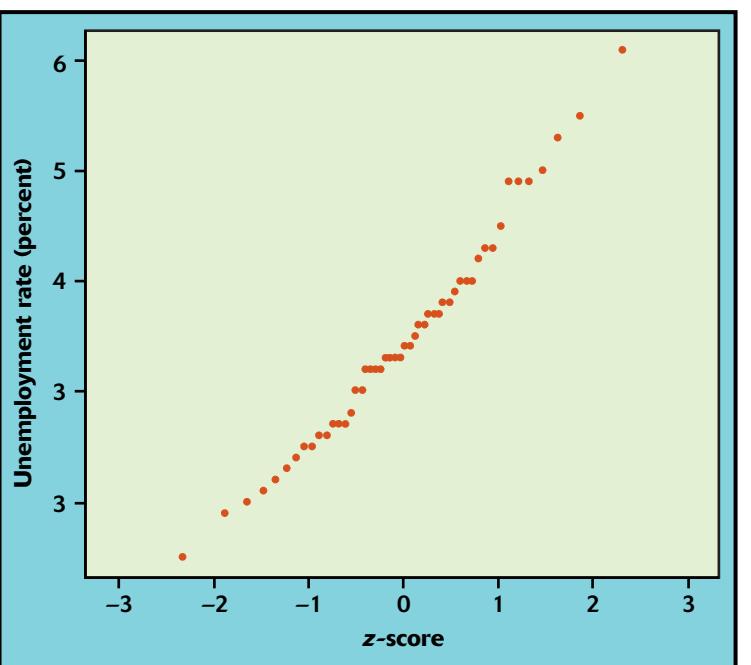


FIGURE 1.29 Normal quantile plot of the December 2000 unemployment rates in the 50 states, for Example 1.25. This distribution is quite Normal.

in Figure 1.6 (page 19) show us that the distribution of unemployment rates is symmetric and single-peaked, but they cannot tell us how closely it follows the distinctive Normal shape.

Figure 1.29 does, of course, show some deviation from a straight line. Real data almost always show some departure from the theoretical Normal model. It is important to confine your examination of a Normal quantile plot to searching for shapes that show *clear departures from Normality*. Don't overreact to minor wiggles in the plot. When we discuss statistical methods that are based on the Normal model, we will pay attention to the sensitivity of each method to departures from Normality. Many common methods work well as long as the data are reasonably symmetric and outliers are not present.

- APPLY YOUR KNOWLEDGE**
- 1.70 Manufacturers measure dimensions of their products to monitor the performance of production processes. A maker of electric meters measures the distance between two mounting holes on 27 consecutive meters after the holes are formed in production.²⁰ Figure 1.30 is a Normal quantile plot of these distances. The measurements are in hundredths of an inch. The large-scale pattern is quite Normal, but the measurements deviate from Normality in an interesting way. What explains the horizontal runs of points in the plot?
- 1.71 Figure 1.31 is a Normal quantile plot of the monthly percent returns for U.S. common stocks from June 1950 to June 2000. Because there are 601 observations, the individual points in the middle of the plot run together. In what way do monthly returns on stocks deviate from a Normal distribution?

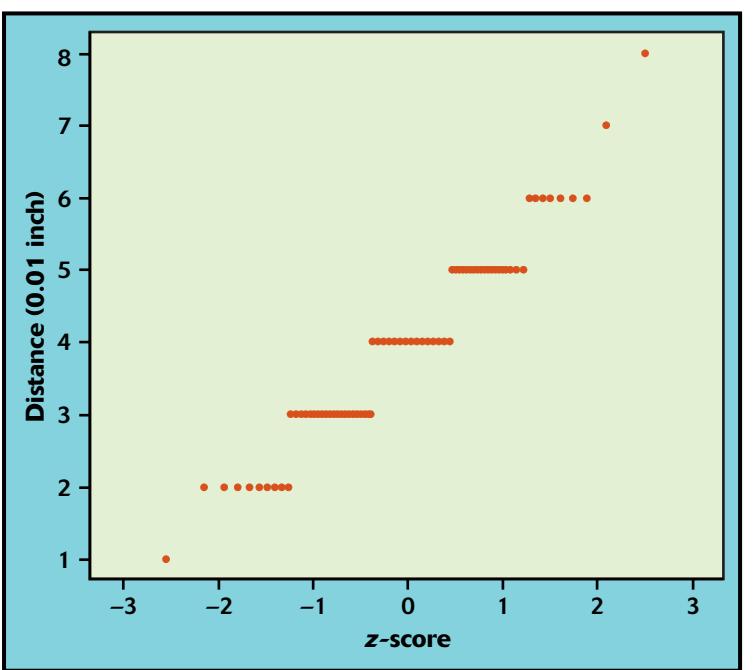


FIGURE 1.30 Normal quantile plot of the distances between two mounting holes on 27 electric meters, for Exercise 1.70.

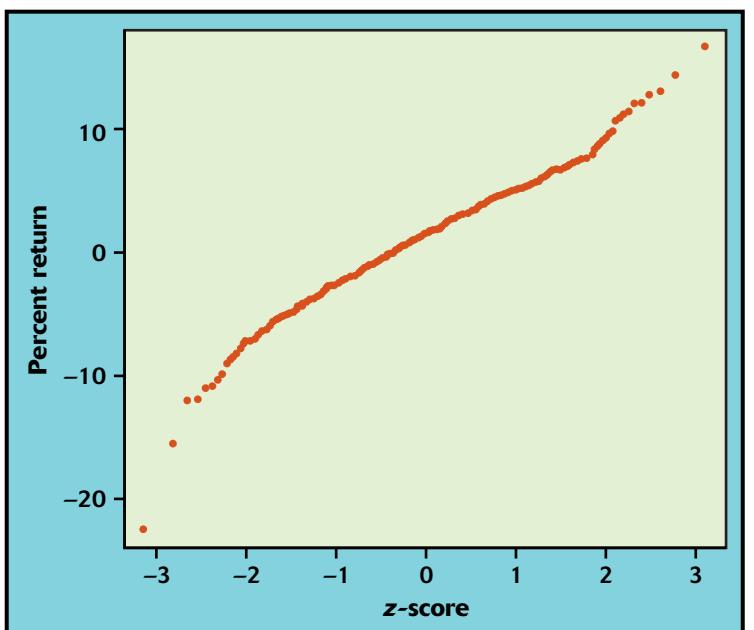


FIGURE 1.31 Normal quantile plot of the percent returns on U.S. common stocks in 601 consecutive months, for Exercise 1.71.

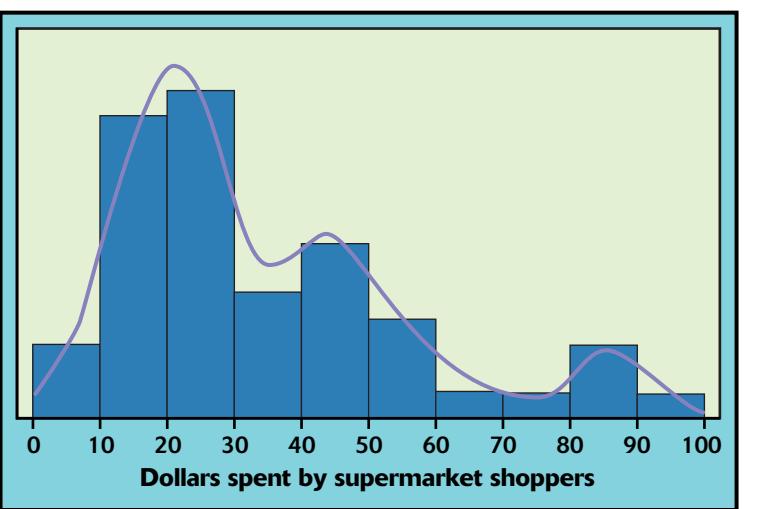


FIGURE 1.32 A density curve for the supermarket shopper data of Exercise 1.11, calculated from the data by density estimation software.

BEYOND THE BASICS: DENSITY ESTIMATION

A density curve gives a compact summary of the overall shape of a distribution. Figure 1.14 (page 51) shows a Normal density curve that provides a good summary of the distribution of miles per gallon ratings for 2001 model year vehicles. Many distributions do not have the Normal shape. There are other families of density curves that are used as mathematical models for various distribution shapes.

density estimation

Modern software offers a more flexible option, **density estimation**. A density estimator does not start with any specific shape, such as the Normal shape. It looks at the data and draws a density curve that describes the overall shape of the data. Figure 1.32 shows a histogram of the right-skewed supermarket shopper data from Exercise 1.11 (page 19). A density estimator produced the density curve drawn over the histogram. You can see that this curve does catch the overall pattern: the strong right skewness, the major peak near \$20, and the two smaller clusters on the right side. Density estimation offers a quick way to picture the overall shapes of distributions. It is another useful tool for exploratory data analysis.

SECTION 1.3 SUMMARY

- We can sometimes describe the overall pattern of a distribution by a **density curve**. A density curve has total area 1 underneath it. An area under a density curve gives the proportion of observations that fall in a range of values.
- A density curve is an idealized description of the overall pattern of a distribution that smooths out the irregularities in the actual data. We write the mean of a density curve as μ and the standard deviation of a density curve as σ to distinguish them from the mean \bar{x} and standard deviation s of the actual data.

- The mean, the median, and the quartiles of a density curve can be located by eye. The **mean** μ is the balance point of the curve. The **median** divides the area under the curve in half. The **quartiles** and the median divide the area under the curve into quarters. The **standard deviation** σ cannot be located by eye on most density curves.
- The mean and median are equal for symmetric density curves. The mean of a skewed curve is located farther toward the long tail than is the median.
- The **Normal distributions** are described by a special family of bell-shaped, symmetric density curves, called **Normal curves**. The mean μ and standard deviation σ completely specify a Normal distribution $N(\mu, \sigma)$. The mean is the center of the curve, and σ is the distance from μ to the change-of-curvature points on either side.
- To **standardize** any observation x , subtract the mean of the distribution and then divide by the standard deviation. The resulting **z -score**

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

says how many standard deviations x lies from the distribution mean.

- All Normal distributions are the same when measurements are transformed to the standardized scale. In particular, all Normal distributions satisfy the **68–95–99.7 rule**, which describes what percent of observations lie within one, two, and three standard deviations of the mean.
- If x has the $N(\mu, \sigma)$ distribution, then the **standardized variable** $z = (x - \mu)/\sigma$ has the **standard Normal distribution** $N(0, 1)$ with mean 0 and standard deviation 1. Table A gives the proportions of standard Normal observations that are less than z for many values of z . By standardizing, we can use Table A for any Normal distribution.
- The adequacy of a Normal model for describing a distribution of data is best assessed by a **Normal quantile plot**, which is available in most statistical software packages. A pattern on such a plot that deviates substantially from a straight line indicates that the data are not Normal.

SECTION 1.3 EXERCISES

- 1.72 Figure 1.33 shows two Normal curves, both with mean 0. Approximately what is the standard deviation of each of these curves?
- 1.73 The Environmental Protection Agency requires that the exhaust of each model of motor vehicle be tested for the level of several pollutants. The level of oxides of nitrogen (NOX) in the exhaust of one light truck model was found to vary among individual trucks according to a Normal distribution with mean $\mu = 1.45$ grams per mile driven and standard deviation $\sigma = 0.40$ grams per mile. Sketch the density curve of this Normal distribution, with

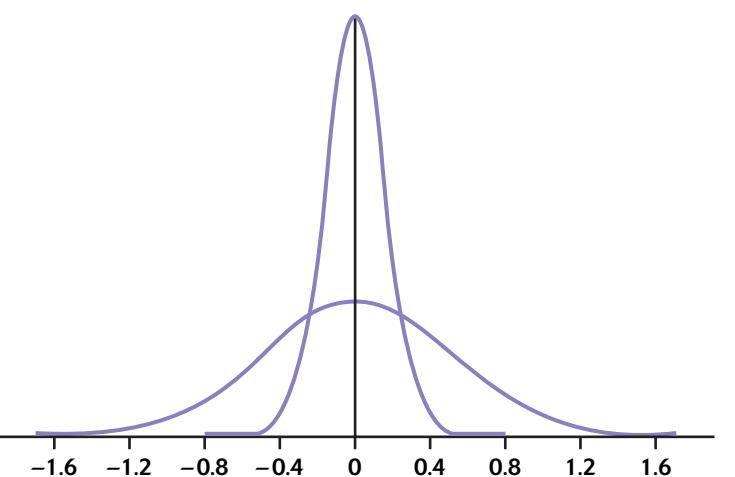


FIGURE 1.33 Two Normal curves with the same mean but different standard deviations, for Exercise 1.72.

the scale of grams per mile marked on the horizontal axis. Also, give an interval that contains the middle 95% of NOX levels in the exhaust of trucks using this Normal model.

- 1.74 **Length of pregnancies.** Some health insurance companies treat pregnancy as a “preexisting condition” when it comes to paying for maternity expenses for a new policyholder. Sometimes the exact date of conception is unknown, so the insurance company must count back from the doctor’s expected due date to judge whether or not conception occurred before or after the new policy began. The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days. Use the 68–95–99.7 rule to answer the following questions.
- Between what values do the lengths of the middle 95% of all pregnancies fall?
 - How short are the shortest 2.5% of all pregnancies?
 - How likely is it that a woman with an expected due date 218 days after her policy began conceived the child after her policy began?

- 1.75 Use Table A to find the proportion of observations from a standard Normal distribution that falls in each of the following regions. In each case, sketch a standard Normal curve and shade the area representing the region.

- $z \leq -2.25$
- $z \geq -2.25$
- $z > 1.77$
- $-2.25 < z < 1.77$

- 1.76
 - Find the number z such that the proportion of observations that are less than z in a standard Normal distribution is 0.8.
 - Find the number z such that 35% of all observations from a standard Normal distribution are greater than z .

- 1.77 **NCAA rules for athletes.** The National Collegiate Athletic Association (NCAA) requires Division I athletes to score at least 820 on the combined mathematics and verbal parts of the SAT exam to compete in their first college year. (Higher scores are required for students with poor high school grades.) In 2000, the scores of the 1,260,000 students taking the SATs were approximately Normal with mean 1019 and standard deviation 209. What percent of all students had scores less than 820?
- 1.78 **More NCAA rules.** The NCAA considers a student a “partial qualifier” eligible to practice and receive an athletic scholarship, but not to compete, if the combined SAT score is at least 720. Use the information in the previous exercise to find the percent of all SAT scores that are less than 720.
- 1.79 **The stock market.** The yearly rate of return on stock indexes (which combine many individual stocks) is approximately Normal. Between 1950 and 2000, U.S. common stocks had a mean yearly return of about 13%, with a standard deviation of about 17%. Take this Normal distribution to be the distribution of yearly returns over a long period.
- In what range do the middle 95% of all yearly returns lie?
 - The market is down for the year if the return is less than zero. In what percent of years is the market down?
 - In what percent of years does the index gain 25% or more?
- 1.80 **Length of pregnancies.** The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days.
- What percent of pregnancies last less than 240 days (that's about 8 months)?
 - What percent of pregnancies last between 240 and 270 days (roughly between 8 months and 9 months)?
 - How long do the longest 20% of pregnancies last?
- 1.81 **Quartiles of Normal distributions.** The median of any Normal distribution is the same as its mean. We can use Normal calculations to find the quartiles for Normal distributions.
- What is the area under the standard Normal curve to the left of the first quartile? Use this to find the value of the first quartile for a standard Normal distribution. Find the third quartile similarly.
 - Your work in (a) gives the z -scores for the quartiles of any Normal distribution. What are the quartiles for the lengths of human pregnancies? (Use the distribution in Exercise 1.80.)
- 1.82 **Deciles of Normal distributions.** The *deciles* of any distribution are the points that mark off the lowest 10% and the highest 10%. On a density curve, these are the points with areas 0.1 and 0.9 to their left under the curve.
- What are the deciles of the standard Normal distribution?
 - The weights of 9-ounce potato chip bags are approximately Normal with mean 9.12 ounces and standard deviation 0.15 ounces. What are the deciles of this distribution?

The remaining exercises for this section require the use of software that will make Normal quantile plots.

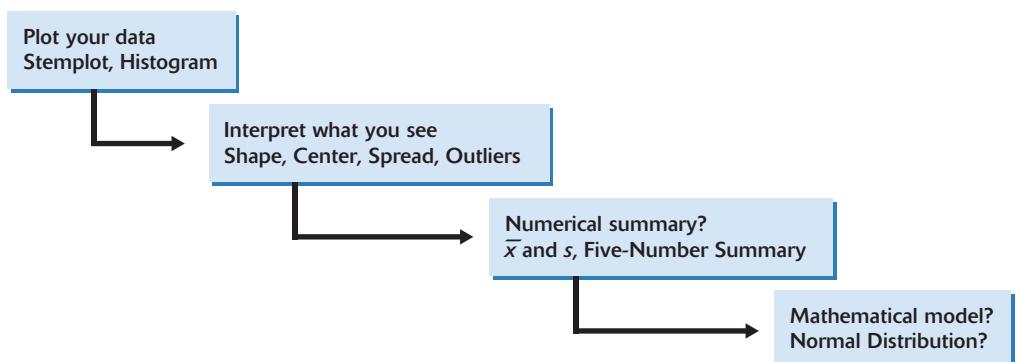
- 1.83 **Are monthly returns on a stock Normal?** Is the distribution of monthly returns on Philip Morris stock approximately Normal with the exception of possible outliers? Make a Normal quantile plot of the data in Table 1.10 and report your conclusions.
- 1.84 **Normal random numbers.** Use software to generate 100 observations from the standard Normal distribution. Make a histogram of these observations. How does the shape of the histogram compare with a Normal density curve? Make a Normal quantile plot of the data. Does the plot suggest any important deviations from Normality? (Repeating this exercise several times is a good way to become familiar with how Normal quantile plots look when data actually are close to Normal.)
- 1.85 **Uniform random numbers.** Use software to generate 100 observations from the distribution described in Exercise 1.60 (page 55). (The software will probably call this a “Uniform distribution.”) Make a histogram of these observations. How does the histogram compare with the density curve in Figure 1.18? Make a Normal quantile plot of your data. According to this plot, how does the Uniform distribution deviate from Normality?

STATISTICS IN SUMMARY

Data analysis is the art of describing data using graphs and numerical summaries. The purpose of data analysis is to describe the most important features of a set of data. This chapter introduces data analysis by presenting statistical ideas and tools for describing the distribution of a single variable. The Statistics in Summary figure below will help you organize the big ideas. The question marks at the last two stages remind us that the usefulness of numerical summaries and models such as Normal distributions depends on what we find when we examine the data using graphs. Here is a review list of the most important skills you should have acquired from your study of this chapter.

A. DATA

1. Identify the individuals and variables in a set of data.
2. Identify each variable as categorical or quantitative. Identify the units in which each quantitative variable is measured.



B. DISPLAYING DISTRIBUTIONS

1. Make a bar graph, pie chart, and/or Pareto chart of the distribution of a categorical variable. Interpret bar graphs, pie charts, and Pareto charts.
2. Make a histogram of the distribution of a quantitative variable.
3. Make a stemplot of the distribution of a small set of observations. Round leaves or split stems as needed to make an effective stemplot.

C. INSPECTING DISTRIBUTIONS (QUANTITATIVE VARIABLE)

1. Look for the overall pattern and for major deviations from the pattern.
2. Assess from a histogram or stemplot whether the shape of a distribution is roughly symmetric, distinctly skewed, or neither. Assess whether the distribution has one or more major peaks.
3. Describe the overall pattern by giving numerical measures of center and spread in addition to a verbal description of shape.
4. Decide which measures of center and spread are more appropriate: the mean and standard deviation (especially for symmetric distributions) or the five-number summary (especially for skewed distributions).
5. Recognize outliers.

D. TIME PLOTS

1. Make a time plot of data, with the time of each observation on the horizontal axis and the value of the observed variable on the vertical axis.
2. Recognize strong trends or other patterns in a time plot.

E. MEASURING CENTER

1. Find the mean \bar{x} of a set of observations.
2. Find the median M of a set of observations.
3. Understand that the median is more resistant (less affected by extreme observations) than the mean. Recognize that skewness in a distribution moves the mean away from the median toward the long tail.

F. MEASURING SPREAD

1. Find the quartiles Q_1 and Q_3 for a set of observations.
2. Give the five-number summary and draw a boxplot; assess center, spread, symmetry, and skewness from a boxplot.
3. Using a calculator or software, find the standard deviation s for a set of observations.
4. Know the basic properties of s : $s \geq 0$ always; $s = 0$ only when all observations are identical and increases as the spread increases; s has the same units as the original measurements; s is pulled strongly up by outliers or skewness.

G. DENSITY CURVES

1. Know that areas under a density curve represent proportions of all observations and that the total area under a density curve is 1.
2. Approximately locate the median (equal-areas point) and the mean (balance point) on a density curve.
3. Know that the mean and median both lie at the center of a symmetric density curve and that the mean moves farther toward the long tail of a skewed curve.

H. NORMAL DISTRIBUTIONS

1. Recognize the shape of Normal curves and be able to estimate by eye both the mean and the standard deviation from such a curve.
2. Use the 68–95–99.7 rule and symmetry to state what percent of the observations from a Normal distribution fall between two points when both points lie at the mean or one, two, or three standard deviations on either side of the mean.
3. Find the standardized value (z -score) of an observation. Interpret z -scores and understand that any Normal distribution becomes standard Normal $N(0, 1)$ when standardized.
4. Given that a variable has the Normal distribution with a stated mean μ and standard deviation σ , calculate the proportion of values above a stated number, below a stated number, or between two stated numbers.
5. Given that a variable has the Normal distribution with a stated mean μ and standard deviation σ , calculate the point having a stated proportion of all values above it. Also calculate the point having a stated proportion of all values below it.
6. Assess the Normality of a set of data by inspecting a Normal quantile plot.

CHAPTER 1 REVIEW EXERCISES

- 1.86 **Household and family income.** In government data, a household contains all people who live together in a residence. A family consists of two or more people living together and related by blood, marriage, or adoption. In 1999, the mean and median of household incomes were \$40,816 and \$54,842. The mean and median of family incomes were \$48,950 and \$62,636.
- (a) Which of each pair is the mean and which is the median? How do you know?
 - (b) Why are the mean and median incomes higher for families than for households?
- 1.87 **What influences buying?** Product preference depends in part on the age, income, and gender of the consumer. A market researcher selects a large sample of potential car buyers. For each consumer, she records gender, age, household income, and automobile preference. Which of these variables are categorical and which are quantitative?

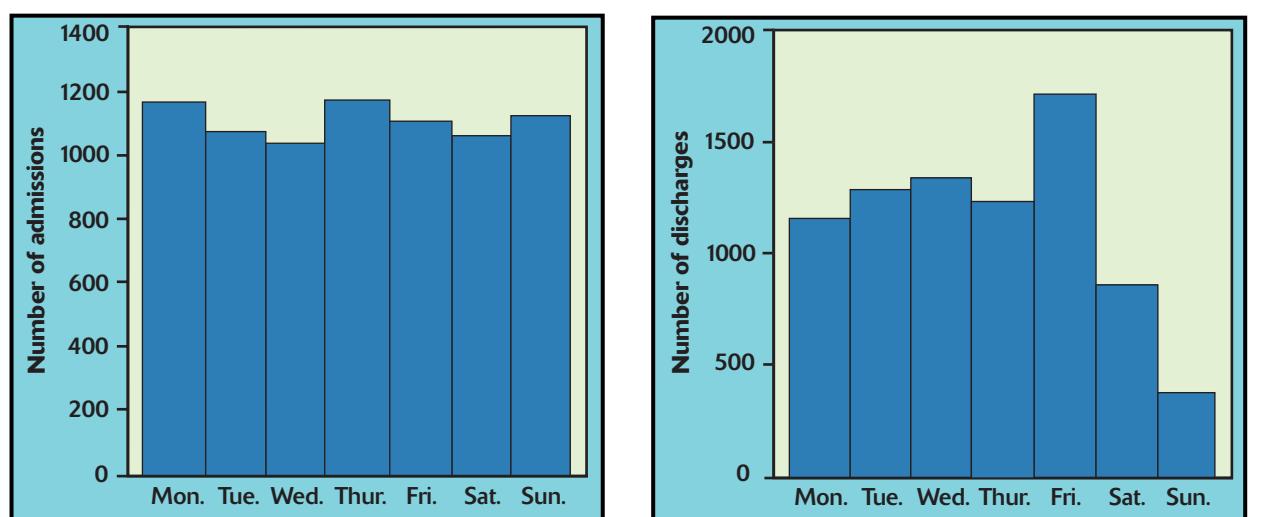


FIGURE 1.34 Bar graphs of the number of heart attack victims admitted and discharged on each day of the week by hospitals in Ontario, Canada, for Exercise 1.88.

- 1.88 **Never on Sunday?** The Canadian province of Ontario carries out statistical studies of the working of Canada's national health care system in the province. The bar graphs in Figure 1.34 come from a study of admissions and discharges from community hospitals in Ontario.²¹ They show the number of heart attack patients admitted and discharged on each day of the week during a 2-year period.
- Explain why you expect the number of patients admitted with heart attacks to be roughly the same for all days of the week. Do the data show that this is true?
 - Describe how the distribution of the day on which patients are discharged from the hospital differs from that of the day on which they are admitted. What do you think explains the difference?
- 1.89 **Yankee salaries.** Few companies release their employees' salaries. The baseball players' union, however, makes player salaries public. Table 1.11 contains the salaries of the New York Yankees as of opening day of the 2001 season.²² Describe the distribution of Yankee salaries, giving a graph and numerical measures to back up your description.
- 1.90 **Stock returns.** Table 1.10 (page 49) gives the monthly percent returns on Philip Morris stock for the period from June 1990 to July 2001. The data appear in time order reading from left to right across each row in turn, beginning with the 3.0% return in June 1990. Make a time plot of the data. This was a period of increasing action against smoking, so we might expect a trend toward lower returns. But it was also a period in which stocks in general rose sharply, which would produce an increasing trend. What does your time plot show?
- 1.91 **Better corn.** Corn is an important animal food. Normal corn lacks certain amino acids, which are building blocks for protein. Plant scientists have developed new corn varieties that have more of these amino acids. To test a new corn as an animal food, a group of 20 one-day-old male chicks was fed a ration containing the new corn. A control group of another 20 chicks was

TABLE 1.11 2001 salaries for the New York Yankees baseball team

Player	Salary	Player	Salary	Player	Salary
Derek Jeter	12,600,000	Jorge Posada	4,050,000	Shane Spencer	320,000
Bernie Williams	12,357,143	Mike Stanton	2,450,000	Todd Williams	320,000
Roger Clemens	10,300,000	Orlando Hernandez	2,050,000	Carlos Almanzar	270,000
Mike Mussina	10,000,000	Allen Watson	1,700,000	Clay Bellinger	230,000
Mariano Rivera	9,150,000	Ramiro Mendoza	1,600,000	Darrell Einertson	206,000
David Justice	7,000,000	Joe Oliver	1,100,000	Randy Choate	204,750
Andy Pettitte	7,000,000	Henry Rodriguez	850,000	Michael Coleman	204,000
Paul O'Neill	6,500,000	Alfonso Soriano	630,000	D'Angelo Jimenez	200,000
Chuck Knoblauch	6,000,000	Luis Sojo	500,000	Christian Parker	200,000
Tino Martinez	6,000,000	Brian Boehringer	350,000	Scott Seabol	200,000
Scott Brosius	5,250,000				

fed a ration that was identical except that it contained normal corn. Here are the weight gains (in grams) after 21 days:²³

Normal corn				New corn			
380	321	366	356	361	447	401	375
283	349	402	462	434	403	393	426
356	410	329	399	406	318	467	407
350	384	316	272	427	420	477	392
345	455	360	431	430	339	410	326

- (a) Compute five-number summaries for the weight gains of the two groups of chicks. Then make boxplots to compare the two distributions. What do the data show about the effect of the new corn?
 - (b) The researchers actually reported means and standard deviations for the two groups of chicks. What are they? How much larger is the mean weight gain of chicks fed the new corn?
- 1.92 Do SUVs waste gas? Table 1.2 (page 13) gives the highway fuel consumption (in miles per gallon) for 32 midsize 2001 model year cars. Here are the highway mileages for 19 four-wheel-drive 2001 sport-utility vehicle models:²⁴

Model	MPG	Model	MPG
Acura MDX	23	Jeep Wrangler	19
Chevrolet Blazer	22	Land Rover	15
Chevrolet Tahoe	18	Mazda Tribute	24
Dodge Durango	17	Mercedes-Benz ML320	21
Ford Expedition	18	Mitsubishi Montero	20
Ford Explorer	19	Nissan Pathfinder	18
Honda Passport	20	Suzuki Vitara	25
Infiniti QX4	19	Toyota RAV4	27
Isuzu Trooper	19	Toyota 4Runner	19
Jeep Grand Cherokee	19		

- (a) Give a graphical and numerical description of highway fuel consumption for sport-utility vehicles. What are the main features of the distribution?
- (b) Make boxplots to compare the highway fuel consumption of midsize cars and sport-utility vehicles. What are the most important differences between the two distributions?
- 1.93 **How much oil?** How much oil the wells in a given field will ultimately produce is key information in deciding whether to drill more wells. Table 1.12 gives the estimated total amount of oil recovered from 64 wells in the Devonian Richmond Dolomite area of the Michigan basin.²⁵
- (a) Graph the distribution and describe its main features.
- (b) Find the mean and median of the amounts recovered. Explain how the relationship between the mean and the median reflects the shape of the distribution.
- (c) Give the five-number summary and explain briefly how it reflects the shape of the distribution.
- 1.94 **The $1.5 \times IQR$ rule.** Exercise 1.58 (page 50) describes the most common rule for identifying suspected outliers. Find the interquartile range IQR for the oil recovery data in the previous exercise. Are there any outliers according to the $1.5 \times IQR$ rule?
- 1.95 **Grading managers.** Some companies “grade on a bell curve” to compare the performance of their managers and professional workers. This forces the use of some low performance ratings, so that not all workers are graded “above average.” Until the threat of lawsuits forced a change, Ford Motor Company’s “performance management process” assigned 10% A grades, 80% B grades, and 10% C grades to the company’s 18,000 managers. It isn’t clear that the “bell curve” of ratings is really a Normal distribution. Nonetheless, suppose that Ford’s performance scores are Normally distributed. This year, managers with scores less than 25 received C’s and those with scores above 475 received A’s. What are the mean and standard deviation of the scores?
- 1.96 Table 1.7 (page 29) reports data on the states. Much more information is available. Do your own exploration of differences among the states. Find in the library or at the U.S. Census Bureau Web site (www.census.gov) the most recent edition of the annual *Statistical Abstract of the United States*. Look up data on (a) the number of businesses started (“business starts”) and (b) the number of business failures for the 50 states. Make

TABLE 1.12 Ultimate recovery (thousands of barrels) for 64 oil wells

21.71	53.2	46.4	42.7	50.4	97.7	103.1	51.9
43.4	69.5	156.5	34.6	37.9	12.9	2.5	31.4
79.5	26.9	18.5	14.7	32.9	196	24.9	118.2
82.2	35.1	47.6	54.2	63.1	69.8	57.4	65.6
56.4	49.4	44.9	34.6	92.2	37.0	58.8	21.3
36.6	64.9	14.8	17.6	29.1	61.4	38.6	32.5
12.0	28.3	204.9	44.5	10.3	37.7	33.7	81.1
12.1	20.1	30.5	7.1	10.1	18.0	3.0	2.0

TABLE 1.13 Population of California counties, 2000

County	Population	County	Population	County	Population
Alameda	1,443,741	Marin	247,289	San Mateo	707,161
Alpine	1,208	Mariposa	17,130	Santa Barbara	399,347
Amador	35,100	Mendocino	86,265	Santa Clara	1,682,585
Butte	203,171	Merced	210,554	Santa Cruz	255,602
Calaveras	40,554	Modoc	9,449	Shasta	163,256
Colusa	18,804	Mono	12,853	Sierra	3,555
Contra Costa	948,816	Monterey	401,762	Siskiyou	44,301
Del Norte	27,507	Napa	124,279	Solano	394,542
El Dorado	156,299	Nevada	92,033	Sonoma	458,614
Fresno	799,407	Orange	2,846,289	Stanislaus	446,997
Glenn	26,453	Placer	248,399	Sutter	78,930
Humboldt	126,518	Plumas	20,824	Tehama	56,039
Imperial	142,361	Riverside	1,545,387	Trinity	13,022
Inyo	17,945	Sacramento	1,223,499	Tulare	368,021
Kern	661,645	San Benito	53,234	Tuolumne	54,501
Kings	129,461	San Bernardino	1,709,434	Ventura	753,197
Lake	58,309	San Diego	2,813,833	Yolo	168,660
Lassen	33,828	San Francisco	776,733	Yuba	60,219
Los Angeles	9,519,338	San Joaquin	563,598		
Madera	123,109	San Luis Obispo	246,681		

graphs and numerical summaries to display the distributions and write a brief description of the most important characteristics of each distribution. Suggest an explanation for any outliers you see.

- 1.97 You are planning a sample survey of households in California. You decide to select households separately within each county and to choose more households from the more populous counties. To aid in the planning, Table 1.13 gives the populations of California counties from the 2000 census.²⁶ Examine the distribution of county populations both graphically and numerically, using whatever tools are most suitable. Write a brief description of the main features of this distribution. Sample surveys often select households from all of the most populous counties but from only some of the less populous. How would you divide California counties into three groups according to population, with the intent of including all of the first group, half of the second, and a smaller fraction of the third in your survey?

The next two exercises require the use of software that will make Normal quantile plots.

- 1.98 Exercise 1.91 (page 79) presents data on the weight gains of chicks fed two types of corn. The researchers used \bar{x} and s to summarize each of the two distributions. Make a Normal quantile plot for each group and report your findings. Is use of \bar{x} and s justified?

- 1.99 Most statistical software packages have routines for generating values of variables having specified distributions. Use your statistical software to generate 25 observations from the $N(20, 5)$ distribution. Compute the mean and standard deviation \bar{x} and s of the 25 values you obtain. How close are \bar{x} and s to the μ and σ of the distribution from which the observations were drawn?

Repeat 20 times the process of generating 25 observations from the $N(20, 5)$ distribution and recording \bar{x} and s . Make a stemplot of the 20 values of \bar{x} and another stemplot of the 20 values of s . Make Normal quantile plots of both sets of data. Briefly describe each of these distributions. Are they symmetric or skewed? Are they roughly Normal? Where are their centers? (The distributions of measures like \bar{x} and s when repeated sets of observations are made from the same theoretical distribution will be very important in later chapters.)



CHAPTER 1 CASE STUDY EXERCISES

CASE STUDY 1.1: **Individual incomes.** Each March, the Bureau of Labor Statistics collects detailed information about more than 50,000 randomly selected households. The data set *individuals.dat* contains data on 55,899 people from the March 2000 survey. The Data Appendix describes this data set in detail.

- A. **All workers.** Give a brief description of the distribution of incomes for these people, using graphs and numbers to report your findings. Because this is a very large randomly selected sample, your results give a good description of individual incomes for all Americans aged 25 to 65 who work outside of agriculture.
- B. **Comparing sectors of the economy.** Do a statistical analysis to compare the incomes of people whose main work experience is (1) in the private sector, (2) in government, and (3) self-employed. Use graphs and numerical descriptions to report your findings.

CASE STUDY 1.2: **Bank employees.** Our analysis of Case 1.2 (page 30) uncovered a racial disparity in the earnings of hourly workers at National Bank. Using the data in the file *hourly.dat*, we found that black workers earn systematically less than whites. The data we have allow one more step. There are 1276 full-time hourly workers and 469 part-time employees. We expect that part-time workers will earn less on the average than full-time workers. If blacks are more often employed part-time, the black/white gap in earnings might be due in part to this difference in job status.

How do earnings differ for the two main job classifications? Use graphs and numerical descriptions to back up your report. Does your analysis show that part-time workers as a group earn less than full-time hourly employees? Now compare the distributions of earnings for black and white employees separately for the two job types and write a brief discussion of your findings.

Prelude

How much for my house?

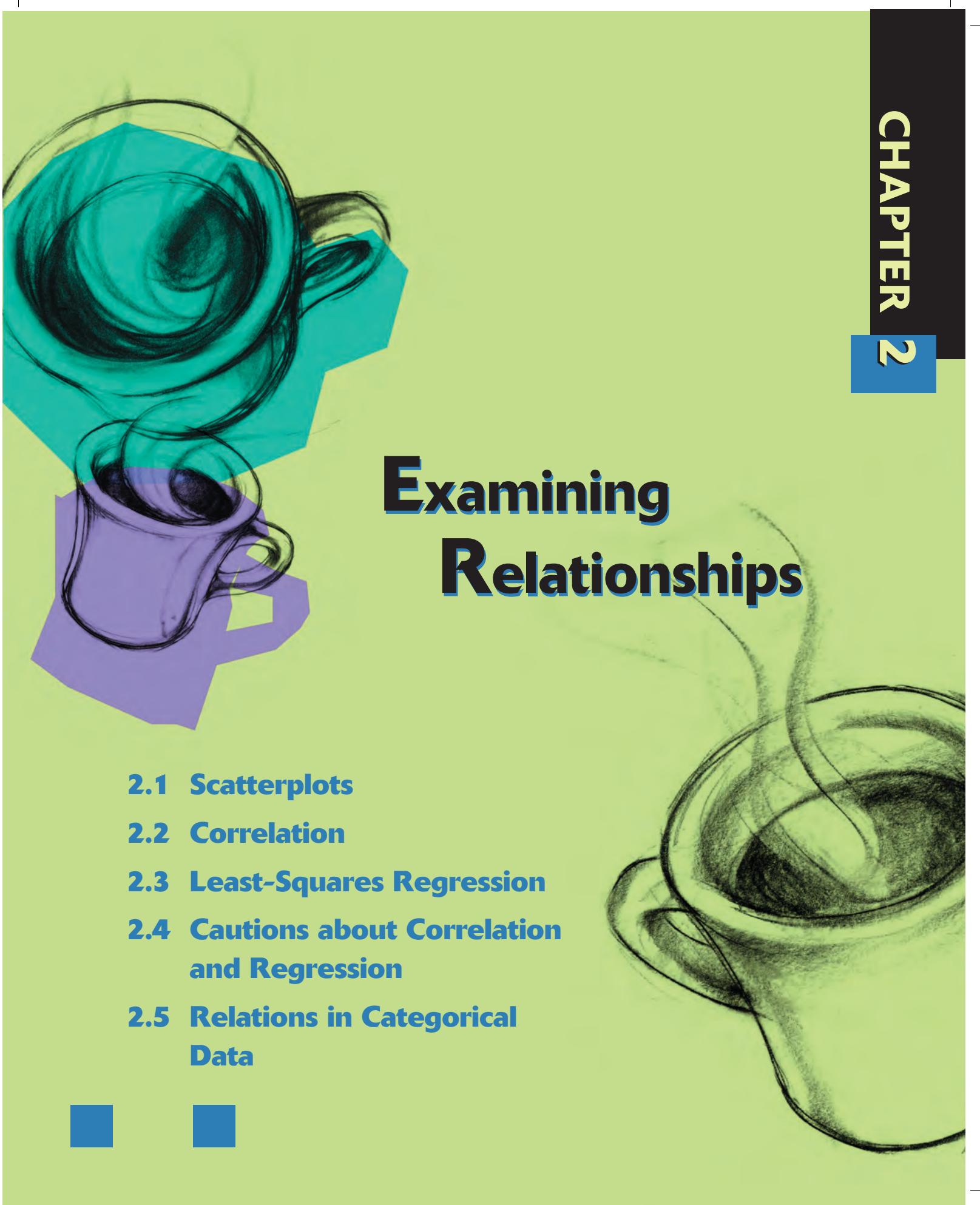
As the mortgage officer for the local bank branch, Matthew helps individuals make decisions about buying and selling houses. A couple planning to buy a new house wonders how much their current house will sell for. The city assessor's office provides *assessed values* for all houses in town for tax purposes. An assessed value is an estimate of the value of a home used to determine property taxes. Matthew knows that most realtors in town tell someone trying to sell a house that they should expect to sell the house for approximately 10% above its assessed value, but he isn't sure about this "rule of thumb."

What is the relationship between selling price and assessed value for houses in town? Matthew took an introductory business statistics course in college, and he recalls studying relationships between two variables like selling price and assessed value. A trip to the city assessor's office provides a sample of the selling prices and assessed values for recently sold houses. A scatterplot shows that the relationship between selling prices and assessed values can be described by a line with a positive slope of approximately 1.07. The value of the slope indicates that selling prices are about 7% above assessed values in Matthew's sample of recently sold houses. While some houses did sell for 10% above their assessed value, others sold for smaller percentages above assessed.

Based on actual sales data, the slope of Matthew's line will provide his customers with a more reasonable assumption about selling prices compared to the "optimistic" 10% rule of thumb the realtors use.

Examining Relationships

- 2.1 Scatterplots**
- 2.2 Correlation**
- 2.3 Least-Squares Regression**
- 2.4 Cautions about Correlation and Regression**
- 2.5 Relations in Categorical Data**



Introduction

A marketing study finds that men spend more money online than women. An insurance group reports that heavier cars have fewer deaths per 10,000 vehicles registered than do lighter cars. These and many other statistical studies look at the relationship between two variables. To understand such a relationship, we must often examine other variables as well. To conclude that men spend more online, for example, the researchers had to eliminate the effect of other variables such as annual income. Our topic in this chapter is relationships between variables. One of our main themes is that the relationship between two variables can be strongly influenced by other variables that are lurking in the background.

Because variation is everywhere, statistical relationships are overall tendencies, not ironclad rules. They allow individual exceptions. Although smokers on the average die younger than nonsmokers, some people live to 90 while smoking three packs a day. To study a relationship between two variables, we measure both variables on the same individuals. Often, we think that one of the variables explains or influences the other.

RESPONSE VARIABLE, EXPLANATORY VARIABLE

A response variable measures an outcome of a study. An explanatory variable explains or influences changes in a response variable.

*independent
variable
dependent variable*

You will often find explanatory variables called **independent variables**, and response variables called **dependent variables**. The idea behind this language is that the response variable depends on the explanatory variable. Because the words “independent” and “dependent” have other meanings in statistics that are unrelated to the explanatory-response distinction, we prefer to avoid those words.

It is easiest to identify explanatory and response variables when we actually set values of one variable to see how it affects another variable.

EXAMPLE 2.1

The best price?

Price is important to consumers and therefore to retailers. Sales of an item typically increase as its price falls, except for some luxury items, where high price suggests exclusivity. The seller’s profits for an item often increase as the price is reduced, due to increased sales, until the point at which lower profit per item cancels rising sales. A retail chain therefore introduces a new DVD player at several different price points and monitors sales. The chain wants to discover the price at which its profits are greatest. Price is the explanatory variable, and total profit from sales of the player is the response variable.



When we don’t set the values of either variable but just observe both variables, there may or may not be explanatory and response variables. Whether there are depends on how we plan to use the data.

EXAMPLE 2.2**Inventory and sales**

Jim is a district manager for a retail chain. He wants to know how the average monthly inventory and monthly sales for the stores in his district are related to each other. Jim doesn't think that either inventory level or sales explains or causes the other. He has two related variables, and neither is an explanatory variable.

Sue manages another district for the same chain. She asks, "Can I predict a store's monthly sales if I know its inventory level?" Sue is treating the inventory level as the explanatory variable and the monthly sales as the response variable.

In Example 2.1, price differences actually *cause* differences in profits from sales of DVD players. There is no cause-and-effect relationship between inventory levels and sales in Example 2.2. Because inventory and sales are closely related, we can nonetheless use a store's inventory level to predict its monthly sales. We will learn how to do the prediction in Section 2.3. Prediction requires that we identify an explanatory variable and a response variable. Some other statistical techniques ignore this distinction. Do remember that calling one variable explanatory and the other response doesn't necessarily mean that changes in one *cause* changes in the other.

Most statistical studies examine data on more than one variable. Fortunately, statistical analysis of several-variable data builds on the tools we used to examine individual variables. The principles that guide our work also remain the same:

- First plot the data, then add numerical summaries.
- Look for overall patterns and deviations from those patterns.
- When the overall pattern is quite regular, use a compact mathematical model to describe it.

APPLY YOUR KNOWLEDGE

- 2.1 In each of the following situations, is it more reasonable to simply explore the relationship between the two variables or to view one of the variables as an explanatory variable and the other as a response variable? In the latter case, which is the explanatory variable and which is the response variable?
 - (a) The amount of time a student spends studying for a statistics exam and the grade on the exam
 - (b) The weight and height of a person
 - (c) The amount of yearly rainfall and the yield of a crop
 - (d) An employee's salary and number of sick days used
 - (e) The economic class of a father and of a son
- 2.2 **Stock prices.** How well does a stock's market price at the beginning of the year predict its price at the end of the year? To find out, record the price of a large group of stocks at the beginning of the year, wait until the end of the year, then record their prices again. What are the explanatory and response variables here? Are these variables categorical or quantitative?
- 2.3 **Hand wipes.** Antibacterial hand wipes can irritate the skin. A company wants to compare two different formulas for new wipes. Investigators choose two groups of adults at random. Each group uses one type of wipes. After

several weeks, a doctor assesses whether or not each person's skin appears abnormally irritated. What are the explanatory and response variables? Are they categorical or quantitative variables?

2.1 Scatterplots



SALES AT A RETAIL SHOP

Three entrepreneurial women open Duck Worth Wearing, a shop selling high-quality secondhand children's clothing, toys, and furniture. Items are consigned to the shop by individuals who then receive a percentage of the selling price of their items. Table 2.1 displays daily data for April 2000 on sales at Duck Worth Wearing.¹ Larger retail establishments keep more elaborate records, particularly relating to sales of specific items or branded lines of goods. The consignment shop does not sell multiple copies of the same item, though records of sales by categories of item (clothing, toys, and so on) might be useful for planning and advertising.

The data in Table 2.1 concern each day's gross sales, broken down by method of payment (cash, check, credit card). "Gross sales" in retailing are overall dollar sales, not adjusted for returns or discounts. The data record the number of items sold and the dollar amount of gross sales for each payment method on each day. We will examine relationships among some of these variables.

The most common way to display the relation between two quantitative variables is a *scatterplot*. Figure 2.1 is an example of a scatterplot.

	A	B	C	D	E	F	G	H	I
1	Date	Gross Sales	Items	Gross Cash	Cash Items	Gross Check	Check Items	Gross Credit Card	Credit Card Items
2	4/1/00	890.50	115	348.20	55	394.30	56	149.00	4
3	4/3/00	197.00	17	42.00	8	44.50	3	110.50	6
4	4/4/00	231.00	26	61.00	9	108.50	10	61.50	7
5	4/5/00	170.00	21	94.00	16	76.00	5	0.00	0
6	4/6/00	202.50	30	59.50	11	104.00	14	39.00	5
7	4/7/00	225.50	35	164.50	26	54.00	8	7.00	1
8	4/8/00	489.70	84	125.70	27	220.60	31	143.40	26
9	4/10/00	234.80	42	110.80	19	97.50	18	26.50	5
10	4/11/00	161.50	21	26.00	5	122.00	14	13.50	2
11	4/12/00	284.00	44	109.00	18	104.00	14	71.00	12
12	4/13/00	422.00	65	180.00	27	126.00	17	116.00	21
13	4/14/00	500.70	59	211.80	36	39.40	14	49.50	9
14	4/15/00	412.40	69	57.50	10	341.90	57	13.00	2
15	4/17/00	346.80	59	115.00	20	126.80	23	105.00	16
16	4/18/00	92.30	19	15.30	8	58.00	8	19.00	3
17	4/19/00	255.80	42	97.30	15	114.50	13	44.00	14
18	4/20/00	118.50	16	61.00	10	57.50	6	0.00	0
19	4/21/00	286.50	39	84.50	15	178.00	21	24.00	3
20	4/22/00	594.00	72	156.50	24	293.50	38	144.00	10
21	4/24/00	263.29	43	87.56	15	113.18	17	62.55	11
22	4/25/00	244.08	45	96.34	19	147.74	26	0.00	0
23	4/26/00	394.28	64	201.53	33	178.95	28	15.80	3
24	4/27/00	241.51	36	107.92	25	153.89	13	0.00	0
25	4/28/00	299.97	40	158.19	21	118.03	15	23.75	4
26	4/29/00	649.04	103	176.04	43	220.40	32	252.60	28
27									

TABLE 2.1 Gross sales and number of items sold.

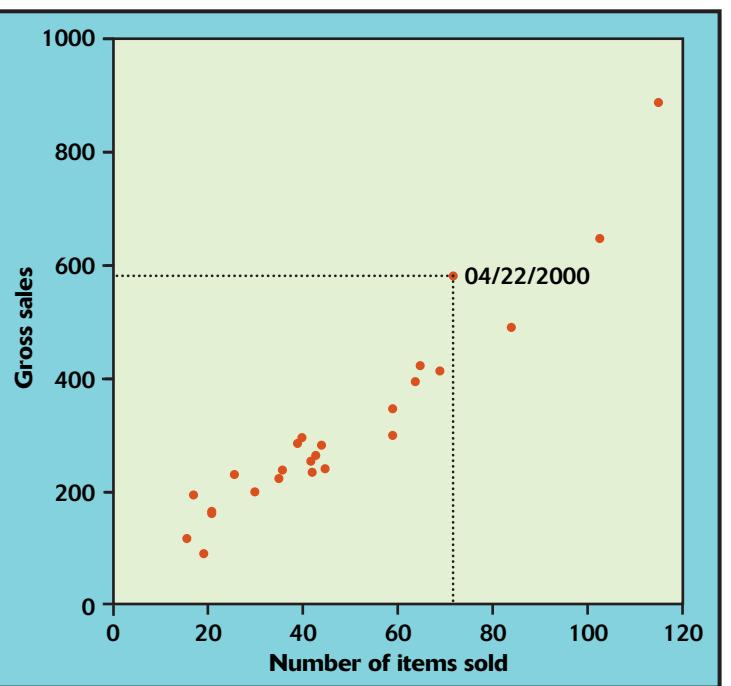


FIGURE 2.1 Scatterplot of the gross sales for each day in April 2000 against the number of items sold for the same day, from Table 2.1. The dotted lines intersect at the point (72, 594), the data for April 22, 2000.

EXAMPLE 2.3



Items sold and gross sales

The owners of Duck Worth Wearing expect a relationship between the number of items sold per day and the gross sales per day. The data appear in columns B and C of Table 2.1.

The number of items sold should help explain gross sales. Therefore, “items sold” is the explanatory variable and “gross sales” is the response variable. We want to see how gross sales change when number of items sold changes, so we put number of items sold (the explanatory variable) on the horizontal axis. Figure 2.1 is the scatterplot. Each point represents a single day of transactions in April 2000. On April 22, 2000, for example, the shop sold 72 items, and the gross sales were \$594. Find 72 on the x (horizontal) axis and 594 on the y (vertical) axis. April 22, 2000, appears as the point (72, 594) above 72 and to the right of 594. Figure 2.1 shows how to locate this point on the plot.

SCATTERPLOT

A scatterplot shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

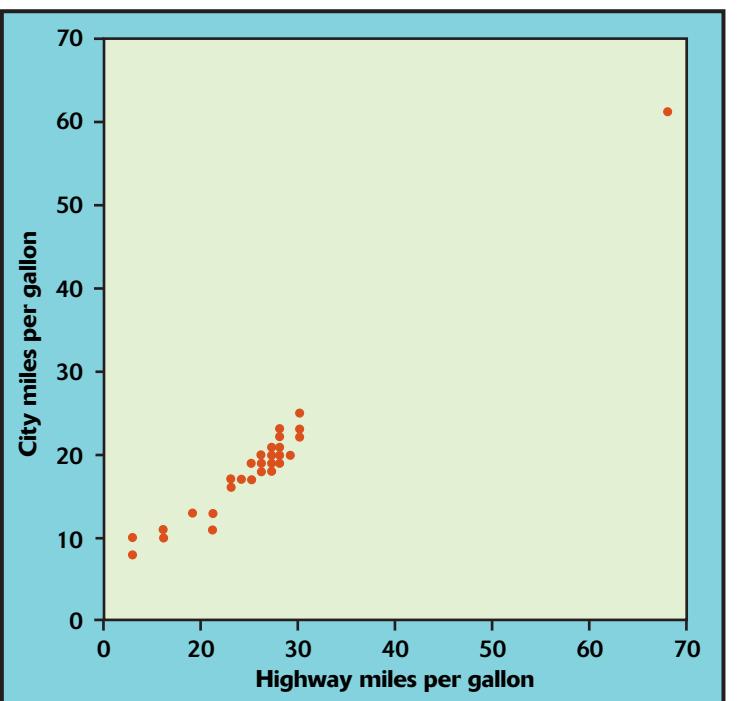


FIGURE 2.2 City and highway fuel consumption for 2001 model two-seater cars, for Exercise 2.5.

Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. As a reminder, we usually call the explanatory variable x and the response variable y . If there is no explanatory-response distinction, either variable can go on the horizontal axis.

- APPLY YOUR KNOWLEDGE**
- 2.4 **Architectural firms.** Table 1.3 (page 14) contains data describing firms engaged in commercial architecture in the Indianapolis, Indiana, area.
- We want to examine the relationship between number of full-time staff members employed and total billings. Which is the explanatory variable?
 - Make a scatterplot of these data. (Be sure to label the axes with the variable names, not just x and y .) What does the scatterplot show about the relationship between these variables?
- 2.5 **Your mileage may vary.** Figure 2.2 plots the city and highway fuel consumption of 2001 model two-seater cars, from the Environmental Protection Agency's *Model Year 2001 Fuel Economy Guide*.
- There is one unusual observation, the Honda Insight. What are the approximate city and highway gas mileages for this car?
 - Describe the pattern of the relationship between city and highway mileage. Explain why you might expect a relationship with this pattern.
 - Does the Honda Insight observation fit the overall relationship portrayed by the other two-seater cars plotted?

Interpreting scatterplots

To interpret a scatterplot, apply the strategies of data analysis learned in Chapter 1.

EXAMINING A SCATTERPLOT

In any graph of data, look for the **overall pattern** and for striking deviations from that pattern.

You can describe the overall pattern of a scatterplot by the **form**, **direction**, and **strength** of the relationship.

An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

*linear relationship
cluster*

Figure 2.1 shows a clear *form*: the data lie in a roughly straight-line, or **linear**, pattern. There are no clear outliers. That is, no points clearly fall outside the overall linear pattern. There appears to be a **cluster** of points at the left, representing low-sales days, with higher-sales days strung out to the right.

The relationship in Figure 2.1 also has a clear *direction*: days on which the shop sells more items tend to have higher gross sales, as we expect. This is a *positive association* between the two variables.

POSITIVE ASSOCIATION, NEGATIVE ASSOCIATION

Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other and below-average values also tend to occur together.

Two variables are **negatively associated** when above-average values of one tend to accompany below-average values of the other, and vice versa.

The *strength* of a relationship in a scatterplot is determined by how closely the points follow a clear form. The overall relationship in Figure 2.1 is fairly moderate—days with similar numbers of items sold show some scatter in their gross sales. Here is an example of an even stronger linear relationship.

EXAMPLE 2.4**Heating a house**

The Sanchez household is about to install solar panels to reduce the cost of heating their house. To know how much the solar panels help, they record consumption of natural gas before the panels are installed. Gas consumption is higher in cold weather, so the relationship between outside temperature and gas consumption is important.

Table 2.2 gives data for 16 months.² The response variable y is the average amount of natural gas consumed each day during the month, in hundreds of cubic feet. The explanatory variable x is the average number of heating degree-days each day during the month. (Heating degree-days are the usual measure of demand for heating. One degree-day is accumulated for each degree a day's average temperature falls below 65° F. An average temperature of 20° F, for example, corresponds to 45 degree-days.)

TABLE 2.2 Average degree-days and natural-gas consumption for the Sanchez household					
Month	Degree-days	Gas (100 cu. ft.)	Month	Degree-days	Gas (100 cu. ft.)
Nov.	24	6.3	July	0	1.2
Dec.	51	10.9	Aug.	1	1.2
Jan.	43	8.9	Sept.	6	2.1
Feb.	33	7.5	Oct.	12	3.1
Mar.	26	5.3	Nov.	30	6.4
Apr.	13	4.0	Dec.	32	7.2
May	4	1.7	Jan.	52	11.0
June	0	1.2	Feb.	30	6.9

The scatterplot in Figure 2.3 shows a *strong positive linear association*. More degree-days means colder weather, so more gas is consumed. It is a strong linear relationship because the points lie close to a line, with little scatter. If we know how cold a month is, we can predict gas consumption quite accurately from the scatterplot.

Of course, not all relationships are linear. What is more, not all relationships have a clear direction that we can describe as positive association or negative association. Exercise 2.7 gives an example that is not linear and has no clear direction.

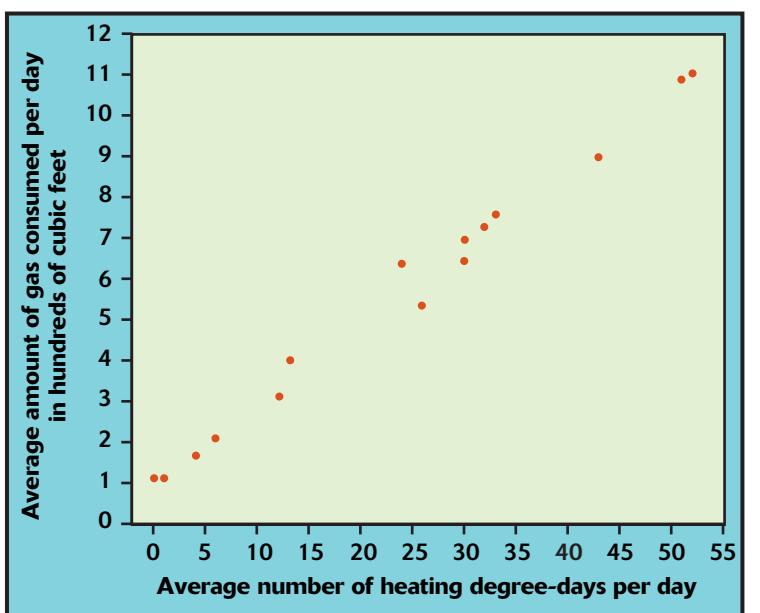


FIGURE 2.3 Scatterplot of the average amount of natural gas used per day by the Sanchez household in 16 months against the average number of heating degree-days per day in those months, from Table 2.2.

APPLY YOUR KNOWLEDGE

- 2.6 More on architectural firms. In Exercise 2.4 you made a scatterplot of number of full-time staff members employed and total billings.
- Describe the direction of the relationship. Are the variables positively or negatively associated?
 - Describe the form of the relationship. Is it linear?
 - Describe the strength of the relationship. Can the total billings of a firm be predicted accurately by the number of full-time staff members? If a firm maintains a full-time staff of 75 members, approximately what will its total billings be from year to year?
- 2.7 Does fast driving waste fuel? How does the fuel consumption of a car change as its speed increases? Here are data for a British Ford Escort. Speed is measured in kilometers per hour, and fuel consumption is measured in liters of gasoline used per 100 kilometers traveled.³

Speed (km/h)	Fuel used (liters/100 km)	Speed (km/h)	Fuel used (liter/100 km)
10	21.00	90	7.57
20	13.00	100	8.27
30	10.00	110	9.03
40	8.00	120	9.87
50	7.00	130	10.79
60	5.90	140	11.77
70	6.30	150	12.83
80	6.95		

- Make a scatterplot. (Which is the explanatory variable?)
- Describe the form of the relationship. Why is it not linear? Explain why the form of the relationship makes sense.
- It does not make sense to describe the variables as either positively associated or negatively associated. Why?
- Is the relationship reasonably strong or quite weak? Explain your answer.

Adding categorical variables to scatterplots

Every business has times of the day, days of the week, and periods in the year when sales are higher or the company is busier than other times. Duck Worth Wearing (Case 2.1) is open Monday through Saturday. The calendar provides an explanation of part of the pattern we observed in discussing Figure 2.1.

EXAMPLE 2.5**Are Saturdays different?**

Figure 2.4 enhances the scatterplot in Figure 2.1 by plotting the Saturdays with a different plot symbol (a blue **x**). The five days with the highest numbers of items sold are the five Saturdays in April 2000. While the Saturdays do stand out, Saturday sales still fit the same overall linear pattern as the weekdays.

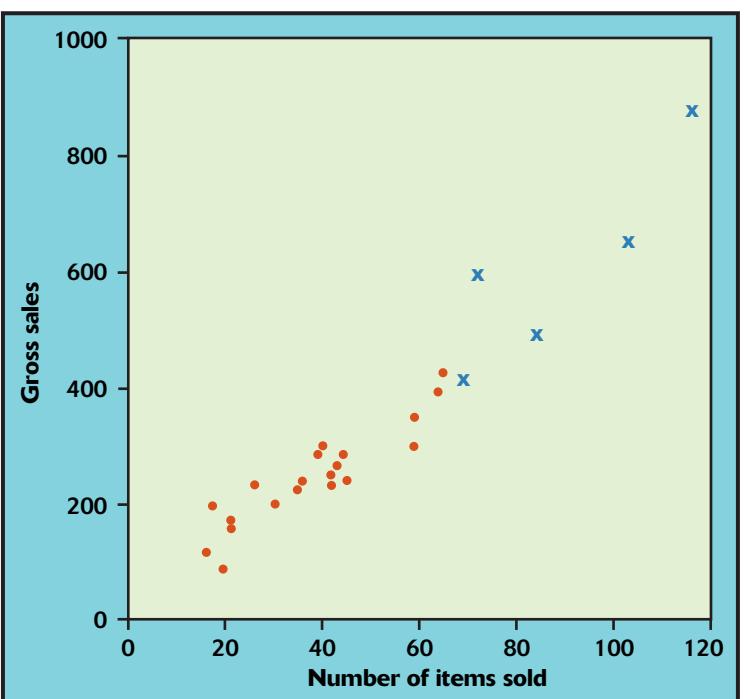


FIGURE 2.4 Gross sales at Duck Worth Wearing against number of items sold, with Saturdays highlighted.

Dividing the days into “Saturday” and “Weekday” introduces a third variable into the scatterplot. This is a categorical variable that has only two values. The two values are displayed by the two different plotting symbols. Use different colors or symbols to plot points when you want to add a categorical variable to a scatterplot.⁴

■ EXAMPLE 2.6

Do solar panels reduce gas use?

After the Sanchez household gathered the information recorded in Table 2.2 and Figure 2.3, they added solar panels to their house. They then measured their natural-gas consumption for 23 more months. To see how the solar panels affected gas consumption, add the degree-days and gas consumption for these months to the scatterplot. Figure 2.5 is the result. We use different colors to distinguish before from after. The “after” data form a linear pattern that is close to the “before” pattern in warm months (few degree-days). In colder months, with more degree-days, gas consumption after installing the solar panels is less than in similar months before the panels were added. The scatterplot shows the energy savings from the panels.



Our gas consumption example suffers from a common problem in drawing scatterplots that you may not notice when a computer does the work. When several individuals have exactly the same data, they occupy the same point on the scatterplot. Look at June and July in Table 2.2. Table 2.2 contains data for 16 months, but there are only 15 points in Figure 2.3.

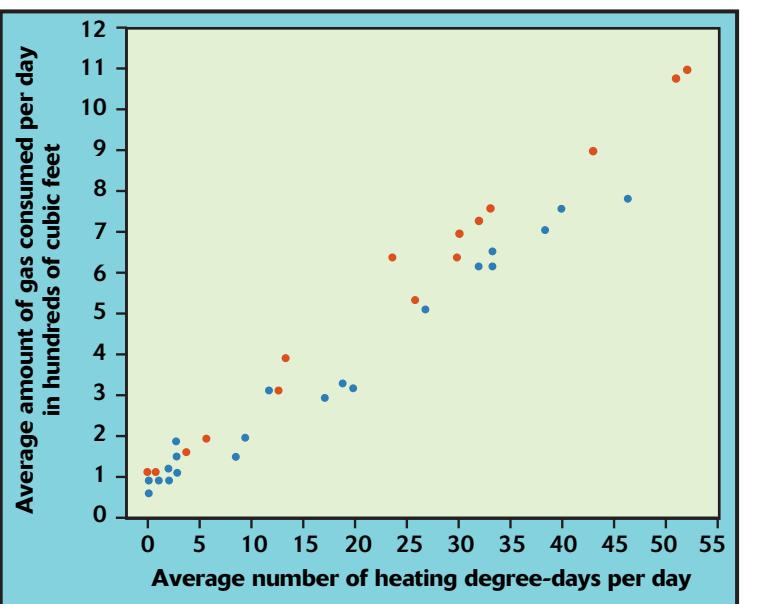


FIGURE 2.5 Natural-gas consumption against degree-days for the Sanchez household. The red observations are for 16 months before installing solar panels. The blue observations are for 23 months with the panels in use.

June and July both occupy the same point. You can use a different plotting symbol to call attention to points that stand for more than one individual. Some computer software does this automatically, but some does not. We recommend that you do use a different symbol for repeated observations when you plot a small number of observations by hand.

- 2.8 Do heavier people burn more energy? In judging the effectiveness of popular diet and exercise programs, researchers wish to take into account metabolic rate, the rate at which the body consumes energy. Table 2.3 gives data on the lean body mass and resting metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a

APPLY YOUR KNOWLEDGE

TABLE 2.3 Lean body mass and metabolic rate

Subject	Sex	Mass (kg)	Rate (cal)	Subject	Sex	Mass (kg)	Rate (cal)
1	M	62.0	1792	11	F	40.3	1189
2	M	62.9	1666	12	F	33.1	913
3	F	36.1	995	13	M	51.9	1460
4	F	54.6	1425	14	F	42.4	1124
5	F	48.5	1396	15	F	34.5	1052
6	F	42.0	1418	16	F	51.1	1347
7	M	47.4	1362	17	F	41.2	1204
8	F	50.6	1502	18	M	51.9	1867
9	F	42.0	1256	19	M	46.9	1439
10	M	48.7	1614				

person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours, the same calories used to describe the energy content of foods. The researchers believe that lean body mass is an important influence on metabolic rate.

- (a) Make a scatterplot of the data for the female subjects. Which is the explanatory variable?
- (b) Is the association between these variables positive or negative? What is the form of the relationship? How strong is the relationship?
- (c) Now add the data for the male subjects to your graph, using a different color or a different plotting symbol. Does the pattern of the relationship that you observed in (b) hold for men also? How do the male subjects as a group differ from the female subjects as a group?

SECTION 2.1 SUMMARY

- To study relationships between variables, we must measure the variables on the same group of individuals.
- If we think that a variable x may explain or even cause changes in another variable y , we call x an **explanatory variable** and y a **response variable**.
- A **scatterplot** displays the relationship between two quantitative variables measured on the same individuals. Mark values of one variable on the horizontal axis (x axis) and values of the other variable on the vertical axis (y axis). Plot each individual's data as a point on the graph.
- Always plot the explanatory variable, if there is one, on the x axis of a scatterplot. Plot the response variable on the y axis.
- Plot points with different colors or symbols to see the effect of a categorical variable in a scatterplot.
- In examining a scatterplot, look for an overall pattern showing the **form**, **direction**, and **strength** of the relationship, and then for **outliers** or other deviations from this pattern.
- **Form:** **Linear relationships**, where the points show a straight-line pattern, are an important form of relationship between two variables. Curved relationships and **clusters** are other forms to watch for.
- **Direction:** If the relationship has a clear direction, we speak of either **positive association** (high values of the two variables tend to occur together) or **negative association** (high values of one variable tend to occur with low values of the other variable).
- **Strength:** The **strength** of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line.

SECTION 2.1 EXERCISES

- 2.9 Size and length of bear markets. Are longer bear markets associated with greater market declines? Figure 2.6 is a scatterplot of the data from

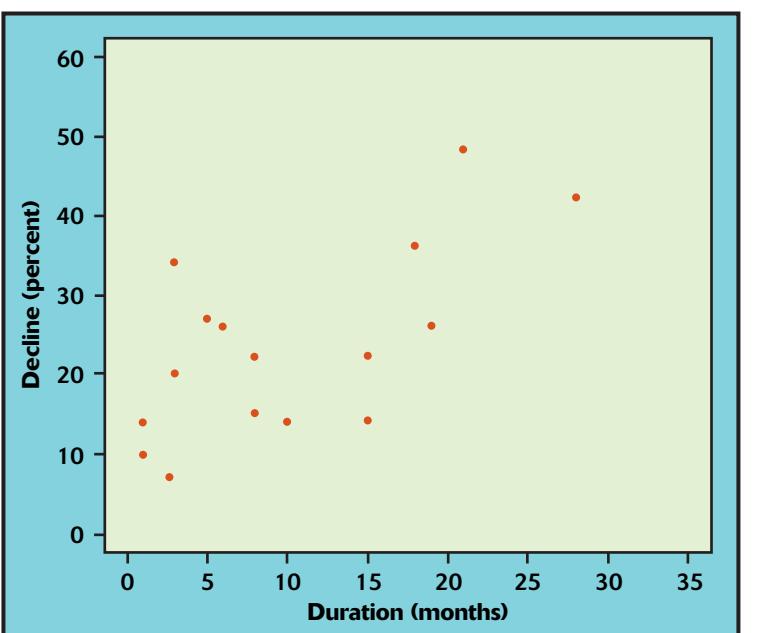


FIGURE 2.6 Scatterplot of percent decline versus duration in months of the bear markets in Table 1.5, for Exercise 2.9.

Table 1.5 (page 26) on percent decline of the market versus duration of the bear market.

- (a) Say in words what a positive association between decline and duration would mean. Does the plot show a positive association?
 - (b) What is the form of the relationship? Is it roughly linear? Is it very strong? Explain your answers.
 - (c) Without looking back at Table 1.5, what are the approximate decline and duration for the bear market that showed the greatest decline?
- 2.10 **Health and wealth.** Figure 2.7 is a scatterplot of data from the World Bank. The individuals are all the world's nations for which data are available. The explanatory variable is a measure of how rich a country is, the gross domestic product (GDP) per person. GDP is the total value of the goods and services produced in a country, converted into dollars. The response variable is life expectancy at birth.⁵ We expect people in richer countries to live longer. Describe the form, direction, and strength of the overall pattern. Does the graph confirm our expectation? The three African nations marked on the graph are outliers; a detailed study would ask what special factors explain the low life expectancy in these countries.
- 2.11 **Rich states, poor states.** One measure of a state's prosperity is the median income of its households. Another measure is the mean personal income per person in the state. Figure 2.8 is a scatterplot of these two variables, both measured in thousands of dollars. Because both variables have the same units, the plot uses equally spaced scales on both axes. The data appear in Table 1.7 (page 29).
- (a) Explain why you expect a positive association between these variables. Also explain why you expect household income to be generally higher than income per person.

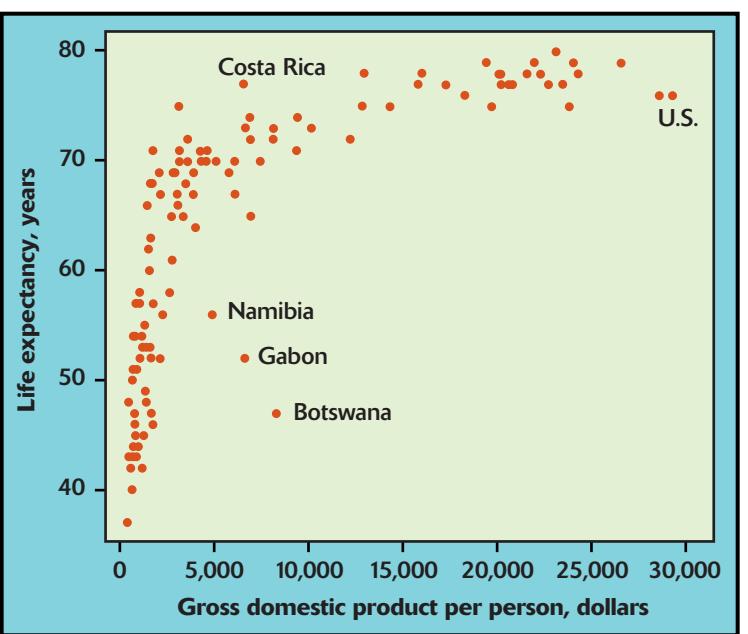


FIGURE 2.7 Scatterplot of life expectancy against gross domestic product per person for all the nations for which data are available, for Exercise 2.10.

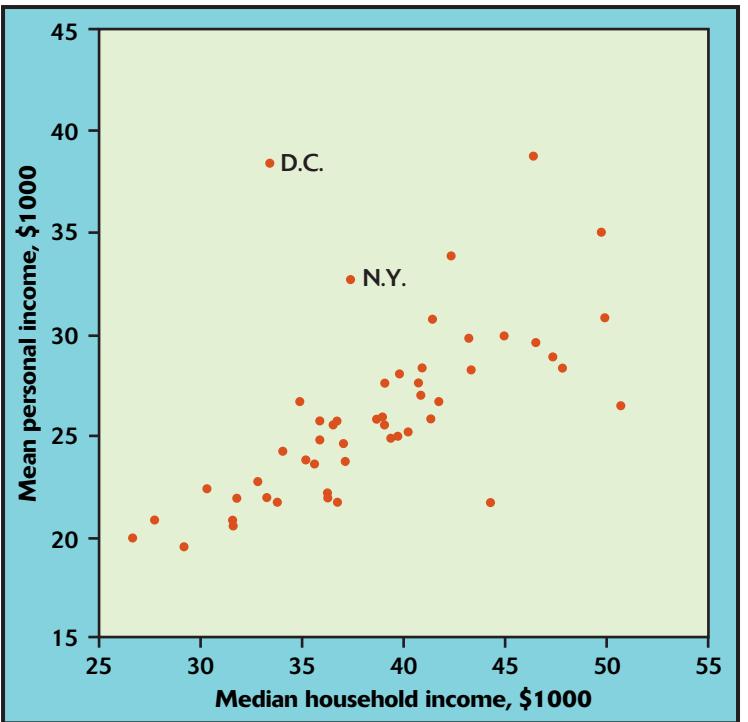


FIGURE 2.8 Scatterplot of mean income per person versus median income per household for the states, for Exercise 2.11.

- (b) Nonetheless, the mean income per person in a state can be higher than the median household income. In fact, the District of Columbia has median income \$33,433 per household and mean income \$38,429 per person. Explain why this can happen.
- (c) Describe the overall pattern of the plot, ignoring outliers.
- (d) We have labeled New York on the graph because it is a large state in which mean individual income is high relative to median household income. There are two points that are outliers clearly more extreme than New York. Which observations do these points represent? What do you know that might help explain why these points are outliers?
- 2.12 **Stocks and bonds.** How is the flow of investors' money into stock mutual funds related to the flow of money into bond mutual funds? Here are data on the net new money flowing into stock and bond mutual funds in the years 1985 to 2000, in billions of dollars.⁶ "Net" means that funds flowing out are subtracted from those flowing in. If more money leaves than arrives, the net flow will be negative. To eliminate the effect of inflation, all dollar amounts are in "real dollars" with constant buying power equal to that of a dollar in the year 2000. Make a plot (with stocks on the horizontal axis) and describe the pattern it reveals.

Year	1985	1986	1987	1988	1989	1990	1991	1992
Stocks	12.8	34.6	28.8	-23.3	8.3	17.1	50.6	97.0
Bonds	100.8	161.8	10.6	-5.8	-1.4	9.2	74.6	87.1

Year	1993	1994	1995	1996	1997	1998	1999	2000
Stocks	151.3	133.6	140.1	238.2	243.5	165.9	194.3	309.0
Bonds	84.6	-72.0	-6.8	3.3	30.0	79.2	-6.2	-48.0

- 2.13 **Calories and salt in hot dogs.** Are hot dogs that are high in calories also high in salt? Here are data for calories and salt content (milligrams of sodium) in 17 brands of meat hot dogs:⁷

Brand	Calories	Sodium (mg)	Brand	Calories	Sodium (mg)
1	173	458	10	136	393
2	191	506	11	179	405
3	182	473	12	153	372
4	190	545	13	107	144
5	172	496	14	195	511
6	147	360	15	135	405
7	146	387	16	140	428
8	139	386	17	138	339
9	175	507			

- (a) Make a scatterplot of these data, with calories on the horizontal axis. Describe the overall form, direction, and strength of the relationship. Are hot dogs that are high in calories generally also high in salt?

- (b) One brand, “Eat Slim Veal Hot Dogs,” is made of veal rather than beef and pork and positions itself as a diet brand. Which brand in the table do you think this is?
- 2.14 **How many corn plants are too many?** Midwestern farmers make many business decisions before planting. Data can help. For example, how much corn per acre should a farmer plant to obtain the highest yield? Too few plants will give a low yield. On the other hand, if there are too many plants, they will compete with each other for moisture and nutrients, and yields will fall. To find the best planting rate, plant at different rates on several plots of ground and measure the harvest. (Be sure to treat all the plots the same except for the planting rate.) Here are data from such an experiment:⁸
- | Plants per acre | Yield (bushels per acre) | | | |
|-----------------|--------------------------|-------|-------|-------|
| 12,000 | 150.1 | 113.0 | 118.4 | 142.6 |
| 16,000 | 166.9 | 120.7 | 135.2 | 149.8 |
| 20,000 | 165.3 | 130.1 | 139.6 | 149.9 |
| 24,000 | 134.7 | 138.4 | 156.1 | |
| 28,000 | 119.0 | 150.5 | | |
- (a) Is yield or planting rate the explanatory variable?
 (b) Make a scatterplot of yield and planting rate.
 (c) Describe the overall pattern of the relationship. Is it linear? Is there a positive or negative association, or neither?
 (d) Find the mean yield for each of the five planting rates. Plot each mean yield against its planting rate on your scatterplot and connect these five points with lines. This combination of numerical description and graphing makes the relationship clearer. What planting rate would you recommend to a farmer whose conditions were similar to those in the experiment?
- 2.15 **Business starts and failures.** Table 2.4 lists the number of businesses started and the number of businesses that failed by state for 1998. We might expect an association to exist between these economic measures.⁹
- (a) Make a scatterplot of business starts against business failures. Take business starts as the explanatory variable.
 (b) The plot shows a positive association between the two variables. Why do we say that the association is positive?
 (c) Find the point for Florida in the scatterplot and circle it.
 (d) There is an outlier at the upper right of the plot. Which state is this?
 (e) We wonder about clusters and gaps in the data display. There is a relatively clear cluster of states at the lower left of the plot. Four states are outside this cluster. Which states are these? Are they mainly from one part of the country?
- transformation* 2.16 **Transforming data.** Data analysts often look for a **transformation** of data that simplifies the overall pattern. Here is an example of how transforming the response variable can simplify the pattern of a scatterplot. The data show the growth of world crude oil production between 1880 and 1990.¹⁰

TABLE 2.4 Business starts and failures

State	Starts	Failures	State	Starts	Failures
AL	2,645	546	MT	397	201
AK	271	177	NE	565	383
AZ	2,868	1,225	NV	1,465	677
AR	1,091	748	NH	708	322
CA	21,582	17,679	NJ	6,412	2,024
CO	3,041	2,483	NM	887	585
CT	2,069	530	NY	13,403	4,233
DE	508	28	NC	4,371	846
DC	537	75	ND	229	144
FL	13,029	2,047	OH	4,829	2,524
GA	5,471	800	OK	1,367	990
HI	593	781	OR	1,823	1,109
ID	639	441	PA	5,525	2,641
IL	5,542	3,291	RI	544	150
IN	2,611	473	SC	2,023	410
IA	1,020	244	SD	281	275
KS	967	1,140	TN	2,835	1,369
KY	1,824	270	TX	10,936	6,785
LA	1,849	377	UT	1,417	388
ME	577	259	VT	261	80
MD	3,139	1,283	VA	3,502	860
MA	3,425	1,200	WA	2,956	2,528
MI	4,293	1,551	WV	623	305
MN	2,111	1,711	WI	2,357	1,005
MS	1,347	177	WY	213	166
MO	2,163	1,321			

Year	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
Millions of barrels	30	77	149	328	689	1412	2150	3803	7674	16,690	21,722	22,100

- (a) Make a scatterplot of production (millions of barrels) against year. Briefly describe the pattern of world crude oil production growth.
- (b) Now take the logarithm of the production in each year (use the \log button on your calculator). Plot the logarithms against year. What is the overall pattern on this plot?

2.17 Categorical explanatory variable. A scatterplot shows the relationship between two quantitative variables. Here is a similar plot to study the relationship between a categorical explanatory variable and a quantitative response variable.

Fidelity Investments, like other large mutual-fund companies, offers many “sector funds” that concentrate their investments in narrow segments

of the stock market. These funds often rise or fall by much more than the market as a whole. We can group them by broader market sector to compare returns. Here are percent total returns for 23 Fidelity “Select Portfolios” funds for the year 2000, a year in which stocks declined and technology shares were especially hard-hit:¹¹

Market sector	Fund returns (percent)					
Consumer	-9.3	29.8	-24.4	-23.1	-11.3	
Financial services	18.3	28.1	28.5	50.2	53.3	
Technology	-1.6	-30.4	-28.8	-17.5	-20.2	-32.3
Utilities and natural resources	31.8	50.3	-18.1	71.3	30.4	-13.5

- (a) Make a plot of total return against market sector (space the four market sectors equally on the horizontal axis). Compute the mean return for each sector, add the means to your plot, and connect the means with line segments.
- (b) Based on the data, which market sectors were good places to invest in 2000? Hindsight is wonderful.
- (c) Does it make sense to speak of a positive or negative association between market sector and total return?
- 2.18 Where the stores are. Target and Wal-Mart are two of the largest retail chains in the United States. Target has 977 stores and Wal-Mart has 2624 stores. The file *ex02-18.dat* on the CD that accompanies this book has the number of Target stores and Wal-Mart stores listed by state.
- (a) Use software to create a scatterplot of the number of Target stores versus the number of Wal-Mart stores for each of the 50 states.
- (b) Identify any unusual observations in your scatterplot by labeling the point with the state abbreviation. Describe specifically what about the observation makes it stand out in the scatterplot.
- (c) Comment on the form, direction, and strength of the relationship between the number of Target stores and the number of Wal-Mart stores per state.

2.2 Correlation

A scatterplot displays the form, direction, and strength of the relationship between two quantitative variables. Linear relations are particularly important because a straight line is a simple pattern that is quite common. We say a linear relation is strong if the points lie close to a straight line, and weak if they are widely scattered about a line. Our eyes are not good judges of how strong a linear relationship is. The two scatterplots in Figure 2.9 depict exactly the same data, but the lower plot is drawn smaller in a large field. The lower plot seems to show a stronger linear relationship. Our eyes can be fooled by changing the plotting scales or the amount of white space around the cloud of points in a scatterplot.¹² We need to follow our strategy for data analysis by using a numerical measure to supplement the graph. *Correlation* is the measure we use.

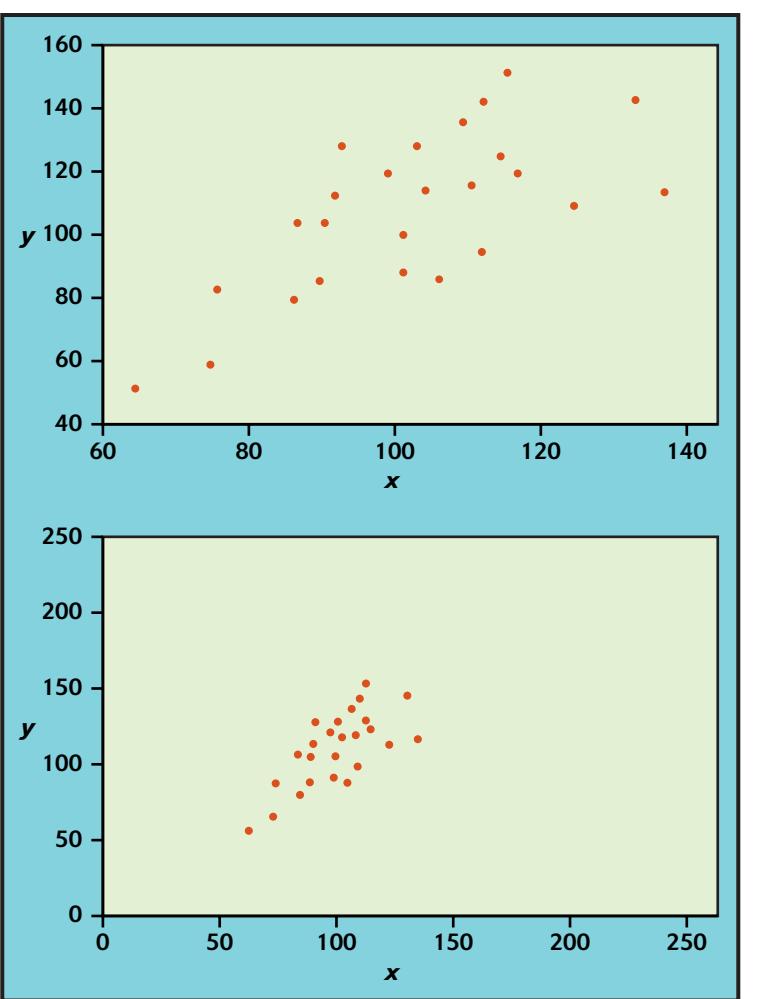


FIGURE 2.9 Two scatterplots of the same data. The straight-line pattern in the lower plot appears stronger because of the surrounding open space.

The correlation r

CORRELATION

The **correlation** measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as r .

Suppose that we have data on variables x and y for n individuals. The values for the first individual are x_1 and y_1 , the values for the second individual are x_2 and y_2 , and so on. The means and standard deviations of the two variables are \bar{x} and s_x for the x -values, and \bar{y} and s_y for the y -values. The correlation r between x and y is

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

As always, the summation sign \sum means “add these terms for all the individuals.” The formula for the correlation r is a bit complex. It helps us to see what correlation is, but in practice you should use software or a calculator that finds r from keyed-in values of two variables x and y . Exercises 2.19 and 2.20 ask you to calculate correlations step-by-step from the definition to solidify the meaning.

The formula for r begins by standardizing the observations. Suppose, for example, that x is height in centimeters and y is weight in kilograms and that we have height and weight measurements for n people. Then \bar{x} and s_x are the mean and standard deviation of the n heights, both in centimeters. The value

$$\frac{x_i - \bar{x}}{s_x}$$

is the standardized height of the i th person, familiar from Chapter 1. The standardized height says how many standard deviations above or below the mean a person’s height lies. Standardized values have no units—in this example, they are no longer measured in centimeters. Standardize the weights also. The correlation r is an average of the products of the standardized height and the standardized weight for the n people.

- 2.19 Four-wheel drive minivans.** The city and highway miles per gallon for all three 2001 model four-wheel drive minivans are

City MPG	18.6	19.2	18.8
Highway MPG	28.7	29.3	29.2

- (a) Make a scatterplot with appropriately labeled axes.
 - (b) Find the correlation r step-by-step. First, find the mean and standard deviation of the city MPG and of the highway MPG. (Use a calculator.) Then find the three standardized values for each variable and use the formula for r .
 - (c) Comment on the direction and strength of the relationship.
- 2.20 The price of a gigabyte.** Here are data on the size in gigabytes (GB) and the price in dollars of several external hard-drive models from one manufacturer:

Size (GB)	8	6	18	30	20	10	3
Price (\$)	310	290	500	800	470	330	150

- (a) Make a scatterplot. Be sure to label the axes appropriately. Comment on the form, direction, and strength of the relationship.
- (b) Find the correlation r step-by-step. First, find the mean and standard deviation of the sizes and of the prices (use your calculator). Then find the seven standardized values for these variables and use the formula for r .
- (c) Enter these data into your calculator and use the calculator’s correlation function to find r . Check that you get the same result as in (b).

APPLY YOUR KNOWLEDGE

Facts about correlation

The formula for correlation helps us see that r is positive when there is a positive association between the variables. Height and weight, for example, have a positive association. People who are above average in height tend also to be above average in weight. Both the standardized height and the standardized weight are positive. People who are below average in height tend also to have below-average weight. Then both standardized height and standardized weight are negative. In both cases, the products in the formula for r are mostly positive and so r is positive. In the same way, we can see that r is negative when the association between x and y is negative. More detailed study of the formula gives more detailed properties of r . Here is what you need to know to interpret correlation.

1. Correlation makes no distinction between explanatory and response variables. It makes no difference which variable you call x and which you call y in calculating the correlation.
2. Correlation requires that both variables be quantitative, so that it makes sense to do the arithmetic indicated by the formula for r . We cannot calculate a correlation between the incomes of a group of people and what city they live in, because city is a categorical variable.
3. Because r uses the standardized values of the observations, r does not change when we change the units of measurement of x , y , or both. Measuring height in inches rather than centimeters and weight in pounds rather than kilograms does not change the correlation between height and weight. The correlation r itself has no unit of measurement; it is just a number.
4. Positive r indicates positive association between the variables, and negative r indicates negative association.
5. The correlation r is always a number between -1 and 1 . Values of r near 0 indicate a very weak linear relationship. The strength of the linear relationship increases as r moves away from 0 toward either -1 or 1 . Values of r close to -1 or 1 indicate that the points in a scatterplot lie close to a straight line. The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect linear relationship, when the points lie exactly along a straight line.
6. Correlation measures the strength of only a linear relationship between two variables. Correlation does not describe curved relationships between variables, no matter how strong they are.
7. Like the mean and standard deviation, the correlation is not resistant: r is strongly affected by a few outlying observations. The correlation for Figure 2.8 (page 98) is $r = 0.6287$ when all 51 observations are included but rises to $r = 0.7789$ when we omit Alaska and the District of Columbia. Use r with caution when outliers appear in the scatterplot.

The scatterplots in Figure 2.10 illustrate how values of r closer to 1 or -1 correspond to stronger linear relationships. To make the meaning of r clearer, the standard deviations of both variables in these plots are equal

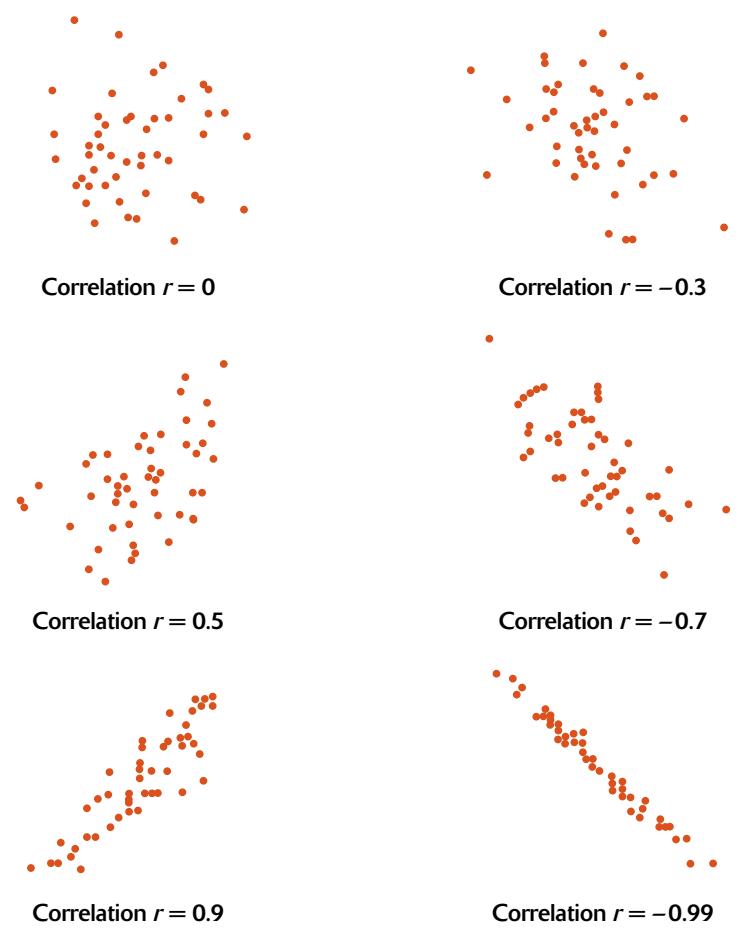


FIGURE 2.10 How correlation measures the strength of a linear relationship. Patterns closer to a straight line have correlations closer to 1 or -1 .

and the horizontal and vertical scales are the same. In general, it is not so easy to guess the value of r from the appearance of a scatterplot. Remember that changing the plotting scales in a scatterplot may mislead our eyes, but it does not change the correlation.

The real data we have examined also illustrate how correlation measures the strength and direction of linear relationships. Figure 2.3 (page 92) shows a very strong positive linear relationship between degree-days and natural-gas consumption. The correlation is $r = 0.9953$. Check this on your calculator or software from the data in Table 2.2. Figure 2.7 (page 98) shows a positive but curved relationship. The correlation, which measures only the strength of the *linear* relationship for this plot, is $r = 0.7177$.

Do remember that **correlation is not a complete description of two-variable data**, even when the relationship between the variables is linear. You should give the means and standard deviations of both x and y along with the correlation. (Because the formula for correlation uses the means and standard deviations, these measures are the proper choice to accompany a

correlation.) Conclusions based on correlations alone may require rethinking in the light of a more complete description of the data.

EXAMPLE 2.7**Forecasting earnings**

Stock analysts regularly forecast the earnings per share (EPS) of companies they follow. EPS is calculated by dividing a company's net income for a given time period by the number of common stock shares outstanding. We have two analysts' EPS forecasts for a computer manufacturer for the next 6 quarters. How well do the two forecasts agree? The correlation between them is $r = 0.9$, but the mean of the first analyst's forecasts is 3 dollars per share lower than the second analyst's mean.

These facts do not contradict each other. They are simply different kinds of information. The means show that the first analyst predicts lower EPS than the second. But because the first analyst's EPS predictions are about 3 dollars per share lower than the second analyst's *for every quarter*, the correlation remains high. Adding or subtracting the same number to all values of either x or y does not change the correlation. The two analysts agree on which quarters will see higher EPS values. The high r shows this agreement, despite the fact that the actual predicted values differ by 3 dollars per share.

APPLY YOUR KNOWLEDGE

- 2.21 **Thinking about correlation.** Figure 2.6 (page 97) is a scatterplot of percent decline versus duration in months for 15 bear markets.
- Is the correlation r for these data near -1 , clearly negative but not near -1 , near 0 , clearly positive but not near 1 , or near 1 ? Explain your answer.
 - Figure 2.2 (page 90) shows the highway and city gas mileage for 2001 model two-seater cars. Is the correlation here closer to 1 than that for Figure 2.6 or closer to zero? Explain your answer.
- 2.22 **Brand names and generic products.**
- If a store always prices its generic "store brand" products at 90% of the brand name products' prices, what would be the correlation between the prices of the brand name products and the store brand products? (Hint: Draw a scatterplot for several prices.)
 - If the store always prices its generic products \$1 less than the corresponding brand name products, then what would be the correlation between the prices of the brand name products and the store brand products?
- 2.23 **Strong association but no correlation.** The gas mileage of an automobile first increases and then decreases as the speed increases. Suppose that this relationship is very regular, as shown by the following data on speed (miles per hour) and mileage (miles per gallon):

Speed	20	30	40	50	60
MPG	24	28	30	28	24

Make a scatterplot of mileage versus speed. Show that the correlation between speed and mileage is $r = 0$. Explain why the correlation is 0 even though there is a strong relationship between speed and mileage.

SECTION 2.2 SUMMARY

- The correlation r measures the strength and direction of the linear association between two quantitative variables x and y . Although you can calculate a correlation for any scatterplot, r measures only straight-line relationships.
- Correlation indicates the direction of a linear relationship by its sign: $r > 0$ for a positive association and $r < 0$ for a negative association.
- Correlation always satisfies $-1 \leq r \leq 1$ and indicates the strength of a relationship by how close it is to -1 or 1 . Perfect correlation, $r = \pm 1$, occurs only when the points on a scatterplot lie exactly on a straight line.
- Correlation ignores the distinction between explanatory and response variables. The value of r is not affected by changes in the unit of measurement of either variable. Correlation is not resistant, so outliers can greatly change the value of r .

SECTION 2.2 EXERCISES



- 2.24 Match the correlation. The *Correlation and Regression* applet at www.whfreeman.com/pbs allows you to create a scatterplot by clicking and dragging with the mouse. The applet calculates and displays the correlation as you change the plot. You will use this applet to make scatterplots with 10 points that have correlation close to 0.7. The lesson is that many patterns can have the same correlation. Always plot your data before you trust a correlation.
- Stop after adding the first two points. What is the value of the correlation? Why does it have this value?
 - Make a lower-left to upper-right pattern of 10 points with correlation about $r = 0.7$. (You can drag points up or down to adjust r after you have 10 points.) Make a rough sketch of your scatterplot.
 - Make another scatterplot with 9 points in a vertical stack at the left of the plot. Add one point far to the right and move it until the correlation is close to 0.7. Make a rough sketch of your scatterplot.
 - Make yet another scatterplot with 10 points in a curved pattern that starts at the lower left, rises to the right, then falls again at the far right. Adjust the points up or down until you have a quite smooth curve with correlation close to 0.7. Make a rough sketch of this scatterplot also.
- 2.25 Mutual-fund performance. Many mutual funds compare their performance with that of a benchmark, an index of the returns on all securities of the kind that the fund buys. The Vanguard International Growth Fund, for example, takes as its benchmark the Morgan Stanley Europe, Australasia, Far East (EAFE) index of overseas stock market performance. Here are the percent returns for the fund and for the EAFE from 1982 (the first full year of the fund's existence) to 2000:¹³

Year	Fund	EAFFE	Year	Fund	EAFFE
1982	5.27	-0.86	1992	-5.79	-11.85
1983	43.08	24.61	1993	44.74	32.94
1984	-1.02	7.86	1994	0.76	8.06
1985	56.94	56.72	1995	14.89	11.55
1986	56.71	69.94	1996	14.65	6.36
1987	12.48	24.93	1997	4.12	2.06
1988	11.61	28.59	1998	16.93	20.33
1989	24.76	10.80	1999	26.34	27.30
1990	-12.05	-23.20	2000	-8.60	-13.96
1991	4.74	12.50			

Make a scatterplot suitable for predicting fund returns from EAFFE returns. Is there a clear straight-line pattern? How strong is this pattern? (Give a numerical measure.) Are there any extreme outliers from the straight-line pattern?

- 2.26 **How many calories?** A food industry group asked 3368 people to guess the number of calories in each of several common foods. Table 2.5 displays the averages of their guesses and the correct number of calories.¹⁴
- We think that how many calories a food actually has helps explain people's guesses of how many calories it has. With this in mind, make a scatterplot of these data.
 - Find the correlation r . Explain why your r is reasonable based on the scatterplot.
 - The guesses are all higher than the true calorie counts. Does this fact influence the correlation in any way? How would r change if every guess were 100 calories higher?
 - The guesses are much too high for spaghetti and the snack cake. Circle these points on your scatterplot. Calculate r for the other eight foods, leaving out these two points. Explain why r changed in the direction that it did.

TABLE 2.5 Guessed and true calories in 10 foods

Food	Guessed calories	Correct calories
8 oz. whole milk	196	159
5 oz. spaghetti with tomato sauce	394	163
5 oz. macaroni with cheese	350	269
One slice wheat bread	117	61
One slice white bread	136	76
2-oz. candy bar	364	260
Saltine cracker	74	12
Medium-size apple	107	80
Medium-size potato	160	88
Cream-filled snack cake	419	160

- 2.27 **Stretching a scatterplot.** Changing the units of measurement can greatly alter the appearance of a scatterplot. Consider the following data:

x	-4	-4	-3	3	4	4
y	0.5	-0.6	-0.5	0.5	0.5	-0.6

- (a) Draw x and y axes each extending from -6 to 6. Plot the data on these axes.
 - (b) Calculate the values of new variables $x^* = x/10$ and $y^* = 10y$, starting from the values of x and y . Plot y^* against x^* on the same axes using a different plotting symbol. The two plots are very different in appearance.
 - (c) Find the correlation between x and y . Then find the correlation between x^* and y^* . How are the two correlations related? Explain why this isn't surprising.
- 2.28 **Where the stores are.** Exercise 2.18 (page 102) looked at data on the number of Target and Wal-Mart stores in each state.
- (a) Use software to calculate the correlation of these data.
 - (b) Predict whether the correlation will increase or decrease if California's data are excluded, then recalculate the correlation without the data for California. Why did the correlation change in this way?
 - (c) Start again with all 50 observations and predict whether the correlation will increase or decrease if the data for Texas are excluded. Recalculate the correlation without the Texas data. Why did the correlation change in this way?
- 2.29 **CEO compensation and stock market performance.** An academic study concludes, “The evidence indicates that the correlation between the compensation of corporate CEOs and the performance of their company’s stock is close to zero.” A business magazine reports this as “A new study shows that companies that pay their CEOs highly tend to perform poorly in the stock market, and vice versa.” Explain why the magazine’s report is wrong. Write a statement in plain language (don’t use the word “correlation”) to explain the study’s conclusion.
- 2.30 **Investment diversification.** A mutual-fund company’s newsletter says, “A well-diversified portfolio includes assets with low correlations.” The newsletter includes a table of correlations between the returns on various classes of investments. For example, the correlation between municipal bonds and large-cap stocks is 0.50, and the correlation between municipal bonds and small-cap stocks is 0.21.¹⁵
- (a) Rachel invests heavily in municipal bonds. She wants to diversify by adding an investment whose returns do not closely follow the returns on her bonds. Should she choose large-cap stocks or small-cap stocks for this purpose? Explain your answer.
 - (b) If Rachel wants an investment that tends to increase when the return on her bonds drops, what kind of correlation should she look for?
- 2.31 **Outliers and correlation.** Both Figures 2.2 and 2.8 contain outliers. Removing the outliers *increases* the correlation r in Figure 2.8, but in Figure 2.2 removing the outlier *decreases* r . Explain why this is true.

- 2.32 **Driving speed and fuel consumption.** The data in Exercise 2.23 were made up to create an example of a strong curved relationship for which, nonetheless, $r = 0$. Exercise 2.7 (page 93) gives actual data on gas used versus speed for a small car. Make a scatterplot if you did not do so in Exercise 2.7. Calculate the correlation, and explain why r is close to 0 despite a strong relationship between speed and gas used.
- 2.33 **Sloppy writing about correlation.** Each of the following statements contains a blunder. Explain in each case what is wrong.
- "There is a high correlation between the gender of American workers and their income."
 - "We found a high correlation ($r = 1.09$) between students' ratings of faculty teaching and ratings made by other faculty members."
 - "The correlation between planting rate and yield of corn was found to be $r = 0.23$ bushel."
- 2.34 **Mutual funds and correlation.** Financial experts use statistical measures to describe the performance of investments, such as mutual funds. In the past, fund companies feared that investors would not understand statistical descriptions, but investor demand for better information is moving standard deviations and correlations into published reports.
- The T. Rowe Price mutual-fund group reports the standard deviation of yearly percent returns for its funds. Recently, Equity Income Fund had standard deviation 9.94%, and Science & Technology Fund had standard deviation 23.77%. Explain to someone who knows no statistics how these standard deviations help investors compare the two funds.
 - Some mutual funds act much like the stock market as a whole, as measured by a market index such as the Standard & Poor's 500-stock index (S&P 500). Others are very different from the overall market. We can use correlation to describe the association. Monthly returns from Fidelity Magellan Fund have correlation $r = 0.85$ with the S&P 500. Fidelity Small Cap Stock Fund has correlation $r = 0.55$ with the S&P 500. Explain to someone who knows no statistics how these correlations help investors compare the two funds.

2.3 Least-Squares Regression

Correlation measures the direction and strength of the straight-line (linear) relationship between two quantitative variables. If a scatterplot shows a linear relationship, we would like to summarize this overall pattern by drawing a line on the scatterplot. A *regression line* summarizes the relationship between two variables, but only in a specific setting: one of the variables helps explain or predict the other. That is, regression describes a relationship between an explanatory variable and a response variable.

REGRESSION LINE

A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x .

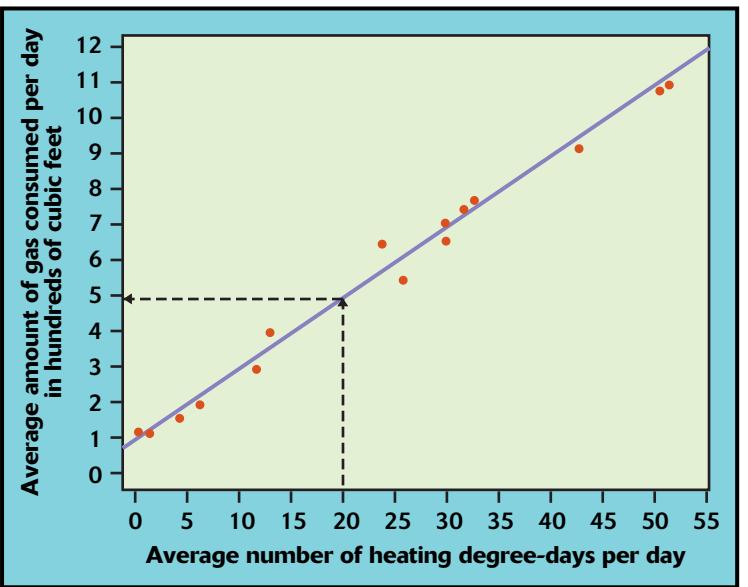


FIGURE 2.11 The Sanchez household gas consumption data, with a regression line for predicting gas consumption from degree-days. The dashed lines illustrate how to use the regression line to predict gas consumption for a month averaging 20 degree-days per day.

EXAMPLE 2.8

Predicting natural-gas consumption

Figure 2.11 is another scatterplot of the natural-gas consumption data from Table 2.2. There is a strong linear relationship between the average outside temperature (measured by heating degree-days) in a month and the average amount of natural gas that the Sanchez household uses per day during the month. The correlation is $r = 0.9953$, close to the $r = 1$ of points that lie exactly on a line. The regression line drawn through the points in Figure 2.11 describes these data very well.

The Sanchez household wants to use this line to predict their natural gas consumption. “If a month averages 20 degree-days per day (that’s 45° F), how much gas will we use?” To predict gas consumption at 20 degree-days, first locate 20 on the x axis. Then go “up and over” as in the figure to find the gas consumption y that corresponds to $x = 20$. We predict that the Sanchez household will use about 4.9 hundreds of cubic feet each day in such a month.

prediction



The least-squares regression line

Different people might draw different lines by eye on a scatterplot. This is especially true when the points are more widely scattered than those in Figure 2.11. We need a way to draw a regression line that doesn’t depend on our guess as to where the line should go. We will use the line to predict y from x , so the prediction errors we make are errors in y , the vertical direction in the scatterplot. If we predict 4.9 hundreds of cubic feet for a

month with 20 degree-days and the actual use turns out to be 5.1 hundreds of cubic feet, our prediction error is

$$\begin{aligned}\text{error} &= \text{observed } y - \text{predicted } \hat{y} \\ &= 5.1 - 4.9 = 0.2\end{aligned}$$

No line will pass exactly through all the points in the scatterplot. We want the *vertical* distances of the points from the line to be as small as possible. Figure 2.12 illustrates the idea. This plot shows three of the points from Figure 2.11, along with the line, on an expanded scale. The line passes above two of the points and below one of them. The vertical distances of the data points from the line appear as vertical line segments. There are many ways to make the collection of vertical distances “as small as possible.” The most common is the *least-squares* method.

LEAST-SQUARES REGRESSION LINE

The **least-squares regression line** of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

One reason for the popularity of the least-squares regression line is that the problem of finding the line has a simple solution. We can give the recipe for the least-squares line in terms of the means and standard deviations of the two variables and their correlation.

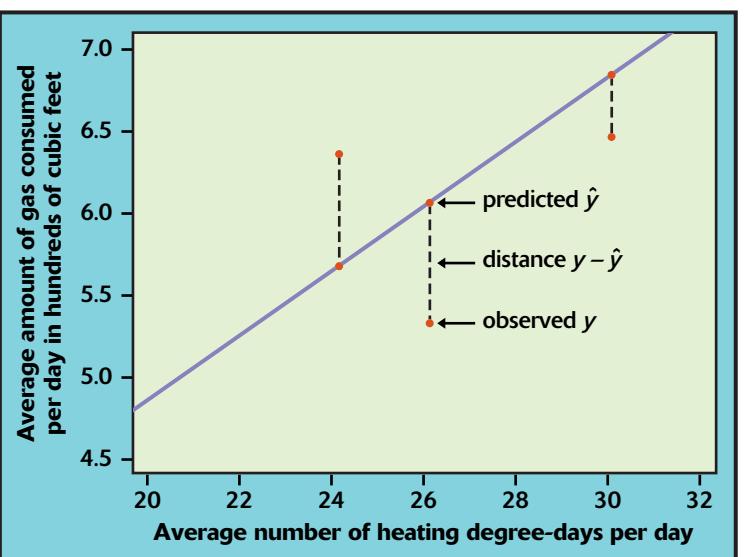


FIGURE 2.12 The least-squares idea. For each observation, find the vertical distance of each point on the scatterplot from a regression line. The least-squares regression line makes the sum of the squares of these distances as small as possible.

EQUATION OF THE LEAST-SQUARES REGRESSION LINE

We have data on an explanatory variable x and a response variable y for n individuals. From the data, calculate the means \bar{x} and \bar{y} and the standard deviations s_x and s_y of the two variables, and their correlation r . The least-squares regression line is the line

$$\hat{y} = a + bx$$

with slope

$$b = r \frac{s_y}{s_x}$$

and intercept

$$a = \bar{y} - b\bar{x}$$

We write \hat{y} (read “y hat”) in the equation of the regression line to emphasize that the line gives a *predicted* response \hat{y} for any x . Because of the scatter of points about the line, the predicted response will usually not be exactly the same as the actually *observed* response y . In practice, you don’t need to calculate the means, standard deviations, and correlation first. Statistical software or your calculator will give the slope b and intercept a of the least-squares line from keyed-in values of the variables x and y . You can then concentrate on understanding and using the regression line.

EXAMPLE 2.9

Using a regression line

The line in Figure 2.11 is in fact the least-squares regression line of gas consumption on degree-days. Enter the data from Table 2.2 into your calculator or software and check that the equation of this line is

$$\hat{y} = 1.0892 + 0.1890x$$

slope

The **slope** of a regression line is almost always important for interpreting the data. The slope is the rate of change, the amount of change in \hat{y} when x increases by 1. The slope $b = 0.1890$ in this example says that each additional degree-day predicts consumption of about 0.19 more hundreds of cubic feet of natural gas per day.

intercept

The **intercept** of the regression line is the value of \hat{y} when $x = 0$. Although we need the value of the intercept to draw the line, it is statistically meaningful only when x can actually take values close to zero. In our example, $x = 0$ occurs when the average outdoor temperature is at least 65° F. We predict that the Sanchez household will use an average of $a = 1.0892$ hundreds of cubic feet of gas per day when there are no degree-days. They use this gas for cooking and heating water, which continue in warm weather.

prediction

The equation of the regression line makes **prediction** easy. Just substitute an x -value into the equation. To predict gas consumption at 20 degree-days, substitute $x = 20$:

$$\begin{aligned}\hat{y} &= 1.0892 + (0.1890)(20) \\ &= 1.0892 + 3.78 = 4.869\end{aligned}$$

plotting a line

To plot the line on the scatterplot, use the equation to find \hat{y} for two values of x , one near each end of the range of x in the data. Plot each \hat{y} above its x and draw the line through the two points.

Figure 2.13 displays the regression output for the gas consumption data from a graphing calculator, a statistical software package, and a spreadsheet. Each output records the slope and intercept of the least-squares line, calculated to more decimal places than we need. The software also provides information that we do not yet need—part of the art of using such tools is to ignore the extra information that is almost always present. You must also inspect the output with care. For example, the graphing calculator presents the least-squares line in the form $y = ax + b$, reversing our use of the symbols a and b .

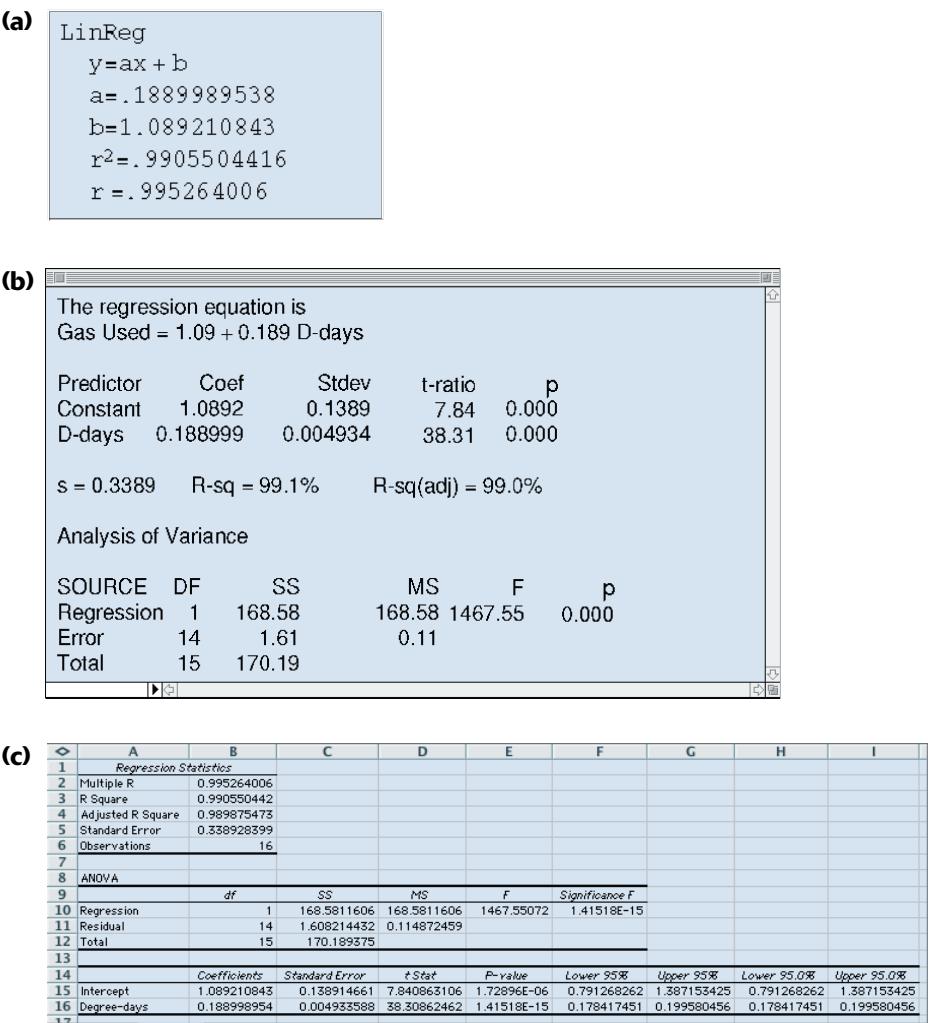


FIGURE 2.13 Least-squares regression output for the gas consumption data from a graphing calculator, a statistical software package, and a spreadsheet. (a) The TI-83 calculator. (b) Minitab. (c) Microsoft Excel.

APPLY YOUR KNOWLEDGE

- 2.35 Example 2.9 gives the equation of the regression line of gas consumption y on degree-days x for the data in Table 2.2 as

$$\hat{y} = 1.0892 + 0.1890x$$

Enter the data from Table 2.2 into your calculator or software.

- (a) Use the regression function to find the equation of the least-squares regression line.
- (b) Find the mean and standard deviation of both x and y and their correlation r . Find the slope b and intercept a of the regression line from these, using the facts in the box Equation of the Least-Squares Regression Line. Verify that in both part (a) and part (b) you get the equation in Example 2.9. (Results may differ slightly because of rounding.)

- 2.36 **Architectural firms.** Table 1.3 (page 14) contains data describing firms engaged in commercial architecture in the Indianapolis, Indiana, area. The regression line for predicting total billings from number of full-time staff members employed is

$$\text{billings} = 0.8334 + (0.0902 \times \text{employed})$$

- (a) Make a scatterplot and draw this regression line on the plot. Using the regression equation, predict the total billings of an architectural firm in Indianapolis with 111 full-time staff members.
- (b) Compare the observed total billings for the firm with 111 full-time staff members with your prediction from part (a) by calculating the prediction error. How accurate was your prediction?

Facts about least-squares regression

Regression is one of the most common statistical settings, and least squares is the most common method for fitting a regression line to data. Here are some facts about least-squares regression lines.

Fact 1. The distinction between explanatory and response variables is essential in regression. Least-squares regression looks at the distances of the data points from the line only in the y direction. If we reverse the roles of the two variables, we get a different least-squares regression line.

EXAMPLE 2.10

Gross sales and item count

Figure 2.14 is a scatterplot of the data in columns B and C of Table 2.1 (page 88). They are daily gross sales and number of items sold per day for Duck Worth Wearing, the children's consignment shop described in Case 2.1. There is a positive linear relationship, with $r = 0.9540$ indicating a strong linear relationship.

The two lines on the plot are the two least-squares regression lines. The regression line of gross sales on item count is solid. The regression line of item count on gross sales is dashed. *Regression of gross sales on item count and regression of item count on gross sales give different lines.* In the regression setting, you must decide which variable is explanatory.

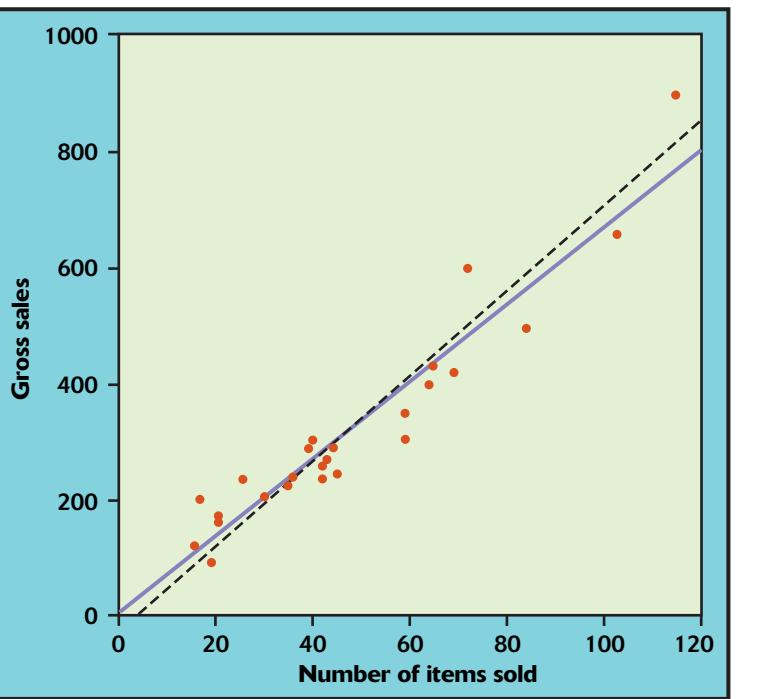


FIGURE 2.14 Scatterplot of the gross sales at Duck Worth Wearing for each day in April 2000 against the number of items sold for the same day, from Table 2.1. The two lines are the two least-squares regression lines: of gross sales on item count (solid) and of item count on gross sales (dashed).

Fact 2. There is a close connection between correlation and the slope of the least-squares line. The slope is

$$b = r \frac{s_y}{s_x}$$

This equation says that along the regression line, a change of one standard deviation in x corresponds to a change of r standard deviations in y . When the variables are perfectly correlated ($r = 1$ or $r = -1$), the change in the predicted response \hat{y} is the same (in standard deviation units) as the change in x . Otherwise, because $-1 \leq r \leq 1$, the change in \hat{y} is less than the change in x . As the correlation grows less strong, the prediction \hat{y} moves less in response to changes in x .

Fact 3. The least-squares regression line always passes through the point (\bar{x}, \bar{y}) on the graph of y against x . So the least-squares regression line of y on x is the line with slope $r s_y / s_x$ that passes through the point (\bar{x}, \bar{y}) . We can describe regression entirely in terms of the basic descriptive measures \bar{x} , s_x , \bar{y} , s_y , and r .

Fact 4. The correlation r describes the strength of a straight-line relationship. In the regression setting, this description takes a specific form: the square of the correlation r^2 is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x .

The idea is that when there is a linear relationship, some of the variation in y is accounted for by the fact that as x changes it pulls y along with it. Look

again at Figure 2.11 on page 112. There is a lot of variation in the observed y 's, the gas consumption data. They range from a low of about 1 to a high of 11. The scatterplot shows that most of this variation in y is accounted for by the fact that outdoor temperature (measured by degree-days x) was changing and pulled gas consumption along with it. There is only a little remaining variation in y , which appears in the scatter of points about the line. The points in Figure 2.14, on the other hand, are more scattered. Linear dependence on item count does explain much of the observed variation in gross sales. You would guess a higher value for the gross sales y knowing that $x = 100$ than you would if you were told that $x = 20$. But there is still considerable variation in y even when x is held relatively constant—look at the variability in the gross sales for points in Figure 2.14 near $x = 40$.

This idea can be expressed in algebra, though we won't do it. It is possible to separate the total variation in the observed values of y into two parts. One part is the variation we expect as x moves and \hat{y} moves with it along the regression line. The other measures the variation of the data points about the line. The squared correlation r^2 is the first of these as a fraction of the whole:

$$r^2 = \frac{\text{variation in } \hat{y} \text{ as } x \text{ pulls it along the line}}{\text{total variation in observed values of } y}$$

EXAMPLE 2.11

Using r^2

The correlation between degree-days and gas use (Figure 2.11) is $r = 0.9953$. So $r^2 = 0.99$ and the straight-line relationship explains 99% of the month-to-month variation in gas use. In Figure 2.14, $r = 0.9540$ and $r^2 = 0.91$. The linear relationship between gross sales and item count explains 91% of the variation in either variable. There are two regression lines, but just one correlation, and r^2 helps interpret both regressions.

The relationship between life expectancy and gross domestic product (GDP) (Figure 2.7, page 98) is positive but curved and has correlation $r = 0.7177$. Here, $r^2 = 0.515$, so that regressing life expectancy on GDP explains only about half of the variation among nations in life expectancy.

When you report a regression, give r^2 as a measure of how successful the regression was in explaining the response. All the software outputs in Figure 2.13 include r^2 , either in decimal form or as a percent. When you see a correlation, square it to get a better feel for the strength of the association. Perfect correlation ($r = -1$ or $r = 1$) means the points lie exactly on a line. Then $r^2 = 1$ and all of the variation in one variable is accounted for by the linear relationship with the other variable. If $r = -0.7$ or $r = 0.7$, $r^2 = 0.49$ and about half the variation is accounted for by the linear relationship. In the r^2 scale, correlation ± 0.7 is about halfway between 0 and ± 1 .

These facts are special properties of least-squares regression. They are not true for other methods of fitting a line to data. Another reason that least squares is the most common method for fitting a regression line to data is that it has many convenient special properties.

APPLY YOUR KNOWLEDGE

- 2.37 The “January effect.” Some people think that the behavior of the stock market in January predicts its behavior for the rest of the year. Take the explanatory variable x to be the percent change in a stock market index in January and the response variable y to be the change in the index for the entire year. We expect a positive correlation between x and y because the change during January contributes to the full year’s change. Calculation from data for the years 1960 to 1997 gives

$$\bar{x} = 1.75\% \quad s_x = 5.36\% \quad r = 0.596$$
$$\bar{y} = 9.07\% \quad s_y = 15.35\%$$

- (a) What percent of the observed variation in yearly changes in the index is explained by a straight-line relationship with the change during January?
- (b) What is the equation of the least-squares line for predicting full-year change from January change?
- (c) The mean change in January is $\bar{x} = 1.75\%$. Use your regression line to predict the change in the index in a year in which the index rises 1.75% in January. Why could you have given this result (up to roundoff error) without doing the calculation?



- 2.38 Is regression useful? In Exercise 2.24 (page 108) you used the *Correlation and Regression* applet to create three scatterplots having correlation about $r = 0.7$ between the horizontal variable x and the vertical variable y . Create three similar scatterplots again, after clicking the “Show least-squares line” box to display the regression line. Correlation $r = 0.7$ is considered reasonably strong in many areas of work. Because there is a reasonably strong correlation, we might use a regression line to predict y from x . In which of your three scatterplots does it make sense to use a straight line for prediction?

Residuals

A regression line is a mathematical model for the overall pattern of a linear relationship between an explanatory variable and a response variable. Deviations from the overall pattern are also important. In the regression setting, we see deviations by looking at the scatter of the data points about the regression line. The vertical distances from the points to the least-squares regression line are as small as possible in the sense that they have the smallest possible sum of squares. Because they represent “leftover” variation in the response after fitting the regression line, these distances are called *residuals*.

RESIDUALS

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

$$\begin{aligned}\text{residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y}\end{aligned}$$

EXAMPLE 2.12**Predicting credit card sales**

Table 2.1 (page 88) contains data on daily sales by Duck Worth Wearing for April 2000. We are now interested in credit card sales. The final two columns in the spreadsheet display gross credit card sales and the number of items purchased with a credit card for each day. We will ignore the four days on which no customer used a credit card.

Figure 2.15 is a scatterplot of these data, with items purchased as the explanatory variable x and gross sales as the response variable y . The plot shows a positive association—days with more items purchased with a credit card tend to be the days with higher credit card sales. The overall pattern is moderately linear. The correlation describes both the direction and strength of the linear relationship. It is $r = 0.7718$.

The line on the plot is the least-squares regression line of credit card sales on item count. Its equation is

$$\hat{y} = 14.3908 + 6.1739x$$

For observation 11, the day with 9 items purchased with credit cards, we predict the credit card sales

$$\hat{y} = 14.3908 + (6.1739)(9) = 69.96$$

This day's actual credit card sales were \$49.50. The residual is

$$\begin{aligned}\text{residual} &= \text{observed } y - \text{predicted } y \\ &= 49.50 - 69.96 = -20.46\end{aligned}$$

The residual is negative because the data point lies below the regression line.

There is a residual for each data point. Finding the residuals with a calculator is a bit unpleasant, because you must first find the predicted response for every x . Statistical software gives you the residuals all at once. Here are the 21 residuals for the credit card data, from regression software:

```
residuals:
108.9135  59.0657   3.8917  -6.2604 -13.5647 -31.5128 -18.7604
-13.2386 -17.4779 -28.0432 -20.4561 -13.7386  -8.1736 -13.9126
-56.8257  -8.9126  67.8700 -19.7540 -19.1126 -15.3365  65.3394
```

Because the residuals show how far the data fall from our regression line, examining the residuals helps assess how well the line describes the data. Although residuals can be calculated from any model fitted to the data, the residuals from the least-squares line have a special property: the mean of the least-squares residuals is always zero.

Compare the scatterplot in Figure 2.15 with the *residual plot* for the same data in Figure 2.16. The horizontal line at zero in Figure 2.16 helps orient us. It corresponds to the regression line in Figure 2.15.

RESIDUAL PLOTS

A residual plot is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of a regression line.

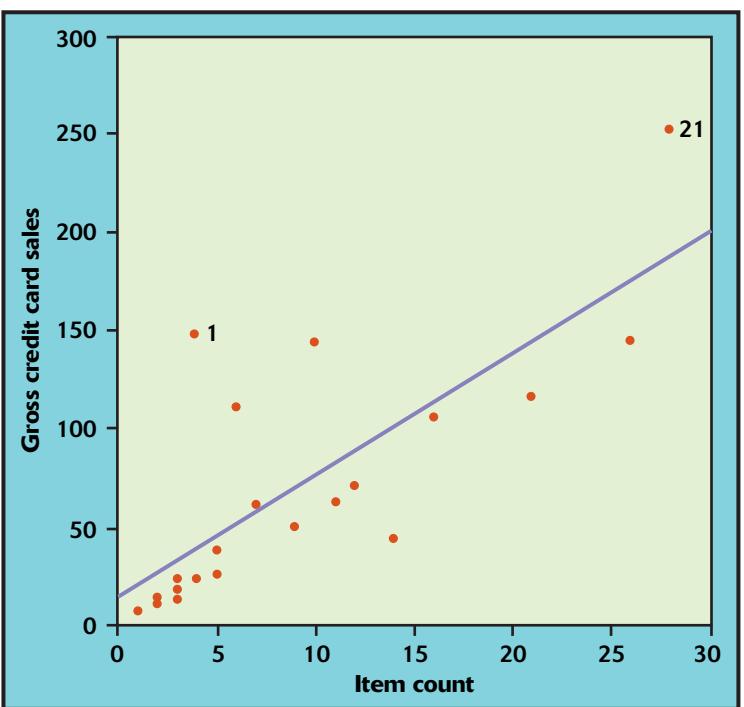


FIGURE 2.15 Scatterplot of gross credit card sales at Duck Worth Wearing versus number of items purchased with a credit card for 21 days, from Table 2.1. The line is the least-squares regression line for predicting sales from item count.

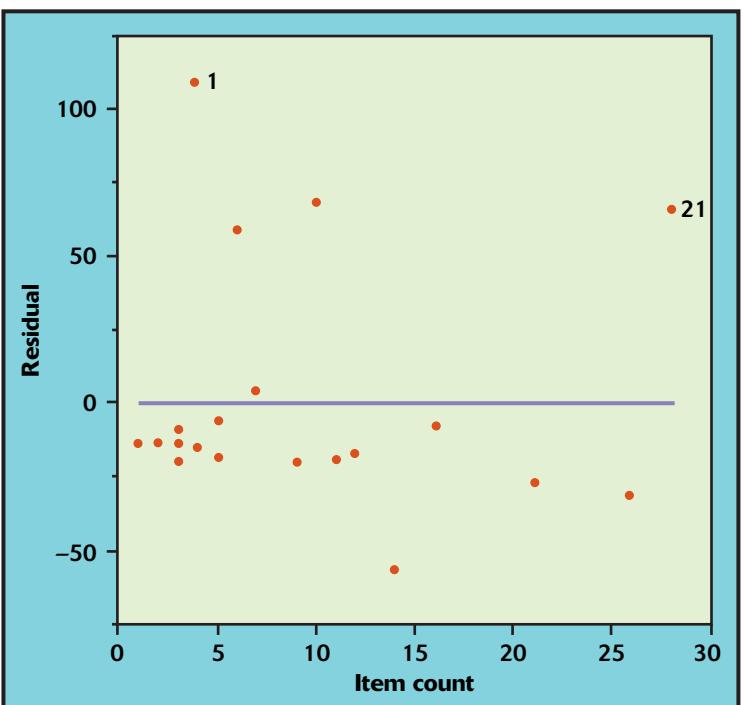


FIGURE 2.16 Residual plot for the regression of sales on item count. Observation 1 is an outlier. Observation 21 has a smaller residual but is more influential.

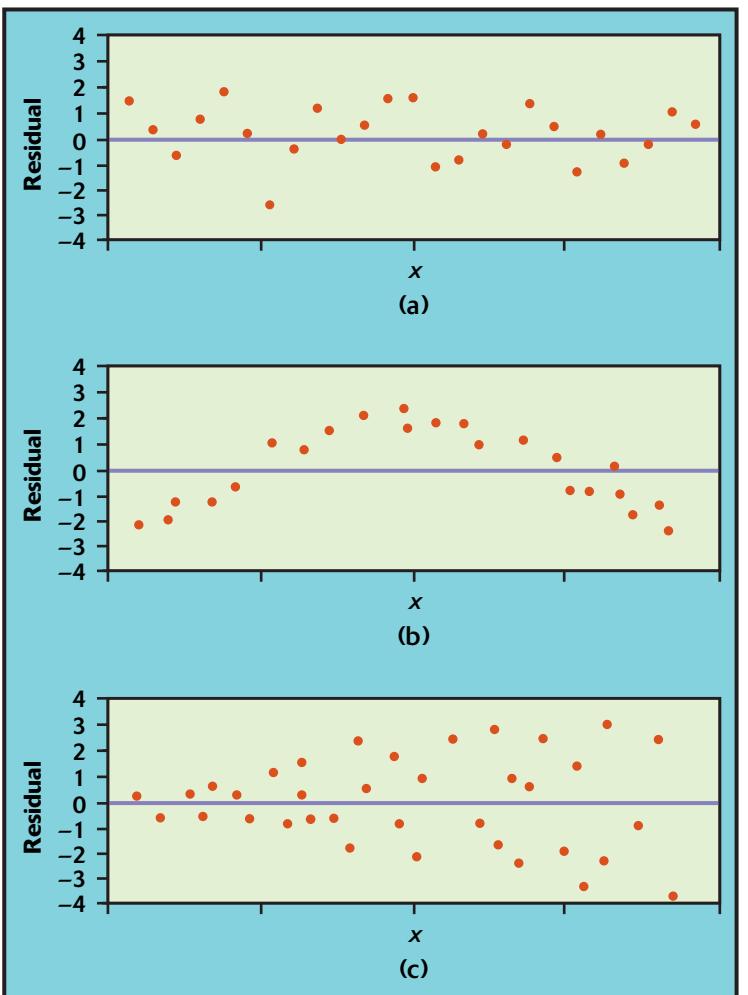


FIGURE 2.17 Idealized patterns in plots of least-squares residuals. Plot (a) indicates that the regression line fits the data well. The data in plot (b) have a curved pattern, so a straight line fits poorly. The response variable y in plot (c) has more spread for larger values of the explanatory variable x , so prediction will be less accurate when x is large.

If the regression line captures the overall relationship between x and y , the residuals should have no systematic pattern. The residual plot will look something like the pattern in Figure 2.17(a). That plot shows a scatter of the points about the fitted line, with no unusual individual observations or systematic change as x increases. Here are some things to look for when you examine the residuals, using either a scatterplot of the data or a residual plot:

- **A curved pattern** shows that the relationship is not linear. Figure 2.17(b) is a simplified example. A straight line is not a good summary for such data. The residuals for Figure 2.7 (page 98) would have this form.

- Increasing or decreasing spread about the line as x increases. Figure 2.17(c) is a simplified example. Prediction of y will be less accurate for larger x in that example.
- Individual points with large residuals, like observation 1 in Figures 2.15 and 2.16. Such points are outliers in the vertical (y) direction because they lie far from the line that describes the overall pattern.
- Individual points that are extreme in the x direction, like observation 21 in Figures 2.15 and 2.16. Such points may not have large residuals, but they can be very important. We address such points next.

Influential observations

Observations 1 and 21 are both unusual in the credit card sales example. They are unusual in different ways. Observation 1 lies far from the regression line. This day's credit card sales figure is so high that we should check for a mistake in recording it. In fact, the amount is correct. Observation 21 is closer to the line but far out in the x direction. This day had more items purchased with credit cards than the other days in the sample. *Because of its extreme position on the item count scale, this point has a strong influence on the position of the regression line.* Figure 2.18 adds a second regression line, calculated after leaving out observation 21. You can see that this one point moves the line quite a bit. We call such points *influential*.

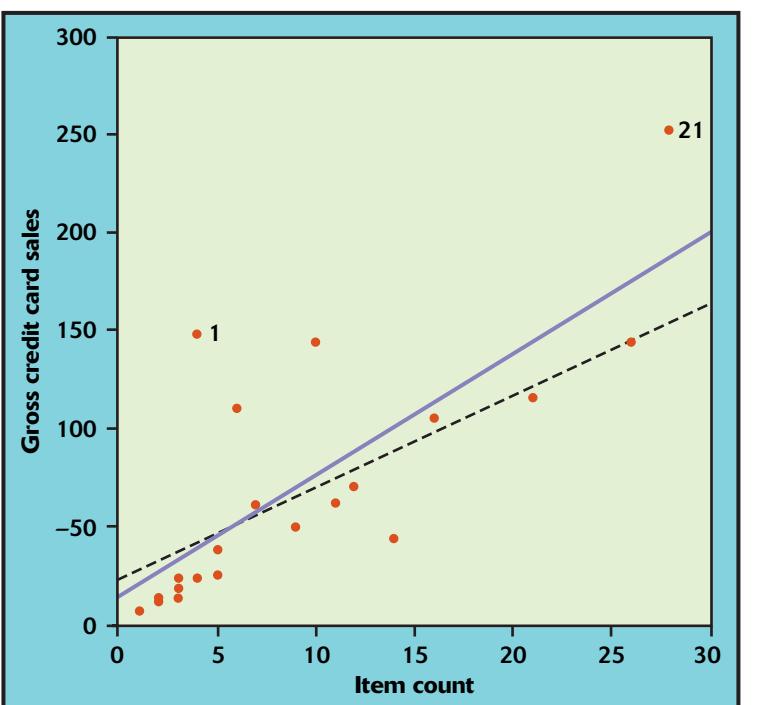


FIGURE 2.18 Two least-squares regression lines of sales on item count. The solid line is calculated from all the data. The dashed line was calculated leaving out observation 21. Observation 21 is an influential observation because leaving out this point moves the regression line quite a bit.

OUTLIERS AND INFLUENTIAL OBSERVATIONS IN REGRESSION

An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the y direction of a scatterplot have large regression residuals, but other outliers need not have large residuals.

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are extreme in the x direction of a scatterplot are often influential for the least-squares regression line.

Observations 1 and 21 are both quite far from the original regression line in Figure 2.18. Observation 21 influences the least-squares line because it is far out in the x direction. Observation 1 is an outlier in the y direction. It has less influence on the regression line because the many other points with similar values of x anchor the line well below the outlying point. Influential observations may have small residuals, because they pull the regression line toward themselves. That is, you can't rely on residuals to point out influential observations. Influential observations can change the interpretation of data.

EXAMPLE 2.13

An influential observation



The strong influence of observation 21 makes the original regression of credit card sales on item count a bit misleading. The original data have $r^2 = 0.5957$. That is, the number of items purchased with credit cards explains about 60% of the variation in daily credit card sales. This relationship is strong enough to be interesting to the store's owners. If we leave out observation 21, r^2 drops to only 43%. The unexamined regression suggests that items sold predict sales quite well, but much of this is due to a single day's results.

What should we do with observation 21? If this unusual day resulted from a special promotion, we might remove it when predicting results for ordinary business days. Otherwise, the influential observation reminds us of the hazard of prediction using small sets of data. More data may include other days with high item counts and so reduce the influence of this one observation.



The best way to grasp the important idea of influence is to use an interactive animation that allows you to move points on a scatterplot and observe how correlation and regression respond. The *Correlation and Regression* applet on the companion Web site for this book, www.whfreeman.com/pbs, allows you to do this. Exercises 2.51 and 2.52 guide the use of this applet.

- 2.39 Driving speed and fuel consumption.** Exercise 2.7 (page 93) gives data on the fuel consumption y of a car at various speeds x . Fuel consumption is measured in liters of gasoline per 100 kilometers driven, and speed is measured in kilometers per hour. Software tells us that the equation of the least-squares regression line is

$$\hat{y} = 11.058 - 0.01466x$$

APPLY YOUR KNOWLEDGE

The residuals, in the same order as the observations, are

10.09	2.24	-0.62	-2.47	-3.33	-4.28	-3.73	-2.94
-2.17	-1.32	-0.42	0.57	1.64	2.76	3.97	

- (a) Make a scatterplot of the observations and draw the regression line on your plot.
 - (b) Would you use the regression line to predict y from x ? Explain your answer.
 - (c) Check that the residuals have sum zero (up to roundoff error).
 - (d) Make a plot of the residuals against the values of x . Draw a horizontal line at height zero on your plot. Notice that the residuals show the same pattern about this line as the data points show about the regression line in the scatterplot in (a).
- 2.40 **Doctors and poverty.** We might expect states with more poverty to have fewer doctors. Table 1.7 (page 29) gives data on the percent of each state's residents living below the poverty line and on the number of M.D.'s per 100,000 residents in each state.
- (a) Make a scatterplot and calculate a regression line suitable for predicting M.D.'s per 100,000 residents from poverty rate. Draw the line on your plot. Surprise: the slope is positive, so poverty and M.D.'s go up together.
 - (b) The District of Columbia is an outlier, with both very many M.D.'s and a high poverty rate. (D.C. is a city rather than a state.) Circle the point for D.C. on your plot and explain why this point may strongly influence the least-squares line.
 - (c) Calculate the regression line for the 50 states, omitting D.C. Add the new line to your scatterplot. Was this point highly influential? Does the number of doctors now go down with increasing poverty, as we initially expected?
- 2.41 **Influential or not?** We have seen that observation 21 in the credit card data in Table 2.1 is an influential observation. Now we will examine the effect of observation 1, which is also an outlier in Figure 2.15.
- (a) Find the least-squares regression line of credit card sales on item count, leaving out observation 1. Example 2.12 gives the regression line from all the data. Plot both lines on the same graph. (You do not have to make a scatterplot of all the points; just plot the two lines.) Would you call observation 1 very influential? Why?
 - (b) How does removing observation 1 change the r^2 for this regression? Explain why r^2 changes in this direction when you drop observation 1.



BEYOND THE BASICS: SCATTERPLOT SMOOTHERS

A scatterplot provides a complete picture of the relationship between two quantitative variables. A complete picture is often too detailed for easy interpretation, so we try to describe the plot in terms of an overall pattern and deviations from that pattern. If the pattern is roughly linear, we can fit a regression line to describe it. Modern software offers a method more flexible

smoothing

than fitting a specific model such as a straight line. Scatterplot smoothing uses algorithms that extract an overall pattern by looking at the y -values for points in the neighborhood of each x -value.

EXAMPLE 2.14**Smoothing a nonlinear pattern**

Figure 2.7 displays the relationship between gross domestic product (GDP) per person and life expectancy for all the world's nations for which data are available. There is a strong pattern with positive association, but the pattern is not linear. Figure 2.19 shows the result of applying a scatterplot smoother to these data. The smoother accurately catches the rapid rise in life expectancy as GDP increases up to roughly \$5000 per person, followed by a much more gradual rise in wealthier nations.

Scatterplot smoothing is not quite automatic—you can decide how smooth you want your result to be. Figure 2.19 shows the results of two choices. The dashed line is very smooth, while the solid line pays more attention to local detail. For example, the solid line is pulled down a bit around \$5000 GDP by the three low outliers.



Scatterplot smoothers help us see overall patterns, especially when these patterns are not simple. They do not, however, provide a handy equation for prediction and other purposes. Least-squares regression, whether fitting a straight line or more complex models, remains more widely used.

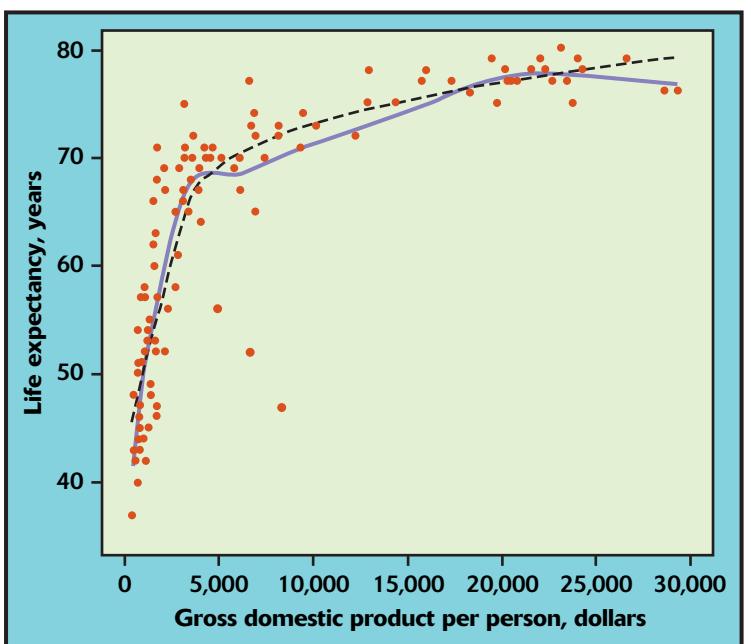


FIGURE 2.19 Two scatterplot smoothing algorithms applied to the data on national gross domestic product per person and life expectancy. Both portray the overall pattern of the plot.

SECTION 2.3 SUMMARY

- A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes.
- The most common method of fitting a line to a scatterplot is least squares. The least-squares regression line is the straight line $\hat{y} = a + bx$ that minimizes the sum of the squares of the vertical distances of the observed points from the line.
- You can use a regression line to predict the value of y for any value of x by substituting this x into the equation of the line.
- The slope b of a regression line $\hat{y} = a + bx$ is the rate at which the predicted response \hat{y} changes along the line as the explanatory variable x changes. Specifically, b is the change in \hat{y} when x increases by 1.
- The intercept a of a regression line $\hat{y} = a + bx$ is the predicted response \hat{y} when the explanatory variable $x = 0$. This prediction is of no statistical use unless x can actually take values near 0.
- The least-squares regression line of y on x is the line with slope rs_y/s_x and intercept $a = \bar{y} - b\bar{x}$. This line always passes through the point (\bar{x}, \bar{y}) .
- Correlation and regression are closely connected. The correlation r is the slope of the least-squares regression line when we measure both x and y in standardized units. The square of the correlation r^2 is the fraction of the variance of one variable that is explained by least-squares regression on the other variable.
- You can examine the fit of a regression line by studying the residuals, which are the differences between the observed and predicted values of y . Be on the lookout for outlying points with unusually large residuals and also for nonlinear patterns and uneven variation about the line.
- Also look for influential observations, individual points that substantially change the regression line. Influential observations are often outliers in the x direction, but they need not have large residuals.

SECTION 2.3 EXERCISES

- 2.42 Review of straight lines. A company manufactures batteries for cell phones. The overhead expenses of keeping the factory operational for a month—even if no batteries are made—total \$500,000. Batteries are manufactured in lots (1000 batteries per lot) costing \$10,000 to make. In this scenario, \$500,000 is the *fixed cost* associated with producing cell phone batteries, and \$10,000 is the *marginal* (or *variable*) cost of producing each lot of batteries. The total monthly cost y of producing x lots of cell phone batteries is given by the equation

$$y = 500,000 + 10,000x$$

- (a) Draw a graph of this equation. (Choose two values of x , such as 0 and 10. Compute the corresponding values of y from the equation. Plot these two points on graph paper and draw the straight line joining them.)

- (b) What will it cost to produce 20 lots of batteries (20,000 batteries)?
 (c) If each lot cost \$20,000 instead of \$10,000 to produce, what is the equation that describes total monthly cost for x lots produced?
- 2.43** **Review of straight lines.** A local consumer electronics store sells exactly 4 DVD players of a particular model each week. The store expects no more shipments of this particular model, and they have 96 such units in their current inventory.
 (a) Give an equation for the number of DVD players of this particular model in inventory after x weeks. What is the slope of this line?
 (b) Draw a graph of this line between now (week 0) and week 10.
 (c) Would you be willing to use this line to predict the inventory after 25 weeks? Do the prediction and think about the reasonableness of the result.
- 2.44** **Review of straight lines.** A cellular telephone company offers two plans. Plan A charges \$20 a month for up to 75 minutes of airtime and \$0.45 per minute above 75 minutes. Plan B charges \$30 a month for up to 250 minutes and \$0.40 per minute above 250 minutes.
 (a) Draw a graph of the Plan A charge against minutes used from 0 to 250 minutes.
 (b) How many minutes a month must the user talk in order for Plan B to be less expensive than Plan A?
- 2.45** **Moving in step?** One reason to invest abroad is that markets in different countries don't move in step. When American stocks go down, foreign stocks may go up. So an investor who holds both bears less risk. That's the theory. Now we read, "The correlation between changes in American and European share prices has risen from 0.4 in the mid-1990s to 0.8 in 2000."¹⁶ Explain to an investor who knows no statistics why this fact reduces the protection provided by buying European stocks.
- 2.46** **Interpreting correlation.** The same article that claims that the correlation between changes in stock prices in Europe and the United States was 0.8 in 2000 goes on to say, "Crudely, that means that movements on Wall Street can explain 80% of price movements in Europe." Is this true? What is the correct percent explained if $r = 0.8$?
- 2.47** **Size and length of bear markets.** Figure 2.6 (page 97) is a scatterplot of the data from Table 1.5 (page 26) on percent decline of the market versus duration of the bear market. Calculation shows that the mean and standard deviation of the durations are
- $$\bar{x} = 10.73 \quad s_x = 8.20$$
- For the declines,
- $$\bar{y} = 24.67 \quad s_y = 11.20$$
- The correlation between duration and decline is $r = 0.6285$.
- (a) Find the equation of the least-squares line for predicting decline from duration.
 (b) What percent of the observed variation in these declines can be attributed to the linear relationship between decline and duration?

- (c) One bear market has a duration of 15 months but a very low decline of 14%. What is the predicted decline for a bear market with duration = 15? What is the residual for this particular bear market?
- 2.48 **Cash or credit?** We might expect credit card purchases to differ from cash purchases at the same store. Let's compare regressions for credit card and cash purchases using the consignment shop data in Table 2.1 (page 88). 
- Make a scatterplot of daily gross sales y versus items sold x for credit card purchases. Using a separate plot symbol or color, add daily gross sales and items sold for cash. It is somewhat difficult to compare the two patterns by eye.
 - Regression can help. Find the equations of the two least-squares regression lines of gross sales on items sold, for credit card sales and for cash sales. Draw these lines on your plot. What are the slopes of the two regression lines? Explain carefully what comparing the slopes says about credit card purchases versus cash purchases.
- 2.49 **Keeping water clean.** Manufacturing companies (and the Environmental Protection Agency) monitor the quality of the water near their facilities. Measurements of pollutants in water are indirect—a typical analysis involves forming a dye by a chemical reaction with the dissolved pollutant, then passing light through the solution and measuring its “absorbance.” To calibrate such measurements, the laboratory measures known standard solutions and uses regression to relate absorbance to pollutant concentration. This is usually done every day. Here is one series of data on the absorbance for different levels of nitrates. Nitrates are measured in milligrams per liter of water.¹⁷
- | Nitrates | 50 | 50 | 100 | 200 | 400 | 800 | 1200 | 1600 | 2000 | 2000 |
|------------|-----|-----|------|------|------|------|-------|-------|-------|-------|
| Absorbance | 7.0 | 7.5 | 12.8 | 24.0 | 47.0 | 93.0 | 138.0 | 183.0 | 230.0 | 226.0 |
- Chemical theory says that these data should lie on a straight line. If the correlation is not at least 0.997, something went wrong and the calibration procedure is repeated. Plot the data and find the correlation. Must the calibration be done again?
 - What is the equation of the least-squares line for predicting absorbance from concentration? If the lab analyzed a specimen with 500 milligrams of nitrates per liter, what do you expect the absorbance to be? Based on your plot and the correlation, do you expect your predicted absorbance to be very accurate?
- 2.50 **Investing at home and overseas.** Investors ask about the relationship between returns on investments in the United States and on investments overseas. Table 2.6 gives the total returns on U.S. and overseas common stocks over a 30-year period. (The total return is change in price plus any dividends paid, converted into U.S. dollars. Both returns are averages over many individual stocks.)¹⁸
- Make a scatterplot suitable for predicting overseas returns from U.S. returns.
 - Find the correlation and r^2 . Describe the relationship between U.S. and overseas returns in words, using r and r^2 to make your description more precise.

TABLE 2.6 Annual total return on overseas and U.S. stocks

Year	Overseas % return	U.S. % return	Year	Overseas % return	U.S. % return
1971	29.6	14.6	1986	69.4	18.6
1972	36.3	18.9	1987	24.6	5.1
1973	-14.9	-14.8	1988	28.5	16.8
1974	-23.2	-26.4	1989	10.6	31.5
1975	35.4	37.2	1990	-23.0	-3.1
1976	2.5	23.6	1991	12.8	30.4
1977	18.1	-7.4	1992	-12.1	7.6
1978	32.6	6.4	1993	32.9	10.1
1979	4.8	18.2	1994	6.2	1.3
1980	22.6	32.3	1995	11.2	37.6
1981	-2.3	-5.0	1996	6.4	23.0
1982	-1.9	21.5	1997	2.1	33.4
1983	23.7	22.4	1998	20.3	28.6
1984	7.4	6.1	1999	27.2	21.0
1985	56.2	31.6	2000	-14.0	-9.1

(c) Find the least-squares regression line of overseas returns on U.S. returns. Draw the line on the scatterplot. Are you confident that predictions using the regression line will be quite accurate? Why?

(d) Circle the point that has the largest residual (either positive or negative). What year is this? Are there any points that seem likely to be very influential?

- 2.51 **Influence on correlation.** The *Correlation and Regression* applet at www.whfreeman.com/pbs allows you to create a scatterplot and to move points by dragging with the mouse. Click to create a group of 10 points in the lower-left corner of the scatterplot with a strong straight-line pattern (correlation about 0.9).

(a) Add one point at the upper right that is in line with the first 10. How does the correlation change?

(b) Drag this last point down until it is opposite the group of 10 points. How small can you make the correlation? Can you make the correlation negative? You see that a single outlier can greatly strengthen or weaken a correlation. Always plot your data to check for outlying points.

- 2.52 **Influence in regression.** As in the previous exercise, create a group of 10 points in the lower-left corner of the scatterplot with a strong straight-line pattern (correlation at least 0.9). Click the “Show least-squares line” box to display the regression line.

(a) Add one point at the upper right that is far from the other 10 points but exactly on the regression line. Why does this outlier have no effect on the line even though it changes the correlation?

(b) Now drag this last point down until it is opposite the group of 10 points. You see that one end of the least-squares line chases this single point,



TABLE 2.7 Four data sets for exploring correlation and regression

Data Set A											
x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
Data Set B											
x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data Set C											
x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
Data Set D											
x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

while the other end remains near the middle of the original group of 10. What about the last point makes it so influential?

- 2.53 **How many calories?** Table 2.5 (page 109) gives data on the true calories in 10 foods and the average guesses made by a large group of people. Exercise 2.26 explored the influence of two outlying observations on the correlation.
- (a) Make a scatterplot suitable for predicting guessed calories from true calories. Circle the points for spaghetti and snack cake on your plot. These points lie outside the linear pattern of the other 8 points.
 - (b) Find the least-squares regression line of guessed calories on true calories. Do this twice, first for all 10 data points and then leaving out spaghetti and snack cake.
 - (c) Plot both lines on your graph. (Make one dashed so you can tell them apart.) Are spaghetti and snack cake, taken together, influential observations? Explain your answer.
- 2.54 **Always plot your data!** Table 2.7 presents four sets of data prepared by the statistician Frank Anscombe to illustrate the dangers of calculating without first plotting the data.¹⁹
- (a) Without making scatterplots, find the correlation and the least-squares regression line for all four data sets. What do you notice? Use the regression line to predict y for $x = 10$.
 - (b) Make a scatterplot for each of the data sets and add the regression line to each plot.
 - (c) In which of the four cases would you be willing to use the regression line to describe the dependence of y on x ? Explain your answer in each case.
- 2.55 **What's my grade?** In Professor Friedman's economics course, the correlation between the students' total scores before the final examination and their final examination scores is $r = 0.6$. The pre-exam totals for all students in the course have mean 280 and standard deviation 30. The final-exam scores

have mean 75 and standard deviation 8. Professor Friedman has lost Julie's final exam but knows that her total before the exam was 300. He decides to predict her final-exam score from her pre-exam total.

- (a) What is the slope of the least-squares regression line of final-exam scores on pre-exam total scores in this course? What is the intercept?
- (b) Use the regression line to predict Julie's final-exam score.
- (c) Julie doesn't think this method accurately predicts how well she did on the final exam. Calculate r^2 and use the value you get to argue that her actual score could have been much higher (or much lower) than the predicted value.

- 2.56 **Investing at home and overseas.** Exercise 2.50 examined the relationship between returns on U.S. and overseas stocks. Investors also want to know what typical returns are and how much year-to-year variability (called *volatility* in finance) there is. Regression and correlation do not answer these questions.

- (a) Find the five-number summaries for both U.S. and overseas returns, and make side-by-side boxplots to compare the two distributions.
- (b) Were returns generally higher in the United States or overseas during this period? Explain your answer.
- (c) Were returns more volatile (more variable) in the United States or overseas during this period? Explain your answer.

- 2.57 **Missing work.** Data on number of days of work missed and annual salary increase for a company's employees show that in general employees who missed more days of work during the year received smaller raises than those who missed fewer days. Number of days missed explained 64% of the variation in salary increases. What is the numerical value of the correlation between number of days missed and salary increase?

- 2.58 **Your mileage may vary.** Figure 2.2 (page 90) displays a scatterplot of city versus highway MPG for all 2001 model two-seater cars. Calculation shows that the mean and standard deviation of the city MPG are

$$\bar{x} = 18.86 \quad s_x = 8.38$$

For the highway MPG,

$$\bar{y} = 25.83 \quad s_y = 8.53$$

The correlation between city MPG and highway MPG is $r = 0.9840$.

- (a) Find the equation of the least-squares line for predicting city MPG from highway MPG for 2001 model two-seater cars.
- (b) What percent of the observed variation in these city MPG can be attributed to the linear relationship between city MPG and highway MPG?
- (c) One car, the Mercedes-Benz SL600, has a highway rating of 19 miles per gallon. What is the predicted city MPG for a 2001 model two-seater with a highway MPG of 19? What is the residual for this particular car?

- 2.59 **Will I bomb the final?** We expect that students who do well on the midterm exam in a course will usually also do well on the final exam. Gary Smith of Pomona College looked at the exam scores of all 346 students who took his statistics class over a 10-year period.²⁰ The least-squares line for predicting final-exam score from midterm exam score was $\hat{y} = 46.6 + 0.41x$.

Octavio scores 10 points above the class mean on the midterm. How many points above the class mean do you predict that he will score on the final? (Hint: Use the fact that the least-squares line passes through the point (\bar{x}, \bar{y}) and the fact that Octavio's midterm score is $\bar{x} + 10$. This is an example of the phenomenon that gave "regression" its name: students who do well on the midterm will on the average do less well, but still above average, on the final.)

- 2.60 **Four-wheel drive minivans.** Exercise 2.19 (page 104) displays data on all three four-wheel drive 2001 model minivans. The least-squares regression line for these data is

$$\hat{y} = -4.5742 + 0.8065x$$

- (a) If the minivan with a highway MPG of 28.7 had a city MPG of 19.3 rather than 18.6, how would the least-squares regression line change? Find the least-squares line for the altered data. In words, describe the change in the line resulting from the change in this one observation. What name is given to an observation like this one?
- (b) If a fourth observation were added with a highway MPG equal to the average of the highway MPGs of the three minivans ($\bar{x} = 29.0667$) and a city MPG of 20, the least-squares regression line for the data set with four observations would be $\hat{y} = -4.2909 + 0.8065x$. In words, describe the change in the line resulting from adding this particular "new" observation. (This illustrates the effect an outlier has on the least-squares regression line.)
- 2.61 **Where the stores are.** Exercise 2.18 (page 102) looks at data on the number of Target and Wal-Mart stores in each of the 50 states.
- (a) Use software to calculate the least-squares regression line for predicting the number of Target stores from the number of Wal-Mart stores.
- (b) Using the line from (a), predict the number of Target stores in Texas based on the number of Wal-Mart stores in Texas (254 Wal-Mart stores). What is the residual for the Texas point?
- (c) Use software to calculate the least-squares regression line for predicting the number of Wal-Mart stores from the number of Target stores.
- (d) Using the line from (c), predict the number of Wal-Mart stores in Texas based on the number of Target stores in Texas (90 Target stores). What is the residual for the Texas point here? (This illustrates how essential the distinction between explanatory and response variables is in regression.)

2.4 Cautions about Correlation and Regression

Correlation and regression are powerful tools for describing the relationship between two variables. When you use these tools, you must be aware of their limitations, beginning with the fact that **correlation and regression describe only linear relationships**. Also remember that the **correlation r and the least-squares regression line are not resistant**. One influential observation or incorrectly entered data point can greatly change these measures. Always

plot your data before interpreting regression or correlation. Here are some other cautions to keep in mind when you apply correlation and regression or read accounts of their use.

Beware extrapolation

Associations for variables can be trusted only for the range of values for which data have been collected. Even a very strong relationship may not hold outside the data's range.

EXAMPLE 2.15

Predicting the future

Here are data on the number of Target stores in operation at the end of each year in the early 1990s:²¹

Year (x)	1990	1991	1992	1993
Stores (y)	420	463	506	554

These points lie almost exactly on a straight line, with $r = 0.9996$. The least-squares regression line for these data predicts 864 Target stores in operation by the end of 2000. However, the Target 2000 Annual Report shows 977 stores in operation at the end of 2000. The very strong linear trend evident in the 1990–1993 data did not continue to the year 2000.



Predictions made far beyond the range for which data have been collected can't be trusted. Few relationships are linear for *all* values of x . It is risky to stray far from the range of x -values that actually appear in your data.

EXTRAPOLATION

Extrapolation is the use of a regression line for prediction far outside the range of values of the explanatory variable x that you used to obtain the line. Such predictions are often not accurate.

APPLY YOUR KNOWLEDGE

- 2.62 The declining farm population. The number of people living on American farms declined steadily during the 20th century. Here are data on farm population (millions of persons) from 1935 to 1980:

Year	1935	1940	1945	1950	1955
Population	32.1	30.5	24.4	23.0	19.1

Year	1960	1965	1970	1975	1980
Population	15.6	12.4	9.7	8.9	7.2

- (a) Make a scatterplot of these data and find the least-squares regression line of farm population on year.
- (b) According to the regression line, how much did the farm population decline each year on the average during this period? What percent of the observed variation in farm population is accounted for by linear change over time?
- (c) Use the regression equation to predict the number of people living on farms in 1990. Is this result reasonable? Why?

Beware correlations based on averaged data

Many regression and correlation studies work with averages or other measures that combine information from many individuals. You should note this carefully and resist the temptation to apply the results of such studies to individuals. Figure 2.3 (page 92) shows a strong relationship between outside temperature and the Sanchez household's natural-gas consumption. Each point on the scatterplot represents a month. Both degree-days and gas consumed are averages over all the days in the month. The data for individual days would form a cloud around each month's averages—they would have more scatter and lower correlation. Averaging over an entire month smooths out the day-to-day variation due to doors left open, house guests using more gas to heat water, and so on. **Correlations based on averages are usually too high when applied to individuals.** This is another reminder that it is important to note exactly what variables were measured in a statistical study.

APPLY YOUR KNOWLEDGE

- 2.63 Stock market indexes. The Standard & Poor's 500-stock index is an average of the price of 500 stocks. There is a moderately strong correlation (roughly $r = 0.6$) between how much this index changes in January and how much it changes during the entire year. If we looked instead at data on all 500 individual stocks, we would find a quite different correlation. Would the correlation be higher or lower? Why?

Beware the lurking variable

Correlation and regression describe the relationship between two variables. Often the relationship between two variables is strongly influenced by other variables. We try to measure potentially influential variables. We can then use more advanced statistical methods to examine all of the relationships revealed by our data. Sometimes, however, the relationship between two variables is influenced by other variables that we did not measure or even think about. Variables lurking in the background, measured or not, often help explain statistical associations.

LURKING VARIABLE

A **lurking variable** is a variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.

A lurking variable can falsely suggest a strong relationship between x and y , or it can hide a relationship that is really there. Here are examples of each of these effects.

EXAMPLE 2.16**Discrimination in medical treatment?**

Studies show that men who complain of chest pain are more likely to get detailed tests and aggressive treatment such as bypass surgery than are women with similar complaints. Is this association between gender and treatment due to discrimination?

Perhaps not. Men and women develop heart problems at different ages—women are on the average between 10 and 15 years older than men. Aggressive treatments are more risky for older patients, so doctors may hesitate to recommend them. Lurking variables—the patient's age and condition—may explain the relationship between gender and doctors' decisions. As the author of one study of the issue said, "When men and women are otherwise the same and the only difference is gender, you find that treatments are very similar."²²

EXAMPLE 2.17**Measuring inadequate housing**

A study of housing conditions in the city of Hull, England, measured several variables for each of the city's wards. Figure 2.20 is a simplified version of one of the study's findings. The variable x measures how crowded the ward is. The variable y measures the fraction of housing units that have no indoor toilet. We expect inadequate housing to be crowded (high x) and lack toilets (high y). That is, we expect a strong correlation between x and y . In fact the correlation was only $r = 0.08$. How can this be?

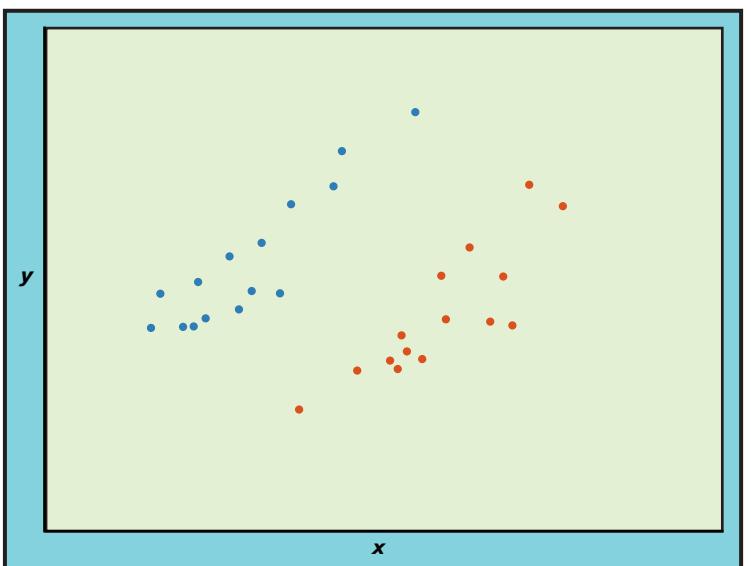


FIGURE 2.20 The variables in this scatterplot have a small correlation even though there is a strong correlation within each of the two clusters.

Figure 2.20 shows that there are two clusters of wards. The wards in the lower cluster have a lot of public housing. These wards are crowded (high values of x), but they have low values of y because public housing always includes indoor toilets. The wards in the upper cluster lack public housing and have high values of both x and y . Within wards of each type, there is a strong positive association between x and y . In Figure 2.20, $r = 0.85$ and $r = 0.91$ in the two clusters. However, because similar values of x correspond to quite different values of y in the two clusters, x alone is of little value for predicting y . Analyzing all wards together ignores the lurking variable—amount of public housing—and hides the nature of the relationship between x and y .²³

APPLY YOUR KNOWLEDGE

- 2.64 **Education and income.** There is a strong positive correlation between years of education and income for economists employed by business firms. In particular, economists with doctorates earn more than economists with only a bachelor's degree. There is also a strong positive correlation between years of education and income for economists employed by colleges and universities. But when all economists are considered, there is a *negative* correlation between education and income. The explanation for this is that business pays high salaries and employs mostly economists with bachelor's degrees, while colleges pay lower salaries and employ mostly economists with doctorates. Sketch a scatterplot with two groups of cases (business and academic) illustrating how a strong positive correlation within each group and a negative overall correlation can occur together. (Hint: Begin by studying Figure 2.20.)

Association is not causation

When we study the relationship between two variables, we often hope to show that changes in the explanatory variable *cause* changes in the response variable. But a strong association between two variables is not enough to draw conclusions about cause and effect. Sometimes an observed association really does reflect cause and effect. The Sanchez household uses more natural gas in colder months because cold weather requires burning more gas to stay warm. In other cases, an association is explained by lurking variables, and the conclusion that x causes y is either wrong or not proved. Here are several examples.

EXAMPLE 2.18**Does television extend life?**

Measure the number of television sets per person x and the average life expectancy y for the world's nations. There is a high positive correlation: nations with many TV sets have higher life expectancies.

The basic meaning of causation is that by changing x we can bring about a change in y . Could we lengthen the lives of people in Rwanda by shipping them TV sets? No. Rich nations have more TV sets than poor nations. Rich nations also have longer life expectancies because they offer better nutrition, clean water, and better health care. There is no cause-and-effect tie between TV sets and length of life.

Correlations such as that in Example 2.18 are sometimes called “nonsense correlations.” The correlation is real. What is nonsense is the conclusion that changing one of the variables causes changes in the other. A lurking variable—such as national wealth in Example 2.18—that influences both x and y can create a high correlation even though there is no direct connection between x and y .

EXAMPLE 2.19**Firestone tires**

Example 1.2 (page 6) presented data on the different Firestone tire models involved in 2969 accidents believed to be caused by defective tires. Most (84.6%) of the accidents involved the Wilderness, Wilderness AT, or ATX tire models, but this is not enough information to conclude that these 3 models are the worst of the 9 models listed in the example. At least two scenarios are possible:

- If these brands account for most of the Firestone tires on the road, then we would expect to see more of them fail because there are more of them on the road.
- If these brands account for only a small portion of the Firestone tires on the road, then it seems that a large percent of these particular models are failing, indicating a potential problem with these 3 models.

The lurking variable here is number of each tire model in use. The failure rate (number failed/number in use) of each tire model is the real variable of interest in identifying defective models.

These and other examples lead us to the most important caution about correlation, regression, and statistical association between variables in general.

ASSOCIATION DOES NOT IMPLY CAUSATION

An association between an explanatory variable x and a response variable y , even if it is very strong, is not by itself good evidence that changes in x actually cause changes in y .

experiment

The best way to get good evidence that x causes y is to do an **experiment** in which we change x and keep lurking variables under control. We will discuss experiments in Chapter 3. When experiments cannot be done, finding the explanation for an observed association is often difficult and controversial. Many of the sharpest disputes in which statistics plays a role involve questions of causation that cannot be settled by experiment. Does gun control reduce violent crime? Does using cell phones cause brain tumors? Has increased free trade widened the gap between the incomes of more-educated and less-educated American workers? All of these questions have become public issues. All concern associations among variables. And all have this in common: they try to pinpoint cause and effect in a setting involving complex relations among many interacting variables.

EXAMPLE 2.20**Does smoking cause lung cancer?**

Despite the difficulties, it is sometimes possible to build a strong case for causation in the absence of experiments. The evidence that smoking causes lung cancer is about as strong as nonexperimental evidence can be.

Doctors had long observed that most lung cancer patients were smokers. Comparison of smokers and “similar” nonsmokers showed a very strong association between smoking and death from lung cancer. Could the association be explained by lurking variables? Might there be, for example, a genetic factor that predisposes people both to nicotine addiction and to lung cancer? Smoking and lung cancer would then be positively associated even if smoking had no direct effect on the lungs. How were these objections overcome?



Let’s answer this question in general terms: What are the criteria for establishing causation when we cannot do an experiment?

- *The association is strong.* The association between smoking and lung cancer is very strong.
- *The association is consistent.* Many studies of different kinds of people in many countries link smoking to lung cancer. That reduces the chance that a lurking variable specific to one group or one study explains the association.
- *Higher doses are associated with stronger responses.* People who smoke more cigarettes per day or who smoke over a longer period get lung cancer more often. People who stop smoking reduce their risk.
- *The alleged cause precedes the effect in time.* Lung cancer develops after years of smoking. The number of men dying of lung cancer rose as smoking became more common, with a lag of about 30 years. Lung cancer kills more men than any other form of cancer. Lung cancer was rare among women until women began to smoke. Lung cancer in women rose along with smoking, again with a lag of about 30 years, and has now passed breast cancer as the leading cause of cancer death among women.
- *The alleged cause is plausible.* Experiments with animals show that tars from cigarette smoke do cause cancer.

Medical authorities do not hesitate to say that smoking causes lung cancer. The U.S. Surgeon General has long stated that cigarette smoking is “the largest avoidable cause of death and disability in the United States.”²⁴ The evidence for causation is overwhelming—but it is not as strong as evidence provided by well-designed experiments.

2.65 **Do firefighters make fires worse?** Someone says, “There is a strong positive correlation between the number of firefighters at a fire and the amount of damage the fire does. So sending lots of firefighters just causes more damage.” Explain why this reasoning is wrong.

2.66 **How’s your self-esteem?** People who do well tend to feel good about themselves. Perhaps helping people feel good about themselves will help

APPLY YOUR KNOWLEDGE

them do better in their jobs and in life. For a time, raising self-esteem became a goal in many schools and companies. Can you think of explanations for the association between high self-esteem and good performance other than “Self-esteem causes better work”?

- 2.67 **Are big hospitals bad for you?** A study shows that there is a positive correlation between the size of a hospital (measured by its number of beds x) and the median number of days y that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital? Why?

BEYOND THE BASICS: DATA MINING

Chapters 1 and 2 of this book are devoted to the important aspect of statistics called *exploratory data analysis* (EDA). We use graphs and numerical summaries to examine data, searching for patterns and paying attention to striking deviations from the patterns we find. In discussing regression, we advanced to using the pattern we find (in this case, a linear pattern) for prediction.

Suppose now that we have a truly enormous data base, such as all purchases recorded by the cash register scanners of our retail chain during the past week. Surely this trove of data contains patterns that might guide business decisions. If we could see clearly the types of activewear preferred in large California cities and compare the preferences of small Midwest cities—right now, not at the end of the season—we might improve profits in both parts of the country by matching stock with demand. This sounds much like EDA, and indeed it is. There is, however, a saying in computer science that a big enough difference of scale amounts to a difference of kind. Exploring really large data bases in the hope of finding useful patterns is called **data mining**. Here are some distinctive features of data mining:

- When you have 100 gigabytes of data, even straightforward calculations and graphics become impossibly time-consuming. So efficient algorithms are very important.
- The structure of the data base and the process of storing the data, perhaps by unifying data scattered across many departments of a large corporation, require careful thought. The fashionable term is *data warehousing*.
- Data mining requires automated tools that work based on only vague queries by the user. The process is too complex to do step-by-step as we have done in EDA.

All of these features point to the need for sophisticated computer science as a basis for data mining. Indeed, data mining is often thought of as a part of computer science. Yet many statistical ideas and tools—mostly tools for dealing with multidimensional data, not the sort of thing that appears in a first statistics course—are very helpful. Like many modern developments, data mining crosses the boundaries of traditional fields of study.

data mining

Do remember that the perils we find with blind use of correlation and regression are yet more perilous in data mining, where the fog of an immense data base prevents clear vision. Extrapolation, ignoring lurking variables, and confusing association with causation are traps for the unwary data miner.

SECTION 2.4 SUMMARY

- Correlation and regression must be interpreted with caution. Plot the data to be sure the relationship is roughly linear and to detect outliers and influential observations.
- Avoid extrapolation, the use of a regression line for prediction for values of the explanatory variable far outside the range of the data from which the line was calculated.
- Remember that correlations based on averages are usually too high when applied to individuals.
- Lurking variables that you did not measure may explain the relations between the variables you did measure. Correlation and regression can be misleading if you ignore important lurking variables.
- Most of all, be careful not to conclude that there is a cause-and-effect relationship between two variables just because they are strongly associated. **High correlation does not imply causation.** The best evidence that an association is due to causation comes from an **experiment** in which the explanatory variable is directly changed and other influences on the response are controlled.

SECTION 2.4 EXERCISES

2.68 Calories and salt in hot dogs. Exercise 2.13 (page 99) gives data on the calories and sodium content for 17 brands of meat hot dogs. The scatterplot shows one outlier.

- Calculate two least-squares regression lines for predicting sodium from calories, one using all of the observations and the other omitting the outlier. Draw both lines on a scatterplot. Does a comparison of the two regression lines show that the outlier is influential? Explain your answer.
- A new brand of meat hot dog has 150 calories per frank. How many milligrams of sodium do you estimate that one of these hot dogs contains?

2.69 Is math the key to success in college? Here is the opening of a newspaper account of a College Board study of 15,941 high school graduates:

Minority students who take high school algebra and geometry succeed in college at almost the same rate as whites, a new study says.

The link between high school math and college graduation is “almost magical,” says College Board President Donald Stewart, suggesting “math is the gatekeeper for success in college.”

TABLE 2.8 Price and consumption of beef, 1970–1993

Year	Price per pound (1993 dollars)	Consumption (lb./capita)	Year	Price per pound (1993 dollars)	Consumption (lb./capita)
1970	3.721	84.62	1982	3.570	77.03
1971	3.789	83.93	1983	3.396	78.64
1972	4.031	85.27	1984	3.274	78.41
1973	4.543	80.51	1985	3.069	79.19
1974	4.212	85.58	1986	2.989	78.83
1975	4.106	88.15	1987	3.032	73.84
1976	3.698	94.36	1988	3.057	72.65
1977	3.477	91.76	1989	3.096	69.34
1978	3.960	87.29	1990	3.107	67.78
1979	4.423	78.10	1991	3.059	66.79
1980	4.098	76.56	1992	2.931	66.48
1981	3.731	77.26	1993	2.934	65.06

“These findings,” he says, “justify serious consideration of a national policy to ensure that all students take algebra and geometry.”²⁵

What lurking variables might explain the association between taking several math courses in high school and success in college? Explain why requiring algebra and geometry may have little effect on who succeeds in college.

- 2.70 **Do artificial sweeteners cause weight gain?** People who use artificial sweeteners in place of sugar tend to be heavier than people who use sugar. Does this mean that artificial sweeteners cause weight gain? Give a more plausible explanation for this association.
- 2.71 **Beef consumption.** Table 2.8 gives data on the amount of beef consumed (pounds per person) and average retail price of beef (dollars per pound) in the United States for the years 1970 to 1993. Because all prices were generally rising during this period, the prices given are “real prices” in 1993 dollars. These are dollars with the buying power that a dollar had in 1993.
- (a) Economists expect consumption of an item to fall when its real price rises. Make a scatterplot of beef consumption y against beef price x . Do you see a relationship of the type expected?
 - (b) Find the equation of the least-squares line and draw the line on your plot. What proportion of the variation in beef consumption is explained by regression on beef price?
 - (c) Although it appears that price helps explain consumption, the scatterplot seems to show some nonlinear patterns. Find the residuals from your regression in (b) and plot them against time. Connect the successive points by line segments to help see the pattern. Are there systematic effects of time remaining after we regress consumption on price? (A partial explanation is that beef production responds to price changes only after some time lag.)

TABLE 2.9 Seafood price per pound, 1970 and 1980

Species	1970 price	1980 price
Cod	13.1	27.3
Flounder	15.3	42.4
Haddock	25.8	38.7
Menhaden	1.8	4.5
Ocean perch	4.9	23.0
Salmon, chinook	55.4	166.3
Salmon, coho	39.3	109.7
Tuna, albacore	26.7	80.1
Clams, soft	47.5	150.7
Clams, blue, hard	6.6	20.3
Lobsters, American	94.7	189.7
Oysters, eastern	61.1	131.3
Sea scallops	135.6	404.2
Shrimp	47.6	149.0

2.72 Seafood prices. The price of seafood varies with species and time. Table 2.9 gives the prices in cents per pound received in 1970 and 1980 by fishermen and vessel owners for several species.

- (a) Plot the data with the 1970 price on the x axis and the 1980 price on the y axis.
- (b) Describe the overall pattern. Are there any outliers? If so, circle them on your graph. Do these unusual points have large residuals from a fitted line? Are they influential in the sense that removing them would change the fitted line?
- (c) Compute the correlation for the entire set of data. What percent of the variation in 1980 prices is explained by the 1970 prices?
- (d) Recompute the correlation discarding the cases that you circled in (b). Do these observations have a strong effect on the correlation? Explain why or why not.
- (e) Does the correlation provide a good measure of the relationship between the 1970 and 1980 prices for this set of data? Explain your answer.

2.73 Using totaled data. Table 2.1 (page 88) gives daily sales data for a consignment shop. The correlation between the total daily sales and the total item counts for the days is $r = 0.9540$.



- (a) Find r^2 and explain in simple language what this number tells us.
- (b) If you calculated the correlation between the sales and item counts of a large number of individual transactions rather than using daily totals, would you expect the correlation to be about 0.95 or quite different? Explain your answer.

2.74 Does herbal tea help nursing-home residents? A group of college students believes that herbal tea has remarkable powers. To test this belief, they make weekly visits to a local nursing home, where they visit with the

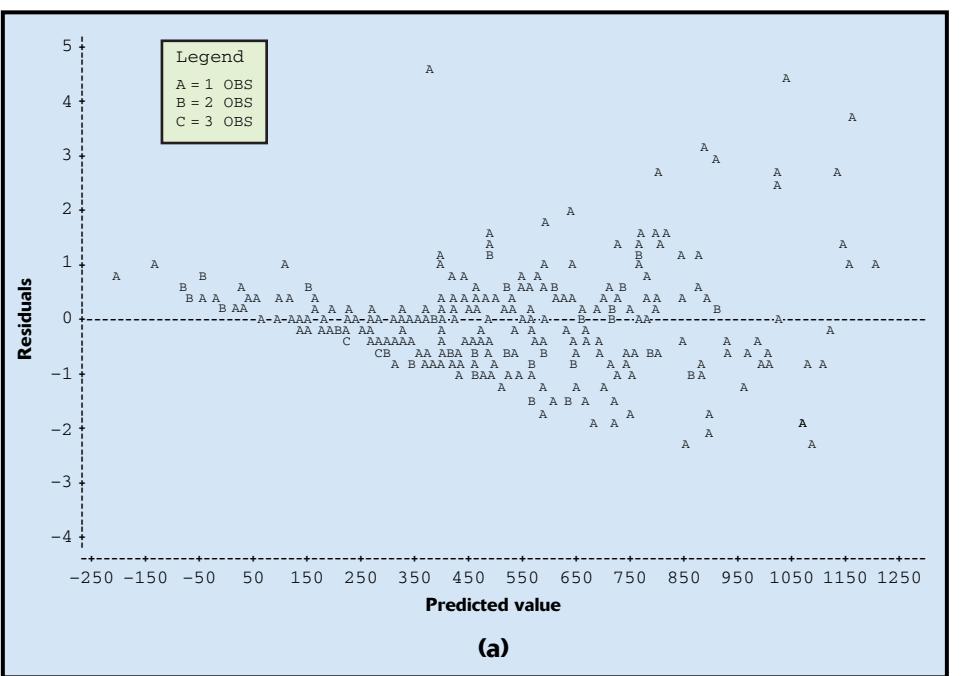
residents and serve them herbal tea. The nursing-home staff reports that after several months many of the residents are healthier and more cheerful. We should commend the students for their good deeds but doubt that herbal tea helped the residents. Identify the explanatory and response variables in this informal study. Then explain what lurking variables account for the observed association.

- 2.75 **Education and income.** There is a strong positive correlation between years of schooling completed x and lifetime earnings y for American men. One possible reason for this association is causation: more education leads to higher-paying jobs. But lurking variables may explain some of the correlation. Suggest some lurking variables that would explain why men with more education earn more.
- 2.76 **Do power lines cause cancer?** It has been suggested that electromagnetic fields of the kind present near power lines can cause leukemia in children. Experiments with children and power lines are not ethical. Careful studies have found no association between exposure to electromagnetic fields and childhood leukemia.²⁶ Suggest several lurking variables that you would want information about in order to investigate the claim that living near power lines is associated with cancer.
- 2.77 **Baseball salaries.** Are baseball players paid according to their performance? To study this question, a statistician analyzed the salaries of over 260 major league hitters along with such explanatory variables as career batting average, career home runs per time at bat, and years in the major leagues. This is a *multiple regression* with several explanatory variables. More detail on multiple regression appears in Chapter 11, but the fit of the model is assessed just as we have done in this chapter, by calculating and plotting the residuals:

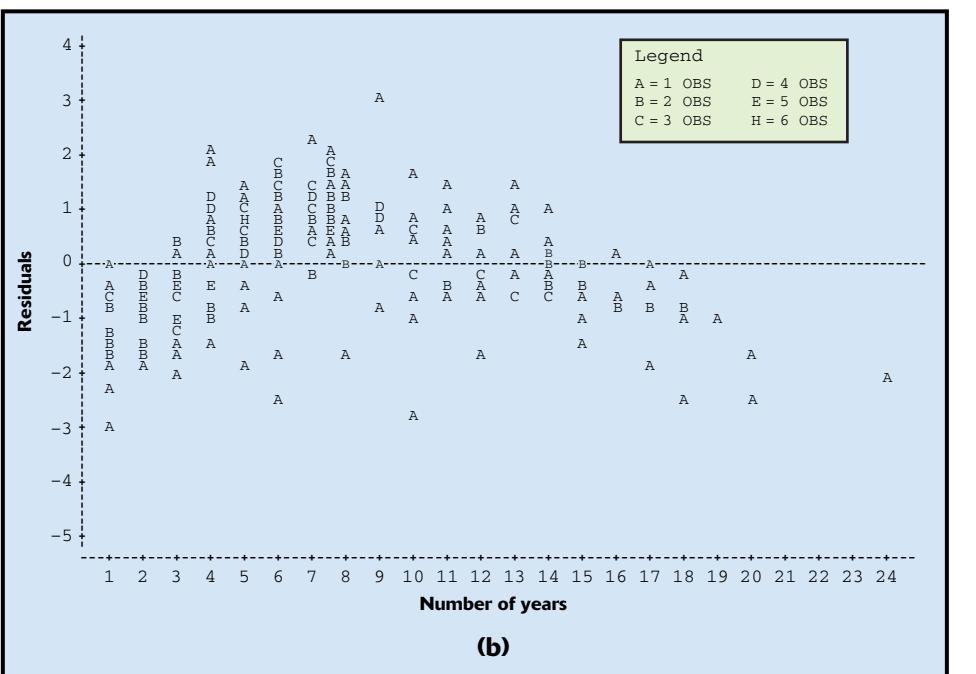
$$\text{residual} = \text{observed } y - \text{predicted } y$$

(This analysis was done by Crystal Richard.)

- (a) Figure 2.21(a) is a plot of the residuals against the predicted salary. When points are too close together to plot separately, the plot uses letters of the alphabet to show how many points there are at each position. Describe the pattern that appears on this residual plot. Will the regression model predict high or low salaries more precisely?
- (b) After studying the residuals in more detail, the statistician decided to predict the logarithm of salary rather than the salary itself. One reason was that while salaries are not Normally distributed (the distribution is skewed to the right), their logarithms are nearly Normal. When the response variable is the logarithm of salary, a plot of the residuals against the predicted value is satisfactory—it looks like Figure 2.17(a) (page 122). Figure 2.21(b) is a plot of the residuals against the number of years the player has been in the major leagues. Describe the pattern that you see. Will the model overestimate or underestimate the salaries of players who are new to the majors? Of players who have been in the major leagues about 8 years? Of players with more than 15 years in the majors?



(a)



(b)

FIGURE 2.21 Two residual plots for the regression of baseball players' salaries on their performance, for Exercise 2.77.

2.5 Relations in Categorical Data*

We have concentrated on relationships in which at least the response variable is quantitative. Now we will shift to describing relationships between two or more categorical variables. Some variables—such as gender, race, and occupation—are categorical by nature. Other categorical variables are created by grouping values of a quantitative variable into classes. Published data often appear in grouped form to save space. To analyze categorical data, we use the *counts* or *percents* of individuals that fall into various categories.



MARITAL STATUS AND JOB LEVEL

We sometimes read in popular magazines that getting married is good for your career. Table 2.10 presents data from one of the studies behind this generalization. To avoid gender effects, the investigators looked only at men. The data describe the marital status and the job level of all 8235 male managers and professionals employed by a large manufacturing firm.²⁷ The firm assigns each position a grade that reflects the value of that particular job to the company. The authors of the study grouped the many job grades into quarters. Grade 1 contains jobs in the lowest quarter of the job grades, and grade 4 contains those in the highest quarter.



*two-way table
row and column
variables*

Table 2.10 is a **two-way table** because it describes two categorical variables. Job grade is the **row variable** because each row in the table describes one job grade group. Marital status is the **column variable** because each column describes one marital status group. The entries in the table are the counts of men in each marital status-by-job grade class. Both marital status and job grade in this table are categorical variables. Job grade has a natural order from lowest to highest. The order of the rows in Table 2.10 reflects the order of the job grades.

Marginal distributions

How can we best grasp the information contained in Table 2.10? First, *look at the distribution of each variable separately*. The distribution of a categorical variable says how often each outcome occurred. The “Total” column at the right of the table contains the totals for each of the rows. These row totals give the distribution of job grade (the row variable) among all men in the study: 955 are at job grade 1, 4239 are at job grade 2, and so on. In the same way, the “Total” row at the bottom of the table gives the *marital status distribution*. If the row and column totals are missing,

*This material is important in statistics, but it is needed later in this book only for Chapter 9. You may omit it if you do not plan to read Chapter 9 or delay reading it until you reach Chapter 9.

TABLE 2.10 Marital status and job level

Job grade	Marital status				Total
	Single	Married	Divorced	Widowed	
1	58	874	15	8	955
2	222	3927	70	20	4239
3	50	2396	34	10	2490
4	7	533	7	4	551
Total	337	7730	126	42	8235

marginal distribution

first thing to do in studying a two-way table is to calculate them. The distributions of job grade alone and marital status alone are called **marginal distributions** because they appear at the right and bottom margins of the two-way table.

Percents are often more informative than counts. We can display the marginal distribution of job grade in terms of percents by dividing each row total by the table total and converting to a percent.

EXAMPLE 2.21

Calculating a marginal distribution

The percent of men in the study who have a grade 1 job is

$$\frac{\text{total with job grade 1}}{\text{table total}} = \frac{955}{8235} = 0.1160 = 11.60\%$$

Do three more such calculations to obtain the marginal distribution of job grade in percents. Here it is:

Job grade	1	2	3	4
Percent	11.60	51.48	30.24	6.69

The total should be 100% because everyone in the study is in one of the job grade categories. The total is in fact 100.01%, due to roundoff error.



Each marginal distribution from a two-way table is a distribution for a single categorical variable. As we saw in Example 1.2 (page 6), we can use a bar graph or a pie chart to display such a distribution. Figure 2.22 is a bar graph of the distribution of job grade. We see that men with jobs of grade 2 make up about half of the total.

In working with two-way tables, you must calculate lots of percents. Here's a tip to help decide what fraction gives the percent you want. Ask, "What group represents the total that I want a percent of?" The count for that group is the denominator of the fraction that leads to the percent. In Example 2.21, we wanted a percent "of men in the study," so the table total is the denominator.

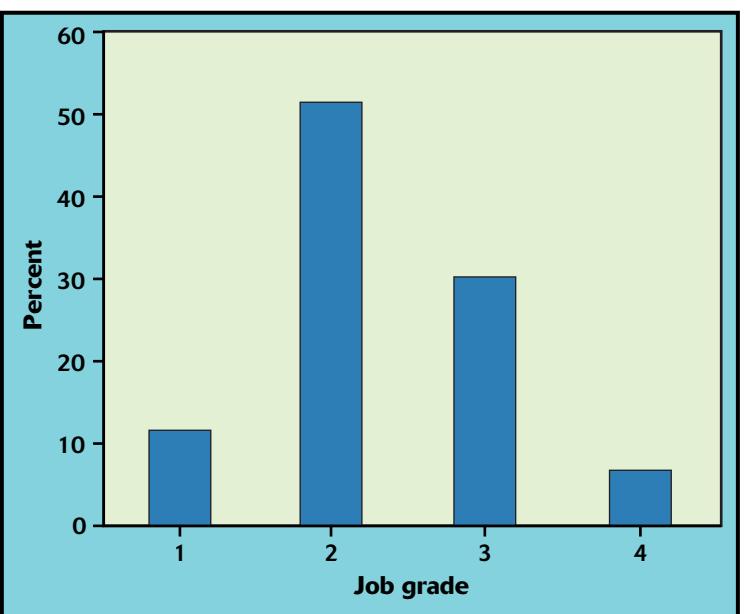


FIGURE 2.22 A bar graph of the distribution of job grade for the men described in Case 2.2. This is one of the marginal distributions for Table 2.10.

APPLY YOUR KNOWLEDGE

- 2.78 **Marital status.** Give the marginal distribution of marital status (in percents) for the men in the study of Case 2.2, starting from the counts in Table 2.10. CASE 2.2
- 2.79 **Smoking by students and their parents.** Advertising by tobacco companies is believed to encourage students to smoke and use other tobacco products. Another source of “encouragement” may be the example set by parents with respect to tobacco use. Here are data from eight high schools on smoking among students and among their parents.²⁸

	Neither parent smokes	One parent smokes	Both parents smoke
Student does not smoke	1168	1823	1380
Student smokes	188	416	400

- (a) How many students do these data describe?
 (b) What percent of these students smoke?
 (c) Give the marginal distribution of parents’ smoking behavior, both in counts and in percents.

Describing relationships

Table 2.10 contains much more information than the two marginal distributions of marital status alone and job grade alone. The nature of the relationship between marital status and job grade cannot be deduced from the separate distributions but requires the full table. Relationships among categorical variables are described by calculating appropriate percents from the counts given. We use percents because counts are often hard to compare. For example, 874 married men have jobs of grade 1, and only 58 single men

have jobs of grade 1. These counts do not accurately describe the association, however, because there are many more married men in the study.

EXAMPLE 2.22

How common are grade 1 jobs?

What percent of single men have grade 1 jobs? This is the count of those who are single and have grade 1 jobs as a percent of the single group total:

$$\frac{58}{337} = 0.1721 = 17.21\%$$

In the same way, find the percent of men in each marital status group who have grade 1 jobs. The comparison of all four groups is

Marital status	Single	Married	Divorced	Widowed
Percent with grade 1 jobs	17.21	11.31	11.90	19.05

These percentages make it clear that grade 1 jobs are less common among married and divorced men than among single and widowed men. Don't leap to any causal conclusions: single men as a group, for example, are probably younger and so have not yet advanced to higher-grade jobs.



- 2.80 **The top jobs.** Using the counts in Table 2.10, find the percent of men in each marital status group who have grade 4 jobs. Draw a bar graph that compares these percents. Explain briefly what the data show.
- 2.81 **Smoking by students.** Using the table in Exercise 2.79 on smoking by students and parents, calculate the percent of students in each group (neither parent smokes, one parent smokes, both parents smoke) that smoke. One might believe that parents' smoking increases smoking in their children. Do the data support that belief? Briefly explain your response.

Conditional distributions

Example 2.22 does not compare the complete distributions of job grade in the four marital status groups. It compares only the percents who have grade 1 jobs. Let's look at the complete picture.

EXAMPLE 2.23

Calculating a conditional distribution

Information about single men occupies the first column in Table 2.10. To find the complete distribution of job grade in this marital status group, look only at that column. Compute each count as a percent of the column total, which is 337. Here is the distribution:

Job grade	1	2	3	4
Percent	17.21	65.88	14.84	2.08

These percents should add to 100% because all single men fall in one of the job grade categories. (In fact, they add to 100.01% because of roundoff error.) The four percents together are the **conditional distribution** of job grade, given that a man is single. We use the term "conditional" because the distribution refers only to men who satisfy the condition that they are single.

conditional distribution

		Marital Status					
		Count	Single	Married	Divorced	Widowed	
Job Grade	1	58	874	15	8	955	
	2	222	3927	70	20	4239	
	3	50	2396	34	10	2490	
	4	7	533	7	4	551	
		337	7730	126	42	8235	

FIGURE 2.23 JMP output of the two-way table of job grade by marital status with the four conditional distributions of job grade, one for each marital status group. The percents in each column add to 100% (up to roundoff error).

Now focus in turn on the second column (married men) and then the third column (divorced men) and finally the fourth column (widowed men) of Table 2.10 to find three more conditional distributions. Statistical software can speed the task of finding each entry in a two-way table as a percent of its column total. Figure 2.23 displays the output.

Each cell in this table contains a count from Table 2.10 along with that count as a percent of the column total. The percents in each column form the conditional distribution of job grade for one marital status group. The percents in each column add to 100% (up to roundoff error) because everyone in the marital status group is accounted for. Figure 2.24 presents all four conditional distributions for comparison with each other and with the overall distribution in Figure 2.22. Comparing these conditional distributions reveals the nature of the association between marital status and job grade. The distributions of job grade in the married, divorced, and widowed groups are quite similar, but grade 4 jobs are less common in the single group.

No single graph (such as a scatterplot) portrays the form of the relationship between categorical variables. No single numerical measure (such as the correlation) summarizes the strength of the association. Bar graphs are flexible enough to be helpful, but you must think about what comparisons you want to display. For numerical measures, we rely on well-chosen percents. You must decide which percents you need. Here is a hint: compare the conditional distributions of the response variable (job grade) for the separate values of the explanatory variable (marital status). That's what we did in Figure 2.24.

APPLY YOUR KNOWLEDGE

- 2.82 Who holds the top jobs? Find the conditional distribution of marital status among men with grade 4 jobs, starting from the counts in Table 2.10. (To do this, look only at the “job grade 4” row in the table.)
- 2.83 Majors for men and women in business. To study the career plans of young women and men, questionnaires were sent to all 722 members of the senior class in the College of Business Administration at the University of Illinois.



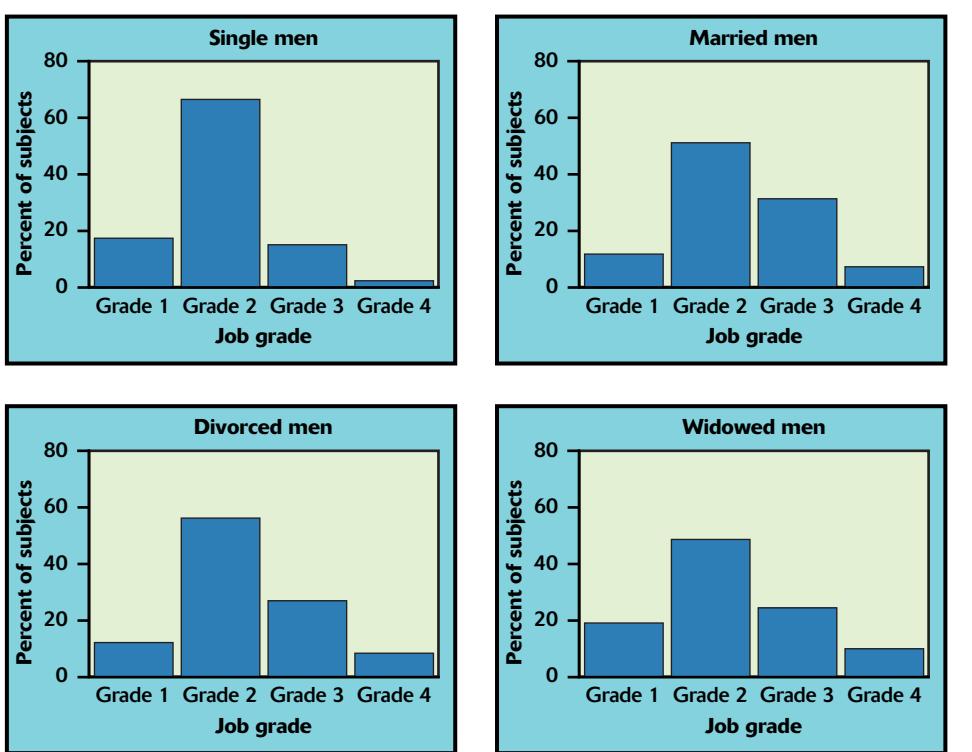


FIGURE 2.24 Bar graphs comparing the distribution of job grade level within each of four marital status groups. These are conditional distributions of job grade given a specific marital status.

One question asked which major within the business program the student had chosen. Here are the data from the students who responded:²⁹

	Female	Male
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

- (a) Find the two conditional distributions of major, one for women and one for men. Based on your calculations, describe the differences between women and men with a graph and in words.
 - (b) What percent of the students did not respond to the questionnaire? The nonresponse weakens conclusions drawn from these data.
- 2.84 Here are the row and column totals for a two-way table with two rows and two columns:

$$\begin{array}{cc|c} a & b & 50 \\ c & d & 50 \\ \hline 60 & 40 & 100 \end{array}$$

Find *two different* sets of counts a , b , c , and d for the body of the table that give these same totals. This shows that the relationship between two variables cannot be obtained from the two individual distributions of the variables.

Simpson's paradox

As is the case with quantitative variables, the effects of lurking variables can change or even reverse relationships between two categorical variables. Here is an example that demonstrates the surprises that can await the unsuspecting user of data.

EXAMPLE 2.24

Which airline is on time?

Air travelers would like their flights to arrive on time. Airlines collect data on on-time arrivals and report them to the government. Following are one month's data for flights from several western cities for two airlines:

	Alaska Airlines	America West
On time	3274	6438
Delayed	501	787
Total	3775	7225

Alaska Airlines is delayed 13.3% ($501/3775$) of the time, and America West is delayed only 10.9% ($787/7225$) of the time. It seems that you should choose America West to minimize delays.

Some cities are more likely to have delays than others, however. If we consider which city the flights left from in our analysis, we obtain a more complete picture of the two airlines' on-time flights. Here are the data broken down by which city each flight left from.³⁰ Check that the entries in the original two-way table are just the sums of the city entries in this table.

	Alaska Airlines		America West	
	On time	Delayed	On time	Delayed
Los Angeles	497	62	694	117
Phoenix	221	12	4840	415
San Diego	212	20	383	65
San Francisco	503	102	320	129
Seattle	1841	305	201	61

Alaska Airlines beats America West for flights from Los Angeles: only 11.1% ($62/559$) delayed compared with 14.4% ($117/811$) for America West. Alaska Airlines wins again for flights from Phoenix, with 5.2% ($12/233$) delayed versus 7.9% ($415/5255$). In fact, if you calculate the percent of delayed flights for each city individually, Alaska Airlines has a lower percent of late flights at *all* of these cities. So Alaska Airlines is the better choice no matter which of the five cities you are flying from.

The city of origin for each flight is a lurking variable when we compare the late-flight percents of the two airlines. When we ignore the lurking variable, America West seems better, even though Alaska Airlines does better for each city. How can Alaska Airlines do better for all cities, yet do worse overall? Look at the data: America West flies most often from sunny

Phoenix, where there are few delays. Alaska Airlines flies most often from Seattle, where fog and rain cause frequent delays. The original two-way table, which did not take account of the city of origin for each flight, was misleading. Example 2.24 illustrates *Simpson's paradox*.

SIMPSON'S PARADOX

An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson's paradox**.

The lurking variables in Simpson's paradox are categorical. That is, they break the individuals into groups, as when flights are classified by city of origin. Simpson's paradox is just an extreme form of the fact that observed associations can be misleading when there are lurking variables.

APPLY YOUR KNOWLEDGE

- 2.85 Which hospital is safer? Insurance companies and consumers are interested in the performance of hospitals. The government releases data about patient outcomes in hospitals that can be useful in making informed health care decisions. Here is a two-way table of data on the survival of patients after surgery in two hospitals. All patients undergoing surgery in a recent time period are included. "Survived" means that the patient lived at least 6 weeks following surgery.

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
Total	2100	800

- (a) What percent of Hospital A patients died? What percent of Hospital B patients died? These are the numbers one might see reported in the media.

Not all surgery cases are equally serious, however. Patients are classified as being in either "poor" or "good" condition before surgery. Here are the data broken down by patient condition. Check that the entries in the original two-way table are just the sums of the "poor" and "good" entries in this pair of tables.

Good Condition		Poor Condition			
	Hospital A	Hospital B			
Died	6	8	Died	57	8
Survived	594	592	Survived	1443	192
Total	600	600	Total	1500	200

- (b) Find the percent of Hospital A patients who died who were classified as "poor" before surgery. Do the same for Hospital B. In which hospital do "poor" patients fare better?

- (c) Repeat (b) for patients classified as “good” before surgery.
- (d) What is your recommendation to someone facing surgery and choosing between these two hospitals?
- (e) How can Hospital A do better in both groups, yet do worse overall?
Look at the data and carefully explain how this can happen.

SECTION 2.5 SUMMARY

- A two-way table of counts organizes data about two categorical variables. Values of the **row variable** label the rows that run across the table, and values of the **column variable** label the columns that run down the table. Two-way tables are often used to summarize large amounts of information by grouping outcomes into categories.
- The **row totals** and **column totals** in a two-way table give the **marginal distributions** of the two individual variables. It is clearer to present these distributions as percents of the table total. Marginal distributions tell us nothing about the relationship between the variables.
- To find the **conditional distribution** of the row variable for one specific value of the column variable, look only at that one column in the table. Find each entry in the column as a percent of the column total.
- There is a conditional distribution of the row variable for each column in the table. Comparing these conditional distributions is one way to describe the association between the row and the column variables. It is particularly useful when the column variable is the explanatory variable.
- **Bar graphs** are a flexible means of presenting categorical data. There is no single best way to describe an association between two categorical variables.
- A comparison between two variables that holds for each individual value of a third variable can be changed or even reversed when the data for all values of the third variable are combined. This is **Simpson’s paradox**. Simpson’s paradox is an example of the effect of lurking variables on an observed association.

SECTION 2.5 EXERCISES

College undergraduates. Exercises 2.86 to 2.90 are based on Table 2.11. This two-way table reports data on all undergraduate students enrolled in U.S. colleges and universities in the fall of 1997 whose age was known.³¹

2.86 College undergraduates.

- (a) How many undergraduate students were enrolled in colleges and universities?
- (b) What percent of all undergraduate students were 18 to 24 years old in the fall of the academic year?

TABLE 2.11**Undergraduate college enrollment, fall 1997
(thousands of students)**

Age	2-year part-time	2-year full-time	4-year part-time	4-year full-time
Under 18	45	170	83	55
18 to 24	1478	1202	4759	562
25 to 39	421	1344	1234	1273
40 and up	121	748	236	611
Total	2064	3464	6311	2501

- (c) Find the percent of the undergraduates enrolled in each of the four types of programs who were 18 to 24 years old. Make a bar graph to compare these percents.
- (d) The 18 to 24 group is the traditional age group for college students. Briefly summarize what you have learned from the data about the extent to which this group predominates in different kinds of college programs.
- 2.87 Two-year college students.**
- (a) An association of two-year colleges asks: “What percent of students enrolled part-time at 2-year colleges are 25 to 39 years old?” Find the percent.
- (b) A bank that makes education loans to adults asks: “What percent of all 25- to 39-year-old students are enrolled part-time at 2-year colleges?” Find the percent.
- 2.88 Students’ ages.**
- (a) Find the marginal distribution of age among all undergraduate students, first in counts and then in percents. Make a bar graph of the distribution in percents.
- (b) Find the conditional distribution of age (in percents) among students enrolled part-time in 2-year colleges and make a bar graph of this distribution.
- (c) Briefly describe the most important differences between the two age distributions.
- (d) The sum of the entries in the “2-year full-time” column is not the same as the total given for that column. Why is this?
- 2.89 Older students.** Call students aged 40 and up “older students.” Compare the presence of older students in the four types of program with numbers, a graph, and a brief summary of your findings.
- 2.90 Nontraditional students.** With a little thought, you can extract from Table 2.11 information other than marginal and conditional distributions. The traditional college age group is ages 18 to 24 years.
- (a) What percent of all undergraduates fall in this age group?
- (b) What percent of students at 2-year colleges fall in this age group?
- (c) What percent of part-time students fall in this group?
- 2.91 Hiring practices.** A company has been accused of age discrimination in hiring for operator positions. Lawyers for both sides look at data on applicants for

the past 3 years. They compare hiring rates for applicants less than 40 years old and those aged 40 years or greater.

Age	Hired	Not hired
< 40	79	1148
≥ 40	1	163

- (a) Find the two conditional distributions of hired/not hired, one for applicants who are less than 40 years old and one for applicants who are not less than 40 years old.
- (b) Based on your calculations, make a graph to show the differences in distribution for the two age categories.
- (c) Describe the company's hiring record in words. Does the company appear to discriminate on the basis of age?
- (d) What lurking variables might be involved here?
- 2.92 Nonresponse in a survey.** A business school conducted a survey of companies in its state. They mailed a questionnaire to 200 small companies, 200 medium-sized companies, and 200 large companies. The rate of nonresponse is important in deciding how reliable survey results are. Here are the data on response to this survey:

	Small	Medium	Large
Response	125	81	40
No response	75	119	160
Total	200	200	200

- (a) What was the overall percent of nonresponse?
- (b) Describe how nonresponse is related to the size of the business. (Use percents to make your statements precise.)
- (c) Draw a bar graph to compare the nonresponse percents for the three size categories.
- 2.93 Helping cocaine addicts.** Cocaine addiction is hard to break. Addicts need cocaine to feel any pleasure, so perhaps giving them an antidepressant drug will help. A 3-year study with 72 chronic cocaine users compared an antidepressant drug called desipramine (manufactured by Hoechst-Marion-Roussel) with lithium and a placebo. (Lithium is a standard drug to treat cocaine addiction. A placebo is a dummy drug, used so that the effect of being in the study but not taking any drug can be seen.) One-third of the subjects, chosen at random, received each drug. Here are the results.³²

	Desipramine	Lithium	Placebo
Relapse	10	18	20
No relapse	14	6	4
Total	24	24	24

- (a) Compare the effectiveness of the three treatments in preventing relapse. Use percents and draw a bar graph.
- (b) Do you think that this study gives good evidence that desipramine actually *causes* a reduction in relapses?

- 2.94 **Employee performance.** Four employees are responsible for handling the cash register at your store. One item that needs to be recorded for each sale is the type of payment: cash, check, or credit card. For a number of transactions this information is missing, though. Are certain employees responsible, or is everyone equally guilty in forgetting to record this information? Below is a table summarizing the last 3372 transactions.³³

	Not recorded	Recorded
Employee 1	68	897
Employee 2	62	679
Employee 3	90	1169
Employee 4	39	497

- (a) What percent of all transactions do not have the payment terms recorded?
 - (b) Compare the reliability of the four employees in recording the payment terms for each transaction they handle. Use percents and draw a bar graph.
 - (c) Do you think these data provide good evidence that a certain employee (or subset of employees) is causing the high percent of transactions without payment terms recorded?
- 2.95 **Demographics and new products.** Companies planning to introduce a new product to the market must define the “target” for the product. Who do we hope to attract with our new product? Age and gender are two of the most important demographic variables. The following two-way table describes the age and marital status of American women in 1999.³⁴ The table entries are in thousands of women.

Age (years)	Marital status				Total
	Never married	Married	Widowed	Divorced	
18 to 24	10,240	2,598	9	184	13,031
25 to 39	7,640	20,129	193	2,930	30,891
40 to 64	3,234	28,923	2,357	6,764	41,278
≥ 65	751	8,270	8,385	1,263	18,667
Total	21,865	59,918	10,944	11,141	103,867

- (a) Find the sum of the entries in the “Married” column. Why does this sum differ from the “Total” entry for that column?
- (b) Give the marginal distribution of marital status for all adult women (use percents). Draw a bar graph to display this distribution.

2.96 **Demographics, continued.**

- (a) Using the data in the previous exercise, compare the conditional distributions of marital status for women aged 18 to 24 and women aged 40 to 64. Briefly describe the most important differences between the two groups of women, and back up your description with percents.
- (b) Your company is planning a magazine aimed at women who have never been married. Find the conditional distribution of age among never-married women and display it in a bar graph. What age group or groups should your magazine aim to attract?

- 2.97 **Discrimination?** Wabash Tech has two professional schools, business and law. Here are two-way tables of applicants to both schools, categorized by gender and admission decision. (Although these data are made up, similar situations occur in reality.)³⁵

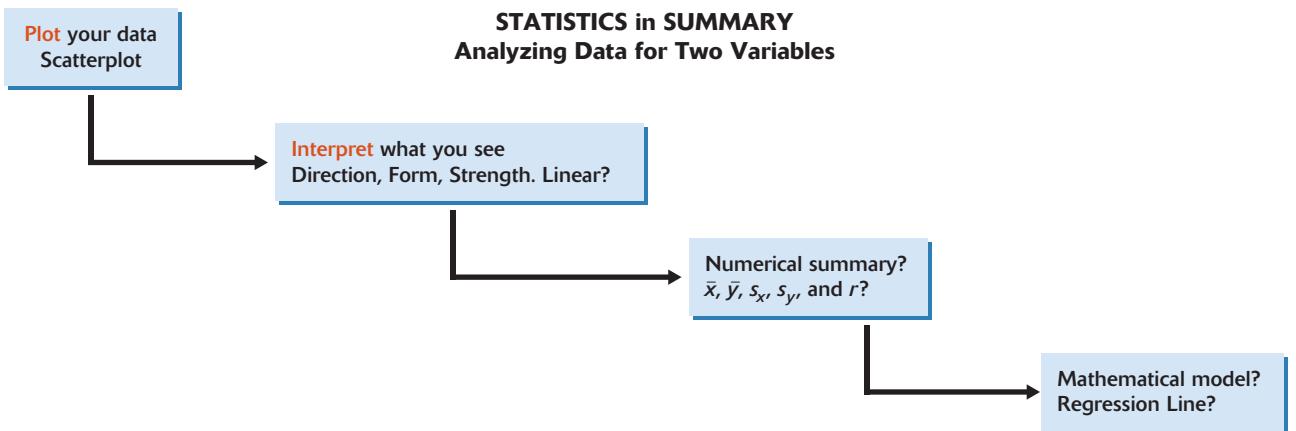
Business		Law			
	Admit	Deny		Admit	Deny
Male	480	120	Male	10	90
Female	180	20	Female	100	200

- (a) Make a two-way table of gender by admission decision for the two professional schools together by summing entries in these tables.
 - (b) From the two-way table, calculate the percent of male applicants who are admitted and the percent of female applicants who are admitted. Wabash admits a higher percent of male applicants.
 - (c) Now compute separately the percents of male and female applicants admitted by the business school and by the law school. Each school admits a higher percent of female applicants.
 - (d) This is Simpson's paradox: both schools admit a higher percent of the women who apply, but overall Wabash admits a lower percent of female applicants than of male applicants. Explain carefully, as if speaking to a skeptical reporter, how it can happen that Wabash appears to favor males when each school individually favors females.
- 2.98 **Obesity and health.** Recent studies have shown that earlier reports underestimated the health risks associated with being overweight. The error was due to overlooking lurking variables. In particular, smoking tends both to reduce weight and to lead to earlier death. Illustrate Simpson's paradox by a simplified version of this situation. That is, make up tables of overweight (yes or no) by early death (yes or no) by smoker (yes or no) such that
- Overweight smokers and overweight nonsmokers both tend to die earlier than those not overweight.
 - But when smokers and nonsmokers are combined into a two-way table of overweight by early death, persons who are not overweight tend to die earlier.

STATISTICS IN SUMMARY

Chapter 1 dealt with data analysis for a single variable. In this chapter, we have studied analysis of data for two or more variables. The proper analysis depends on whether the variables are categorical or quantitative and on whether one is an explanatory variable and the other a response variable.

When you have a categorical explanatory variable and a quantitative response variable, use the tools of Chapter 1 to compare the distributions of the response variable for the different categories of the explanatory variable. Make side-by-side boxplots, stemplots, or histograms and compare medians or means. If both variables are categorical, there is no satisfactory graph (though bar graphs can help). We describe the relationship numerically by comparing percents. The optional Section 2.5 explains how to do this.



Most of this chapter concentrates on relations between two quantitative variables. The Statistics in Summary figure organizes the main ideas in a way that stresses that our tactics are the same as when we faced single-variable data in Chapter 1. Here is a review list of the most important skills you should have gained from studying this chapter.

A. DATA

1. Recognize whether each variable is quantitative or categorical.
2. Identify the explanatory and response variables in situations where one variable explains or influences another.

B. SCATTERPLOTS

1. Make a scatterplot to display the relationship between two quantitative variables. Place the explanatory variable (if any) on the horizontal scale of the plot.
2. Add a categorical variable to a scatterplot by using a different plotting symbol or color.
3. Describe the form, direction, and strength of the overall pattern of a scatterplot. In particular, recognize positive or negative association and linear (straight-line) patterns. Recognize outliers in a scatterplot.

C. CORRELATION

1. Using a calculator or software, find the correlation r between two quantitative variables.
2. Know the basic properties of correlation: r measures the strength and direction of only linear relationships; $-1 \leq r \leq 1$ always; $r = \pm 1$ only for perfect straight-line relations; r moves away from 0 toward ± 1 as the linear relation gets stronger.

D. STRAIGHT LINES

1. Explain what the slope b and the intercept a mean in the equation $y = a + bx$ of a straight line.
2. Draw a graph of a straight line when you are given its equation.

E. REGRESSION

1. Using a calculator or software, find the least-squares regression line of a response variable y on an explanatory variable x from data.
2. Find the slope and intercept of the least-squares regression line from the means and standard deviations of x and y and their correlation.
3. Use the regression line to predict y for a given x . Recognize extrapolation and be aware of its dangers.
4. Use r^2 to describe how much of the variation in one variable can be accounted for by a straight-line relationship with another variable.
5. Recognize outliers and potentially influential observations from a scatterplot with the regression line drawn on it.
6. Calculate the residuals and plot them against the explanatory variable x or against other variables. Recognize unusual patterns.

F. LIMITATIONS OF CORRELATION AND REGRESSION

1. Understand that both r and the least-squares regression line can be strongly influenced by a few extreme observations.
2. Recognize that a correlation based on averages of several observations is usually stronger than the correlation for individual observations.
3. Recognize possible lurking variables that may explain the observed association between two variables x and y .
4. Understand that even a strong correlation does not mean that there is a cause-and-effect relationship between x and y .

G. CATEGORICAL DATA (Optional)

1. From a two-way table of counts, find the marginal distributions of both variables by obtaining the row sums and column sums.
2. Express any distribution in percents by dividing the category counts by their total.
3. Describe the relationship between two categorical variables by computing and comparing percents. Often this involves comparing the conditional distributions of one variable for the different categories of the other variable.
4. Recognize Simpson's paradox and be able to explain it.

CHAPTER 2 REVIEW EXERCISES

- 2.99 Are high interest rates bad for stocks? When interest rates are high, investors may shun stocks because they can get high returns with less risk. Figure 2.25 plots the annual return on U.S. common stocks for the years 1950 to 2000 against the returns on Treasury bills for the same years.³⁶ (The interest rate paid by Treasury bills is a measure of how high interest rates were that year.) Describe the pattern you see. Are high interest rates bad for stocks? Is there a strong relationship between interest rates and stock returns?

- 2.100 Are high interest rates bad for stocks? The scatterplot in Figure 2.25 suggests that returns on common stocks may be somewhat lower in years with high interest rates. Here is part of the Excel output for the regression of stock returns on the bill returns for the same years:

	A	B	C	D	E	F	G	
1								
2								
<i>Regression Statistics</i>								
4	Multiple R	0.097538007						
5	R Square	0.009513663						
6	Adjusted R Square	-0.010700344						
7	Standard Error	17.16003473						
8	Observations	51						
9								
10	ANOVA							
11		df	SS	MS	F	Significance F		
12	Regression	1	138.5899296	138.58993	0.47064706	0.495923077		
13	Residual	49	14428.87281	294.46679				
14	Total	50	14567.46274					
15								
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	
17	Intercept	16.23521332	4.960252104	3.2730621	0.00195431	6.267219563	26.20320707	
18	Tbill return	-0.572668524	0.834748484	-0.6860372	0.49592308	-2.250157389	1.104820341	

- (a) What is the equation of the least-squares line? Use this line to predict the percent return on stocks in a year when Treasury bills return 5%.
- (b) Explain what the slope of the regression line tells us. Does the slope confirm that high interest rates are in general bad for stocks?
- (c) If you knew the return on Treasury bills for next year, do you think you could predict the return on stocks quite accurately? Use both the scatterplot in Figure 2.25 and the regression output to justify your answer.
- 2.101 Influence? The scatterplot in Figure 2.25 contains an outlier: in 1981, inflation reached a peak and the return on Treasury bills was 14.72%.

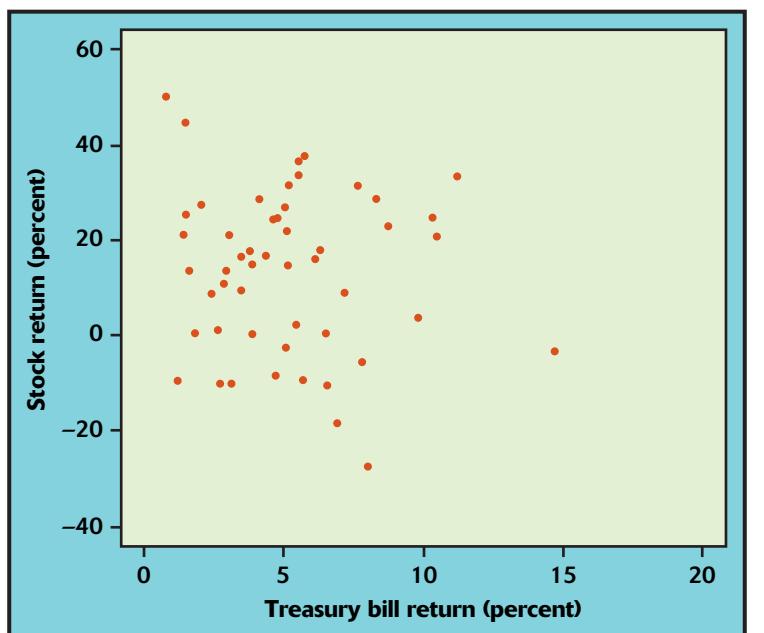


FIGURE 2.25 Annual returns on U.S. common stocks versus returns on Treasury bills for the years 1950 to 2000, for Exercise 2.99.

Would removing this point increase or decrease the correlation between returns on T-bills and stocks? Why? Is the location of this point such that it could strongly influence the regression line? Why?

- 2.102 **A hot stock?** It is usual in finance to describe the returns from investing in a single stock by regressing the stock's returns on the returns from the stock market as a whole. This helps us see how closely the stock follows the market. We analyzed the monthly percent total return y on Philip Morris common stock and the monthly return x on the Standard & Poor's 500-stock index, which represents the market, for the period between July 1990 and May 1997. Here are the results:

$$\begin{aligned}\bar{x} &= 1.304 & s_x &= 3.392 & r &= 0.5251 \\ \bar{y} &= 1.878 & s_y &= 7.554\end{aligned}$$

A scatterplot shows no very influential observations.

- (a) Find the equation of the least-squares line from this information. What percent of the variation in Philip Morris stock is explained by the linear relationship with the market as a whole?
 - (b) Explain carefully what the slope of the line tells us about how Philip Morris stock responds to changes in the market. This slope is called "beta" in investment theory.
 - (c) Returns on most individual stocks have a positive correlation with returns on the entire market. That is, when the market goes up, an individual stock tends also to go up. Explain why an investor should prefer stocks with $\text{beta} > 1$ when the market is rising and stocks with $\text{beta} < 1$ when the market is falling.
- 2.103 **What correlation doesn't say.** Investment reports now often include correlations. Following a table of correlations among mutual funds, a report adds: "Two funds can have perfect correlation, yet different levels of risk. For example, Fund A and Fund B may be perfectly correlated, yet Fund A moves 20% whenever Fund B moves 10%." Write a brief explanation, for someone who knows no statistics, of how this can happen. Include a sketch to illustrate your explanation.
- 2.104 **A computer game.** A multimedia statistics learning system includes a test of skill in using the computer's mouse. The software displays a circle at a random location on the computer screen. The subject tries to click in the circle with the mouse as quickly as possible. A new circle appears as soon as the subject clicks the old one. Table 2.12 gives data for one subject's trials, 20 with each hand. Distance is the distance from the cursor location to the center of the new circle, in units whose actual size depends on the size of the screen. Time is the time required to click in the new circle, in milliseconds.³⁷
- (a) We suspect that time depends on distance. Make a scatterplot of time against distance, using separate symbols for each hand.
 - (b) Describe the pattern. How can you tell that the subject is right-handed?
 - (c) Find the regression line of time on distance separately for each hand. Draw these lines on your plot. Which regression does a better job of predicting time from distance? Give numerical measures that describe the success of the two regressions.
 - (d) It is possible that the subject got better in later trials due to learning. It is also possible that he got worse due to fatigue. Plot the residuals from

TABLE 2.12 Reaction times in a computer game

Time	Distance	Hand	Time	Distance	Hand
115	190.70	right	240	190.70	left
96	138.52	right	190	138.52	left
110	165.08	right	170	165.08	left
100	126.19	right	125	126.19	left
111	163.19	right	315	163.19	left
101	305.66	right	240	305.66	left
111	176.15	right	141	176.15	left
106	162.78	right	210	162.78	left
96	147.87	right	200	147.87	left
96	271.46	right	401	271.46	left
95	40.25	right	320	40.25	left
96	24.76	right	113	24.76	left
96	104.80	right	176	104.80	left
106	136.80	right	211	136.80	left
100	308.60	right	238	308.60	left
113	279.80	right	316	279.80	left
123	125.51	right	176	125.51	left
111	329.80	right	173	329.80	left
95	51.66	right	210	51.66	left
108	201.95	right	170	201.95	left

each regression against the time order of the trials (down the columns in Table 2.12). Is either of these systematic effects of time visible in the data?

- 2.105 Wood products.** A wood product manufacturer is interested in replacing solid-wood building material with less-expensive products made from wood flakes. The company collected the following data to examine the relationship between the length (in inches) and the strength (in pounds per square inch) of beams made from wood flakes:

Length	5	6	7	8	9	10	11	12	13	14
Strength	446	371	334	296	249	254	244	246	239	234

- (a) Make a scatterplot that shows how the length of a beam affects its strength.
- (b) Describe the overall pattern of the plot. Are there any outliers?
- (c) Fit a least-squares line to the entire set of data. Graph the line on your scatterplot. Does a straight line adequately describe these data?
- (d) The scatterplot suggests that the relation between length and strength can be described by *two* straight lines, one for lengths of 5 to 9 inches and another for lengths of 9 to 14 inches. Fit least-squares lines to these two subsets of the data, and draw the lines on your plot. Do they

describe the data adequately? What question would you now ask the wood experts?

- 2.106 Heating a business.** Joan is concerned about the amount of energy she uses to heat her small business. She keeps a record of the natural gas she consumes each month over one year's heating season. The variables are the same as in Example 2.4 (page 91) for the Sanchez home. Here are Joan's data:³⁸

	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
Degree-days per day	15.6	26.8	37.8	36.4	35.5	18.6	15.3	7.9	0.0
Gas consumed per day	520	610	870	850	880	490	450	250	110

- (a) Make a scatterplot of these data. There is a strongly linear pattern with no outliers.
 - (b) Find the equation of the least-squares regression line for predicting gas use from degree-days. Draw this line on your graph. Explain in simple language what the slope of the regression line tells us about how gas use responds to outdoor temperature.
 - (c) Joan adds some insulation during the summer, hoping to reduce her gas consumption. Next February averages 40 degree-days per day, and her gas consumption is 870 cubic feet per day. Predict from the regression equation how much gas the house would have used at 40 degree-days per day last winter before the extra insulation. Did the insulation reduce gas consumption?
- 2.107 Heating a business.** Find the mean and standard deviation of the degree-day and gas consumption data in the previous exercise. Find the correlation between the two variables. Use these five numbers to find the equation of the regression line for predicting gas use from degree-days. Verify that your work agrees with your previous results. Use the same five numbers to find the equation of the regression line for predicting degree-days from gas use. What units does each of these slopes have?
- 2.108 Size and selling price of houses.** Table 2.13 provides information on a random sample of 50 houses sold in Ames, Iowa, in the year 2000.³⁹
- (a) Describe the distribution of selling price with a graph and a numerical summary. What are the main features of this distribution?
 - (b) Make a scatterplot of selling price versus square feet and describe the relationship between these two variables.
 - (c) Calculate the least-squares regression line for these data. On average, how much does each square foot add to the selling price of a house?
 - (d) What would you expect the selling price of a 1600-square-foot house in Ames to be?
 - (e) What percent of the variability in these 50 selling prices can be attributed to differences in square footage?
- 2.109 Age and selling price of houses.**
- (a) Using the data in Table 2.13, calculate the least-squares regression line for predicting selling price from age at time of sale, that is, in 2000.
 - (b) What would you expect for the selling price of a house built in 2000? 1999? 1998? 1997? Describe specifically how age relates to selling price.

TABLE 2.13 Houses sold in Ames, Iowa

Selling price (\$)	Square footage	Age (years)	Selling price (\$)	Square footage	Age (years)
268,380	1897	1	169,900	1686	35
131,000	1157	15	180,000	2054	34
112,000	1024	35	127,000	1386	50
112,000	935	35	242,500	2603	10
122,000	1236	39	152,900	1582	3
127,900	1248	32	171,600	1790	1
157,600	1620	33	195,000	1908	6
135,000	1124	33	83,100	1378	72
145,900	1248	35	125,000	1668	55
126,000	1139	39	60,500	1248	100
142,000	1329	40	85,000	1229	59
107,500	1040	45	117,000	1308	60
110,000	951	42	57,000	892	90
187,000	1628	1	110,000	1981	72
94,000	816	43	127,250	1098	70
99,500	1060	24	119,000	1858	80
78,000	800	68	172,500	2010	60
55,790	492	79	123,000	1680	86
70,000	792	80	161,715	1670	1
53,600	980	62	179,797	1938	1
157,000	1629	3	117,250	1120	36
166,730	1889	0	116,500	914	4
340,000	2759	6	117,000	1008	23
195,000	1811	3	177,500	1920	32
215,850	2400	27	132,000	1146	37

- (c) Would you trust this regression line for predicting the selling price of a house that was built in 1900? 1899? 1850? Explain your responses. (What would this line predict for a house built in 1850?)
- (d) Calculate and interpret the correlation between selling price and age.
- 2.110 Beta.** Exercise 2.102 introduced the financial concept of a stock's "beta." Beta measures the volatility of a stock relative to the overall market. A beta of less than 1 indicates lower risk than the market; a beta of more than 1 indicates higher risk than the market. The information on Apple Computer, Inc., at finance.yahoo.com lists a beta of 1.24; the information on Apple Computer, Inc., at www.nasdaq.com lists a beta of 1.69. Both Web sites use Standard & Poor's 500-stock index to measure changes in the overall market. Using the 60 monthly returns from March 1996 to February 2001, we find the least-squares regression line $\hat{y} = 0.30 + 1.35x$ (y is Apple return, x is Standard & Poor's 500-stock index return).

- (a) The correlation for our 60 months of data is 0.3481. Interpret this correlation in terms of using the movement of the overall market to predict movement in Apple stock.
- (b) We have three different beta values for Apple stock: 1.24, 1.35, and 1.69. What is a likely explanation for this discrepancy?

The following exercises concern material in the optional Section 2.5.

- 2.111 **Aspirin and heart attacks.** Does taking aspirin regularly help prevent heart attacks? “Nearly five decades of research now link aspirin to the prevention of stroke and heart attacks.” So says the Bayer Aspirin Web site, www.bayeraspirin.com. The most important evidence for this claim comes from the Physicians’ Health Study. The subjects were 22,071 healthy male doctors at least 40 years old. Half the subjects, chosen at random, took aspirin every other day. The other half took a placebo, a dummy pill that looked and tasted like aspirin. Here are the results.⁴⁰ (The row for “None of these” is left out of the two-way table.)

	Aspirin group	Placebo group
Fatal heart attacks	10	26
Other heart attacks	129	213
Strokes	119	98
Total	11,037	11,034

What do the data show about the association between taking aspirin and heart attacks and stroke? Use percents to make your statements precise. Do you think the study provides evidence that aspirin actually reduces heart attacks (cause and effect)?

- 2.112 **More smokers live at least 20 more years!** You can see the headlines “More smokers than nonsmokers live at least 20 more years after being contacted for study!” A medical study contacted randomly chosen people in a district in England. Here are data on the 1314 women contacted who were either current smokers or who had never smoked. The tables classify these women by their smoking status and age at the time of the survey and whether they were still alive 20 years later.⁴¹

Age 18 to 44		Age 45 to 64		Age 65 +	
	Smoker	Not		Smoker	Not
Dead	19	13	Dead	78	52
Alive	269	327	Alive	162	147

Age 18 to 44		Age 45 to 64		Age 65 +	
	Smoker	Not		Smoker	Not
Dead	42	165	Dead	7	28
Alive	7	28	Alive	162	147

- (a) From these data make a two-way table of smoking (yes or no) by dead or alive. What percent of the smokers stayed alive for 20 years? What percent of the nonsmokers survived? It seems surprising that a higher percent of smokers stayed alive.
- (b) The age of the women at the time of the study is a lurking variable. Show that within each of the three age groups in the data, a higher percent of nonsmokers remained alive 20 years later. This is another example of Simpson’s paradox.

- (c) The study authors give this explanation: "Few of the older women (over 65 at the original survey) were smokers, but many of them had died by the time of follow-up." Compare the percent of smokers in the three age groups to verify the explanation.



CHAPTER 2 CASE STUDY EXERCISES

CASE STUDY 2.1: Wal-Mart stores. The file *ex02-18.dat* contains data on the number of the nation's 2624 Wal-Mart stores in each state. Table 1.7 gives the populations of the states.

- A. Stores and population. How well does least-squares regression on state population predict the number of Wal-Mart stores in a state? Do an analysis that includes graphs, calculations, and a discussion of your findings. If a state's population increases by a million people, about how many more Wal-Marts would you expect? What other state-to-state differences might account for differences in number of Wal-Mart stores per state?
- B. California. Is California unusual with respect to number of Wal-Mart stores per million people? Does looking at graphs suggest that California may be influential for either correlation or the least-squares line? Redo parts of your analysis without California and draw conclusions about the influence of the largest state on this statistical study.

CASE STUDY 2.2: Predicting coffee exports. The data set *coffeexports.dat*, described in the Data Appendix, contains information on the number of 60-kilo bags of coffee exported from 48 countries in 2000 and 2001. In this Case Study, you will consider three prediction scenarios.

- A. Using December 2000 exports to predict 2001 exports. What is the correlation between December 2000 exports and total exports for 2001? Make a scatterplot with the least-squares line added for predicting 2001 exports using December 2000 exports. What does the value of r^2 tell you? Identify any unusual values by country name.
- B. Using 2000 exports to predict 2001 exports. What is the correlation between 2000 exports and 2001 exports? Make a scatterplot with the least-squares line added for predicting 2001 exports using 2000 exports. What does the value of r^2 tell you in this setting? For the countries listed as having unusual values in Part A, find their data points on the scatterplot in Part B. Are these countries' values unusual in this data set? Do you recommend using 2000 total exports or December 2000 exports for predicting 2001 exports? Briefly explain your choice.
- C. Using December 2000 exports to predict December 2001 exports. What is the correlation between December 2000 exports and December 2001 exports? Make a scatterplot with the least-squares line added for predicting December 2001 exports using December 2000 exports. What does the value of r^2 tell you here? This r^2 -value should be higher than the r^2 -values in Parts A and B. Explain intuitively why one might expect the highest r^2 to be in Part C.

Prelude

Which ad works better?

Will a new TV advertisement sell more Crest toothpaste than the current ad? Procter & Gamble, the maker of Crest, would like to learn the answer without running the risk of replacing the current ad with a new one that might not work as well.

With help from A. C. Nielsen Company, a large market research firm, Procter & Gamble made a direct comparison of the effectiveness of the two commercials. Nielsen enlisted the cooperation of 2500 households in Springfield, Missouri, as well as of all the major stores in town. Each household's TV sets are wired so that Nielsen can replace the regular CBS network broadcast with its own. Only the commercials are different: when CBS broadcasts the current Crest ad, Nielsen shows a new ad. Half of the households, chosen at random, receive the new Crest commercial and the other half see the current ad.

Which group will buy more Crest? Each household has an ID card that can be read by the checkout scanners in the stores. Everything each household buys is noted by the scanners and reported to Nielsen's computers. It's now easy to see if the households that saw the new commercial bought more Crest than those that received the current ad.

Nielsen statisticians can say not only which ad sold more Crest but which demographic groups found it most convincing. Carefully designed production of data combined with statistical analysis enables advertisers to plan their marketing more precisely than in the past.¹



CHAPTER

3

Producing Data

- 3.1 Designing Samples**
- 3.2 Designing Experiments**
- 3.3 Toward Statistical Inference**



Introduction

Numerical data are the raw material for sound conclusions. Executives and investors want data, not merely opinions, about a firm's financial condition. As the Prelude illustrates, marketing managers and advertising agencies can also base their decisions on data rather than relying on subjective impressions. Statistics is concerned with producing data as well as with interpreting already available data.

Chapters 1 and 2 explored the art of data analysis. They showed how to uncover the features of a set of data by applying numerical and graphical techniques. It is helpful to distinguish between two purposes in analyzing data. Sometimes we just want a careful description of some situation. A retail chain considering locations for new stores looks carefully at data on the local population: age, income, ethnic mix, and other variables help managers choose locations where stores will prosper. This is *exploratory data analysis*. We simply want to know what the data say about each potential store site. Our tools are graphs and numerical summaries. Our conclusions apply only to the specific sites that our data describe.

A second purpose in analyzing data is to provide clear answers to specific questions about some setting too broad to examine in complete detail. “Are American consumers planning to reduce their spending?” “Which TV ad will sell more toothpaste?” “Do a majority of college students prefer Pepsi to Coke when they taste both without knowing which they are drinking?” We cannot afford to ask *all* consumers about their spending plans or to show both TV ads to a national audience. If we carefully design the production of data on a smaller scale, we can use our data to draw conclusions about a wider setting. *Statistical inference* uses data to answer specific questions about a setting that goes beyond the data in hand, and it attaches a known degree of confidence to the answers. The success of inference depends on designing the production of data with both the specific questions and the wider setting in mind. This chapter introduces statistical ideas for producing data and concludes with a conceptual introduction to statistical inference. The rest of the book discusses inference in detail, building on the twin foundations of data analysis (Chapters 1 and 2) and data production.

Observation and experiment

We want to know “What percent of American adults agree that the economy is getting better?” To answer the question, we interview American adults. We can't afford to ask all adults, so we put the question to a **sample** chosen to represent the entire adult population. How do we choose a sample that truly represents the opinions of the entire population? Statistical designs for choosing samples are the topic of Section 3.1.

Our goal in choosing a sample is a picture of the population, disturbed as little as possible by the act of gathering information. Sample surveys are one kind of *observational study*. Other kinds of observational studies might watch the behavior of consumers looking at store displays or the interaction between managers and employees.

sample

In other settings, we gather data from an *experiment*. In doing an experiment, we don't just observe individuals or ask them questions. We actively impose some treatment to observe the response. To answer the question "Which TV ad will sell more toothpaste?" we show each ad to a separate group of consumers and note whether they buy the toothpaste. Experiments, like samples, provide useful data only when properly designed. We discuss statistical design of experiments in Section 3.2. The distinction between experiments and observational studies is one of the most important ideas in statistics.

OBSERVATION VERSUS EXPERIMENT

An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses.

An **experiment** deliberately imposes some treatment on individuals to observe their responses.

Observational studies are essential sources of data about topics from public attitudes toward business to the behavior of online shoppers. But an observational study, even one based on a statistical sample, is a poor way to gauge the effect of an intervention. To see the response to a change, we must actually impose the change. When our goal is to understand cause and effect, experiments are the only source of fully convincing data.

EXAMPLE 3.1

Do cell phones cause brain cancer?

As cell phone use grows, concerns about exposure to radiation from longterm use of the phones have also grown. Reports of cell phone users suffering from brain cancer make headlines and plant seeds of fear in some wireless phone users. Are the observed cases of cell phone users with brain cancer evidence that cell phones *cause* brain cancer? Careful observational studies look at large groups, not at isolated cases. To date, they find no consistent relationship between cell phone use and cancer. But observation cannot answer the cause-and-effect question directly. Even if observational data did show a relationship, cell phone users might have some other common characteristics that contribute to the development of brain cancer. For example, they are more likely than non-users to live in metropolitan areas. That means higher exposure to many types of pollution, some of which might cause cancer.

To see if cell phone radiation actually causes brain cancer, we might carry out an experiment. Choose two similar groups of people. Expose one group to cell phone radiation regularly and forbid the other group any close contact with cell phones. Wait a few years, then compare the rates of brain cancer in the two groups. This experiment would answer our question, but it is neither practical nor ethical to force people to accept or avoid cell phones. Experimenters can and do use mice or rats, exposing some to cell phone radiation while keeping others free from it. The experiments have not (yet) been conclusive—those that show more tumors in mice exposed to radiation used mice bred to be prone to cancer and radiation more intense than that produced by phones.²

When we simply observe cell phone use and brain cancer, any effect of radiation on the occurrence of brain cancer is *confounded* (mixed up) with such lurking variables as age, occupation, and place of residence.

CONFOUNDING

Two variables (explanatory variables or lurking variables) are **confounded** when their effects on a response variable cannot be distinguished from each other.

Observational studies of the effect of one variable on another often fail because the explanatory variable is confounded with lurking variables. We will see that well-designed experiments take steps to defeat confounding.

APPLY YOUR KNOWLEDGE

- 3.1 **Gender and consumer choices.** Men and women differ in their choices for many product categories. Are there gender differences in preferences for health insurance plans as well? A market researcher interviews a large sample of consumers, both men and women. She asks each person which of a list of features he or she considers essential in a health plan. Is this study an experiment? Why or why not?
- 3.2 **Teaching economics.** An educational software company wants to compare the effectiveness of its computer animation for teaching about supply, demand, and market clearing with that of a textbook presentation. The company tests the economic knowledge of each of a group of first-year college students, then divides them into two groups. One group uses the animation, and the other studies the text. The company retests all the students and compares the increase in economic understanding in the two groups. Is this an experiment? Why or why not? What are the explanatory and response variables?
- 3.3 **Does job training work?** A state institutes a job-training program for manufacturing workers who lose their jobs. After five years, the state reviews how well the program works. Critics claim that because the state's unemployment rate for manufacturing workers was 6% when the program began and 10% five years later, the program is ineffective. Explain why higher unemployment does not necessarily mean that the training program failed. In particular, identify some lurking variables whose effect on unemployment may be confounded with the effect of the training program.

3.1 Designing Samples

A major food producer wants to know what fraction of the public worries about use of genetically modified corn in taco shells. A quality engineer must estimate what fraction of the bearings rolling off an assembly line is defective. Government economists inquire about household income. In all these situations, we want to gather information about a large group of

people or things. Time, cost, and inconvenience prevent inspecting every bearing or contacting every household. In such cases, we gather information about only part of the group to draw conclusions about the whole.

POPULATION, SAMPLE

The **population** in a statistical study is the entire group of individuals about which we want information.

A **sample** is a part of the population from which we actually collect information, used to draw conclusions about the whole.

Notice that the population is defined in terms of our desire for knowledge. If we wish to draw conclusions about all U.S. college students, that group is our population even if only local students are available for questioning. The sample is the part from which we draw conclusions about the whole. The **design** of a sample refers to the method used to choose the sample from the population. Poor sample designs can produce misleading conclusions.

sample design

EXAMPLE 3.2

A “good” sample isn’t a good idea

A mill produces large coils of thin steel for use in manufacturing home appliances. A quality engineer wants to submit a sample of 5-centimeter squares to detailed laboratory examination. She asks a technician to cut a sample of 10 such squares. Wanting to provide “good” pieces of steel, the technician carefully avoids the visible defects in the steel when cutting the sample. The laboratory results are wonderful, but the customers complain about the material they are receiving.



EXAMPLE 3.3

Call-in opinion polls

Television news programs like to conduct call-in polls of public opinion. The program announces a question and asks viewers to call one telephone number to respond “Yes” and another for “No.” Telephone companies charge for these calls. The ABC network program *Nightline* once asked whether the United Nations should continue to have its headquarters in the United States. More than 186,000 callers responded, and 67% said “No.”

People who spend the time and money to respond to call-in polls are not representative of the entire adult population. In fact, they tend to be the same people who call radio talk shows. People who feel strongly, especially those with strong negative opinions, are more likely to call. It is not surprising that a properly designed sample showed that 72% of adults want the UN to stay.³



Call-in opinion polls are an example of *voluntary response sampling*. A voluntary response sample can easily produce 67% “No” when the truth about the population is close to 72% “Yes.”

VOLUNTARY RESPONSE SAMPLE

A voluntary response sample consists of people who choose themselves by responding to a general appeal. Voluntary response samples are biased because people with strong opinions, especially negative opinions, are most likely to respond.

convenience sampling

Voluntary response is one common type of bad sample design. Another is **convenience sampling**, which chooses the individuals easiest to reach. Here is an example of convenience sampling.

EXAMPLE 3.4**Interviewing at the mall**

Manufacturers and advertising agencies often use interviews at shopping malls to gather information about the habits of consumers and the effectiveness of ads. A sample of mall shoppers is fast and cheap. “Mall interviewing is being propelled primarily as a budget issue,” one expert told the *New York Times*. But people contacted at shopping malls are not representative of the entire U.S. population. They are richer, for example, and more likely to be teenagers or retired. Moreover, mall interviewers tend to select neat, safe-looking individuals from the stream of customers. Decisions based on mall interviews may not reflect the preferences of all consumers.⁴



Both voluntary response samples and convenience samples produce samples that are almost guaranteed not to represent the entire population. These sampling methods display *bias*, or systematic error, in favoring some parts of the population over others.

BIAS

The design of a study is **biased** if it systematically favors certain outcomes.

APPLY YOUR KNOWLEDGE

- 3.4 **Sampling employed women.** A sociologist wants to know the opinions of employed adult women about government funding for day care. She obtains a list of the 520 members of a local business and professional women's club and mails a questionnaire to 100 of these women selected at random. Only 48 questionnaires are returned. What is the population in this study? What is the sample from whom information is actually obtained? What is the rate (percent) of nonresponse?
- 3.5 **What is the population?** For each of the following sampling situations, identify the population as exactly as possible. That is, say what kind of individuals the population consists of and say exactly which individuals fall in the population. If the information given is not sufficient, complete the description of the population in a reasonable way.

- (a) Each week, the Gallup Poll questions a sample of about 1500 adult U.S. residents to determine national opinion on a wide variety of issues.
 - (b) The 2000 census tried to gather basic information from every household in the United States. Also, a “long form” requesting much additional information was sent to a sample of about 17% of households.
 - (c) A machinery manufacturer purchases voltage regulators from a supplier. There are reports that variation in the output voltage of the regulators is affecting the performance of the finished products. To assess the quality of the supplier’s production, the manufacturer sends a sample of 5 regulators from the last shipment to a laboratory for study.
- 3.6 **Is that movie any good?** You wonder if that new “blockbuster” movie is really any good. Some of your friends like the movie, but you decide to check the Internet Movie Database (www.imdb.com) to see others’ ratings. You find that 2497 people chose to rate this movie, with an average rating of only 3.7 out of 10. You are surprised that most of your friends liked the movie, while many people gave low ratings to the movie online. Are you convinced that a majority of those who saw the movie would give it a low rating? What type of sample are your friends? What type of sample are the raters on the Internet Movie Database?

Simple random samples

In a voluntary response sample, people choose whether to respond. In a convenience sample, the interviewer makes the choice. In both cases, personal choice produces bias. The statistician’s remedy is to allow impersonal chance to choose the sample. A sample chosen by chance allows neither favoritism by the sampler nor self-selection by respondents. Choosing a sample by chance attacks bias by giving all individuals an equal chance to be chosen. Rich and poor, young and old, black and white, all have the same chance to be in the sample.

The simplest way to use chance to select a sample is to place names in a hat (the population) and draw out a handful (the sample). This is the idea of *simple random sampling*.

SIMPLE RANDOM SAMPLE

A simple random sample (SRS) of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

An SRS not only gives each individual an equal chance to be chosen (thus avoiding bias in the choice) but also gives every possible sample an equal chance to be chosen. There are other random sampling designs that give each individual, but not each sample, an equal chance. Exercise 3.23 describes one such design, called systematic random sampling.

The idea of an SRS is to choose our sample by drawing names from a hat. In practice, computer software can choose an SRS almost instantly from



a list of the individuals in the population. For example, the *Simple Random Sample* applet available on the Web site for this book can choose an SRS of any size up to $n = 40$ from a population of any size up to 500. If you don't use software, you can randomize by using a *table of random digits*.

RANDOM DIGITS

A **table of random digits** is a long string of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with these two properties:

1. Each entry in the table is equally likely to be any of the 10 digits 0 through 9.
2. The entries are independent of each other. That is, knowledge of one part of the table gives no information about any other part.

Table B at the back of the book is a table of random digits. You can think of Table B as the result of asking an assistant (or a computer) to mix the digits 0 to 9 in a hat, draw one, then replace the digit drawn, mix again, draw a second digit, and so on. The assistant's mixing and drawing save us the work of mixing and drawing when we need to randomize. Table B begins with the digits 19223950340575628713. To make the table easier to read, the digits appear in groups of five and in numbered rows. The groups and rows have no meaning—the table is just a long list of randomly chosen digits. Because the digits in Table B are random:

- Each entry is equally likely to be any of the 10 possibilities 0, 1, ..., 9.
- Each pair of entries is equally likely to be any of the 100 possible pairs 00, 01, ..., 99.
- Each triple of entries is equally likely to be any of the 1000 possibilities 000, 001, ..., 999; and so on.

These "equally likely" facts make it easy to use Table B to choose an SRS. Here is an example that shows how.

EXAMPLE 3.5

How to choose an SRS

Joan's accounting firm serves 30 small business clients. Joan wants to interview a sample of 5 clients in detail to find ways to improve client satisfaction. To avoid bias, she chooses an SRS of size 5.

Step 1: Label. Give each client a numerical label, using as few digits as possible. Two digits are needed to label 30 clients, so we use labels

01, 02, 03, ..., 29, 30

It is also correct to use labels 00 to 29 or even another choice of 30 two-digit labels. Here is the list of clients, with labels attached.

01	A-1 Plumbing	16	JL Records
02	Accent Printing	17	Johnson Commodities
03	Action Sport Shop	18	Keiser Construction
04	Anderson Construction	19	Liu's Chinese Restaurant
05	Bailey Trucking	20	MagicTan
06	Balloons Inc.	21	Peerless Machine
07	Bennett Hardware	22	Photo Arts
08	Best's Camera Shop	23	River City Books
09	Blue Print Specialties	24	Riverside Tavern
10	Central Tree Service	25	Rustic Boutique
11	Classic Flowers	26	Satellite Services
12	Computer Answers	27	Scotch Wash
13	Darlene's Dolls	28	Sewing Center
14	Fleisch Realty	29	Tire Specialties
15	Hernandez Electronics	30	Von's Video Store

Step 2: Table. Enter Table B anywhere and read two-digit groups. Suppose we enter at line 130, which is

69051 64817 87174 09517 84534 06489 87201 97245

The first 10 two-digit groups in this line are

69 05 16 48 17 87 17 40 95 17

Each successive two-digit group is a label. The labels 00 and 31 to 99 are not used in this example, so we ignore them. The first 5 labels between 01 and 30 that we encounter in the table choose our sample. Of the first 10 labels in line 130, we ignore 5 because they are too high (over 30). The others are 05, 16, 17, 17, and 17. The clients labeled 05, 16, and 17 go into the sample. Ignore the second and third 17s because that client is already in the sample. Now run your finger across line 130 (and continue to line 131 if needed) until 5 clients are chosen.

The sample is the clients labeled 05, 16, 17, 20, 19. These are Bailey Trucking, JL Records, Johnson Commodities, MagicTan, and Liu's Chinese Restaurant.

CHOOSING AN SRS

Choose an SRS in two steps:

Step 1: Label. Assign a numerical label to every individual in the population.

Step 2: Table. Use Table B to select labels at random.

You can assign labels in any convenient manner, such as alphabetical order for names of people. Be certain that all labels have the same number of digits. Only then will all individuals have the same chance to be

chosen. Use the shortest possible labels: one digit for a population of up to 10 members, two digits for 11 to 100 members, three digits for 101 to 1000 members, and so on. As standard practice, we recommend that you begin with label 1 (or 01 or 001, as needed). You can read digits from Table B in any order—across a row, down a column, and so on—because the table has no order. As standard practice, we recommend reading across rows.

APPLY YOUR KNOWLEDGE

- 3.7 A firm wants to understand the attitudes of its minority managers toward its system for assessing management performance. Below is a list of all the firm's managers who are members of minority groups. Use Table B at line 139 to choose 6 to be interviewed in detail about the performance appraisal system.

Agarwal	Dewald	Huang	Puri
Anderson	Fernandez	Kim	Richards
Baxter	Fleming	Liao	Rodriguez
Bowman	Garcia	Mourning	Santiago
Brown	Gates	Naber	Shen
Castillo	Goel	Peters	Vega
Cross	Gomez	Pliego	Wang

- 3.8 Thirty individuals in your target audience have been using a new product. Each person has filled out short evaluations of the product periodically during the test period. At the end of the period, you decide to select 4 of the individuals at random for a lengthy interview. The list of participants appears below. Choose an SRS of 4, using the table of random digits beginning at line 145.

Armstrong	Gonzalez	Kemphorne	Robertson
Aspin	Green	Laskowsky	Sanchez
Bennett	Gupta	Liu	Sosa
Bock	Gutierrez	Montoya	Tran
Breiman	Harter	Patnaik	Trevino
Collins	Henderson	Pirelli	Wu
Dixon	Hughes	Rao	
Edwards	Johnson	Rider	

- 3.9 You must choose an SRS of 10 of the 440 retail outlets in New York that sell your company's products. How would you label this population? Use Table B, starting at line 105, to choose your sample.

Stratified samples

The general framework for statistical sampling is a *probability sample*.

PROBABILITY SAMPLE

A **probability sample** is a sample chosen by chance. We must know what samples are possible and what chance, or probability, each possible sample has.

Some probability sampling designs (such as an SRS) give each member of the population an equal chance to be selected. This may not be true in more elaborate sampling designs. In every case, however, the use of chance to select the sample is the essential principle of statistical sampling.

Designs for sampling from large populations spread out over a wide area are usually more complex than an SRS. For example, it is common to sample important groups within the population separately, then combine these samples. This is the idea of a *stratified random sample*.

STRATIFIED RANDOM SAMPLE

To select a **stratified random sample**, first divide the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.

Choose the strata based on facts known before the sample is taken. For example, a population of consumers might be divided into strata according to age, gender, and other demographic information. A stratified design can produce more exact information than an SRS of the same size by taking advantage of the fact that individuals in the same stratum are similar to one another. If all individuals in each stratum are identical, for example, just one individual from each stratum is enough to completely describe the population.

EXAMPLE 3.6

Who wrote that song?

A radio station that broadcasts a piece of music owes a royalty to the composer. The organization of composers (called ASCAP) collects these royalties for all its members by charging stations a license fee for the right to play members' songs. ASCAP has 4 million songs in its catalog and collects \$435 million in fees each year. How should ASCAP distribute this income among its members? By sampling: ASCAP tapes about 60,000 hours from the 53 million hours of local radio programs across the country each year.

Radio stations are stratified by type of community (metropolitan, rural), geographic location (New England, Pacific, etc.), and the size of the license fee paid to ASCAP, which reflects the size of the audience. In all, there are 432 strata. Tapes are made at random hours for randomly selected members of each stratum. The tapes are reviewed by experts who can recognize almost every piece of music ever written, and the composers are then paid according to their popularity.⁵

Multistage samples

Another common type of probability sample chooses the sample in stages. This is usual practice for national samples of households or people. For example, government data on employment and unemployment are gathered by the **Current Population Survey** (Figure 3.1), which conducts interviews in about 55,000 households each month. It is not practical to maintain a list of all U.S. households from which to select an SRS. Moreover, the cost of sending interviewers to the widely scattered households in an SRS would

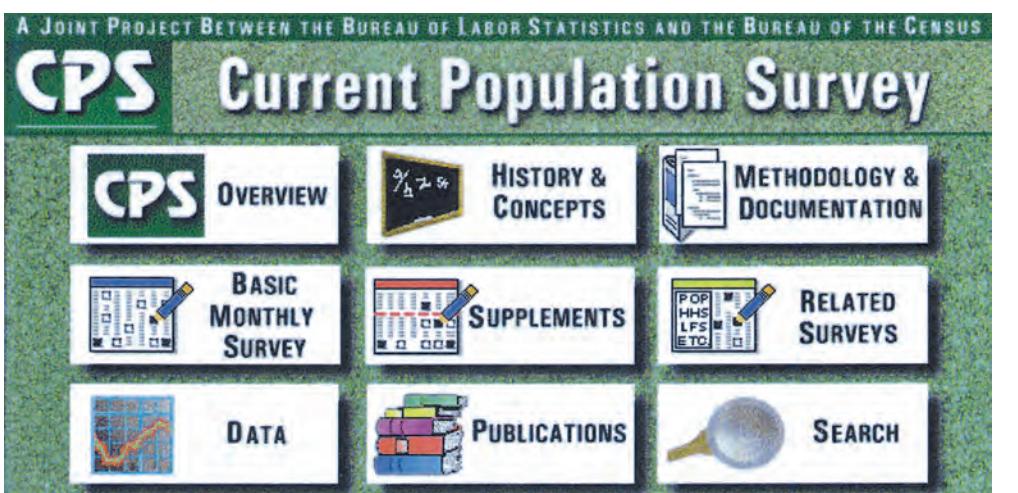


FIGURE 3.1 The monthly Current Population Survey is one of the most important sample surveys in the United States. You can find information and data from the survey on its Web page, www.bls.census.gov/cps.

multistage sample

be too high. The Current Population Survey therefore uses a **multistage sample design**. The final sample consists of clusters of nearby households. Most opinion polls and other national samples are also multistage, though interviewing in most national samples today is done by telephone rather than in person, eliminating the economic need for clustering. The Current Population Survey sampling design is roughly as follows:⁶

Stage 1. Divide the United States into 2007 geographical areas called Primary Sampling Units, or PSUs. Select a sample of 754 PSUs. This sample includes the 428 PSUs with the largest population and a stratified sample of 326 of the others.

Stage 2. Divide each PSU selected at Stage 1 into small areas called “census blocks.” Stratify the blocks using ethnic and other information and take a stratified sample of the blocks in each PSU.

Stage 3. Group the housing units in each block into clusters of four nearby units. Interview the households in a random sample of these clusters.

Analysis of data from sampling designs more complex than an SRS takes us beyond basic statistics. But the SRS is the building block of more elaborate designs, and analysis of other designs differs more in complexity of detail than in fundamental concepts.

- 3.10 Who goes to the workshop? A small advertising firm has 30 junior associates and 10 senior associates. The junior associates are

Abel	Fisher	Huber	Miranda	Reinmann
Chen	Ghosh	Jimenez	Moskowitz	Santos
Cordoba	Griswold	Jones	Neyman	Shaw
David	Hein	Kim	O'Brien	Thompson
Deming	Hernandez	Klotz	Pearl	Utts
Elashoff	Holland	Lorenz	Potter	Warga

APPLY YOUR KNOWLEDGE

The senior associates are

Andrews	Fernandez	Kim	Moore	West
Besicovitch	Gupta	Lightman	Vicario	Yang

The firm will send 4 junior associates and 2 senior associates to a workshop on current trends in market research. It decides to choose those who will go by random selection. Use Table B to choose a stratified random sample of 4 junior associates and 2 senior associates. Start at line 121 to choose your sample.

- 3.11 **Sampling by accountants.** Accountants use stratified samples during audits to verify a company's records of such things as accounts receivable. The stratification is based on the dollar amount of the item and often includes 100% sampling of the largest items. One company reports 5000 accounts receivable. Of these, 100 are in amounts over \$50,000; 500 are in amounts between \$1000 and \$50,000; and the remaining 4400 are in amounts under \$1000. Using these groups as strata, you decide to verify all of the largest accounts and to sample 5% of the midsize accounts and 1% of the small accounts. How would you label the two strata from which you will sample? Use Table B, starting at line 115, to select *only the first 5* accounts from each of these strata.

Cautions about sample surveys

Random selection eliminates bias in the choice of a sample from a list of the population. When the population consists of human beings, however, accurate information from a sample requires much more than a good sampling design.⁷ To begin, we need an accurate and complete list of the population. Because such a list is rarely available, most samples suffer from some degree of *undercoverage*. A sample survey of households, for example, will miss not only homeless people but prison inmates and students in dormitories, too. An opinion poll conducted by telephone will miss the 6% of American households without residential phones. The results of national sample surveys therefore have some bias if the people not covered—who most often are poor people—differ from the rest of the population.

A more serious source of bias in most sample surveys is *nonresponse*, which occurs when a selected individual cannot be contacted or refuses to cooperate. Nonresponse to sample surveys often reaches 50% or more, even with careful planning and several callbacks. Because nonresponse is higher in urban areas, most sample surveys substitute other people in the same area to avoid favoring rural areas in the final sample. If the people contacted differ from those who are rarely at home or who refuse to answer questions, some bias remains.

UNDERCOVERAGE AND NONRESPONSE

Undercoverage occurs when some groups in the population are left out of the process of choosing the sample.

Nonresponse occurs when an individual chosen for the sample can't be contacted or refuses to cooperate.

EXAMPLE 3.7**How bad is nonresponse?**

The Current Population Survey has the lowest nonresponse rate of any poll we know: only about 6% or 7% of the households in the CPS sample don't respond. People are more likely to respond to a government survey such as the CPS, and the CPS contacts its sample in person before doing later interviews by phone.

The General Social Survey, conducted by the University of Chicago's National Opinion Research Center, is the nation's most important social science survey. The GSS also contacts its sample in person, and it is run by a university. Despite these advantages, its most recent survey had a 30% rate of nonresponse.

What about polls done by the media and by market research and opinion-polling firms? We don't know their rates of nonresponse, because they don't say. That itself is a bad sign. The Pew Research Center imitated a careful telephone survey and published the results: out of 2879 households called, 1658 were never at home, refused, or would not finish the interview. That's a nonresponse rate of 58%.⁸

response bias

In addition, the behavior of the respondent or of the interviewer can cause **response bias** in sample results. Respondents may lie, especially if asked about illegal or unpopular behavior. The sample then underestimates the presence of such behavior in the population. An interviewer whose attitude suggests that some answers are more desirable than others will get these answers more often. The race or sex of the interviewer can influence responses to questions about race relations or attitudes toward feminism. Answers to questions that ask respondents to recall past events are often inaccurate because of faulty memory. For example, many people "telescope" events in the past, bringing them forward in memory to more recent time periods. "Have you visited a dentist in the last 6 months?" will often draw a "Yes" from someone who last visited a dentist 8 months ago.⁹ Careful training of interviewers and careful supervision to avoid variation among the interviewers can greatly reduce response bias. Good interviewing technique is another aspect of a well-done sample survey.

wording effects

The **wording of questions** is the most important influence on the answers given to a sample survey. Confusing or leading questions can introduce strong bias, and even minor changes in wording can change a survey's outcome. Here are two examples.

EXAMPLE 3.8**Should we ban disposable diapers?**

A survey paid for by makers of disposable diapers found that 84% of the sample opposed banning disposable diapers. Here is the actual question:

It is estimated that disposable diapers account for less than 2% of the trash in today's landfills. In contrast, beverage containers, third-class mail and yard wastes are estimated to account for about 21% of the trash in landfills. Given this, in your opinion, would it be fair to ban disposable diapers?¹⁰

This question gives information on only one side of an issue, then asks an opinion. That's a sure way to bias the responses. A different question that described how long disposable diapers take to decay and how many tons they contribute to landfills each year would draw a quite different response.

EXAMPLE 3.9**A few words make a big difference**

How do Americans feel about government help for the poor? Only 13% think we are spending too much on “assistance to the poor,” but 44% think we are spending too much on “welfare.” How do the Scots feel about the movement to become independent from England? Well, 51% would vote for “independence for Scotland,” but only 34% support “an independent Scotland separate from the United Kingdom.”¹¹

It seems that “assistance to the poor” and “independence” are nice, hopeful words. “Welfare” and “separate” are negative words. Small changes in how a question is worded can make a big difference in the response.

Never trust the results of a sample survey until you have read the exact questions posed. The amount of nonresponse and the date of the survey are also important. Good statistical design is a part, but only a part, of a trustworthy survey.

APPLY YOUR KNOWLEDGE*sampling frame*

- 3.12 **Random digit dialing.** The list of individuals from which a sample is actually selected is called the **sampling frame**. Ideally, the frame should list every individual in the population, but in practice this is often difficult. A frame that leaves out part of the population is a common source of undercoverage.
 - (a) Suppose that a sample of households in a community is selected at random from the telephone directory. What households are omitted from this frame? What types of people do you think are likely to live in these households? These people will probably be underrepresented in the sample.
 - (b) It is usual in telephone surveys to use random digit dialing equipment that selects the last four digits of a telephone number at random after being given the exchange (the first three digits). Which of the households that you mentioned in your answer to (a) will be included in the sampling frame by random digit dialing?
- 3.13 **Ring-no-answer.** A common form of nonresponse in telephone surveys is “ring-no-answer.” That is, a call is made to an active number, but no one answers. The Italian National Statistical Institute looked at nonresponse to a government survey of households in Italy during two periods, January 1 to Easter and July 1 to August 31. All calls were made between 7 and 10 p.m., but 21.4% gave “ring-no-answer” in one period versus 41.5% “ring-no-answer” in the other period.¹² Which period do you think had the higher rate of no answers? Why? Explain why a high rate of nonresponse makes sample results less reliable.
- 3.14 **Campaign contributions.** Here are two wordings for the same question. The first question was asked by presidential candidate Ross Perot, and the second by a *Time/CNN* poll, both in March 1993.¹³
 - A. Should laws be passed to eliminate all possibilities of special interests giving huge sums of money to candidates?
 - B. Should laws be passed to prohibit interest groups from contributing to campaigns, or do groups have a right to contribute to the candidates they support?

One of these questions drew 40% favoring banning contributions; the other drew 80% with this opinion. Which question produced the 40% and which got 80%? Explain why the results were so different.

BEYOND THE BASICS: CAPTURE-RECAPTURE SAMPLING

Pacific salmon return to reproduce in the river where they were hatched three or four years earlier. How many salmon made it back this year? The answer will help determine quotas for commercial fishing on the west coast of Canada and the United States. Biologists estimate the size of animal populations with a special kind of repeated sampling, called *capture-recapture sampling*. More recently, capture-recapture methods have been used on human populations as well.

EXAMPLE 3.10

Counting salmon

The old method of counting returning salmon involved placing a “counting fence” in a stream and counting all the fish caught by the fence. This is expensive and difficult. For example, fences are often damaged by high water. Sampling using small nets is more practical.¹⁴

During this year’s spawning run in the Chase River in British Columbia, Canada, you net 200 coho salmon, tag the fish, and release them. Later in the week, your nets capture 120 coho salmon in the river, of which 12 have tags.

The proportion of your second sample that have tags should estimate the proportion in the entire population of returning salmon that are tagged. So if N is the unknown number of coho salmon in the Chase River this year, we should have approximately

$$\text{proportion tagged in sample} = \text{proportion tagged in population}$$

$$\frac{12}{120} = \frac{200}{N}$$

Solve for N to estimate that the total number of salmon in this year’s spawning run in the Chase River is approximately

$$N = 200 \times \frac{120}{12} = 2000$$

The capture-recapture idea employs of a sample proportion to estimate a population proportion. The idea works well if both samples are SRSs from the population and the population remains unchanged between samples. In practice, complications arise. For example, some tagged fish might be caught by bears or otherwise die between the first and second samples. Variations on capture-recapture samples are widely used in wildlife studies and are now finding other applications. One way to estimate the census undercount in a district is to consider the census as “capturing and marking” the households that respond. Census workers then visit the district, take an SRS of households, and see how many of those counted by the census show up in the sample. Capture-recapture estimates the total count of households in the district. As with estimating wildlife populations,

there are many practical pitfalls. Our final word is as before: the real world is less orderly than statistics textbooks imply.

SECTION 3.1 SUMMARY

- We can produce data intended to answer specific questions by **observational studies or experiments**. **Sample surveys** that select a part of a population of interest to represent the whole are one type of observational study. **Experiments**, unlike observational studies, actively impose some treatment on the subjects of the experiment.
- A sample survey selects a **sample** from the **population** of all individuals about which we desire information. We base conclusions about the population on data about the sample.
- The **design** of a sample refers to the method used to select the sample from the population. **Probability sampling** designs use random selection to select the sample, thus avoiding bias due to personal choice.
- The basic probability sample is a **simple random sample (SRS)**. An SRS gives every possible sample of a given size the same chance to be chosen.
- Choose an SRS by labeling the members of the population and using a **table of random digits** to select the sample. Software can automate this process.
- To choose a **stratified random sample**, divide the population into **strata**, groups of individuals that are similar in some way that is important to the response. Then choose a separate SRS from each stratum.
- Failure to use probability sampling often results in **bias**, or systematic errors in the way the sample represents the population. **Voluntary response samples**, in which the respondents choose themselves, are particularly prone to large bias.
- In human populations, even probability samples can suffer from bias due to **undercoverage** or **nonresponse**, from **response bias**, or from misleading results due to **poorly worded questions**. Sample surveys must deal expertly with these potential problems in addition to using a probability sampling design.

SECTION 3.1 EXERCISES

- 3.15 What is the population? For each of the following sampling situations, identify the population as exactly as possible. That is, say what kind of individuals the population consists of and say exactly which individuals fall in the population. If the information given is not sufficient, complete the description of the population in a reasonable way.
- (a) A business school researcher wants to know what factors affect the survival and success of small businesses. She selects a sample of 150 eating-and-drinking establishments from those listed in the telephone directory Yellow Pages for a large city.

- (b) A member of Congress wants to know whether his constituents support proposed legislation on health care. His staff reports that 228 letters have been received on the subject, of which 193 oppose the legislation.
- (c) An insurance company wants to monitor the quality of its procedures for handling loss claims from its auto insurance policyholders. Each month the company examines in detail an SRS from all auto insurance claims filed that month.
- 3.16 **Instant opinion.** The Excite Poll can be found online at poll.excite.com. The question appears on the screen, and you simply click buttons to vote “Yes,” “No,” or “Not sure.” On January 25, 2000, the question was “Should female athletes be paid the same as men for the work they do?” In all, 13,147 respondents (44%) said “Yes,” another 15,182 (50%) said “No,” and the remaining 1448 said “Don’t know.”
- (a) What is the sample size for this poll?
- (b) That’s a much larger sample than standard sample surveys. In spite of this, we can’t trust the result to give good information about any clearly defined population. Why?
- (c) It is still true that more men than women use the Web. How might this fact affect the poll results?
- 3.17 **Mail to Congress.** You are on the staff of a member of Congress who is considering a bill that would require all employers to provide health insurance for their employees. You report that 1128 letters dealing with the issue have been received, of which 871 oppose the legislation. “I’m surprised that most of my constituents oppose the bill. I thought it would be quite popular,” says the congresswoman. Are you convinced that a majority of the voters opposes the bill? State briefly how you would explain the statistical issue to the congresswoman.
- 3.18 **Design your own bad sample.** Your college wants to gather student opinion about parking for students on campus. It isn’t practical to contact all students.
- (a) Give an example of a way to choose a sample of students that is poor practice because it depends on voluntary response.
- (b) Give another example of a bad way to choose a sample that doesn’t use voluntary response.
- 3.19 **Quality control sampling.** A manufacturer of chemicals chooses 3 containers from each lot of 25 containers of a reagent to test for purity and potency. Below are the control numbers stamped on the bottles in the current lot. Use Table B at line 111 to choose an SRS of 3 of these bottles.
- | | | | | |
|-------|-------|-------|-------|-------|
| A1096 | A1097 | A1098 | A1101 | A1108 |
| A1112 | A1113 | A1117 | A2109 | A2211 |
| A2220 | B0986 | B1011 | B1096 | B1101 |
| B1102 | B1103 | B1110 | B1119 | B1137 |
| B1189 | B1223 | B1277 | B1286 | B1299 |
- 3.20 **Apartment living.** You are planning a report on apartment living in a college town. You decide to select three apartment complexes at random for in-depth

interviews with residents. Use Table B, starting at line 117, to select a simple random sample of three of the following apartment complexes.

Ashley Oaks	Country View	Mayfair Village
Bay Pointe	Country Villa	Nobb Hill
Beau Jardin	Crestview	Pemberly Courts
Bluffs	Del-Lynn	Peppermill
Brandon Place	Fairington	Pheasant Run
Briarwood	Fairway Knolls	Richfield
Brownstone	Fowler	Sagamore Ridge
Burberry	Franklin Park	Salem Courthouse
Cambridge	Georgetown	Village Manor
Chauncey Village	Greenacres	Waterford Court
Country Squire	Lahr House	Williamsburg

- 3.21 Sampling from a census tract. The Census Bureau divides the entire country into “census tracts” that contain about 4000 people. Each tract is in turn divided into small “blocks,” which in urban areas are bounded by local streets. An SRS of blocks from a census tract is often the next-to-last stage in a multistage sample. Figure 3.2 shows part of census tract 8051.12, in Cook County, Illinois, west of Chicago. The 44 blocks in this tract are divided into three “block groups.” Group 1 contains 6 blocks numbered 1000 to 1005; group 2 (outlined in Figure 3.2) contains 12 blocks numbered 2000 to 2011; group 3 contains 26 blocks numbered 3000 to 3025. Use Table B, beginning at line 125, to choose an SRS of 5 of the 44 blocks in this census tract. Explain carefully how you labeled the blocks.

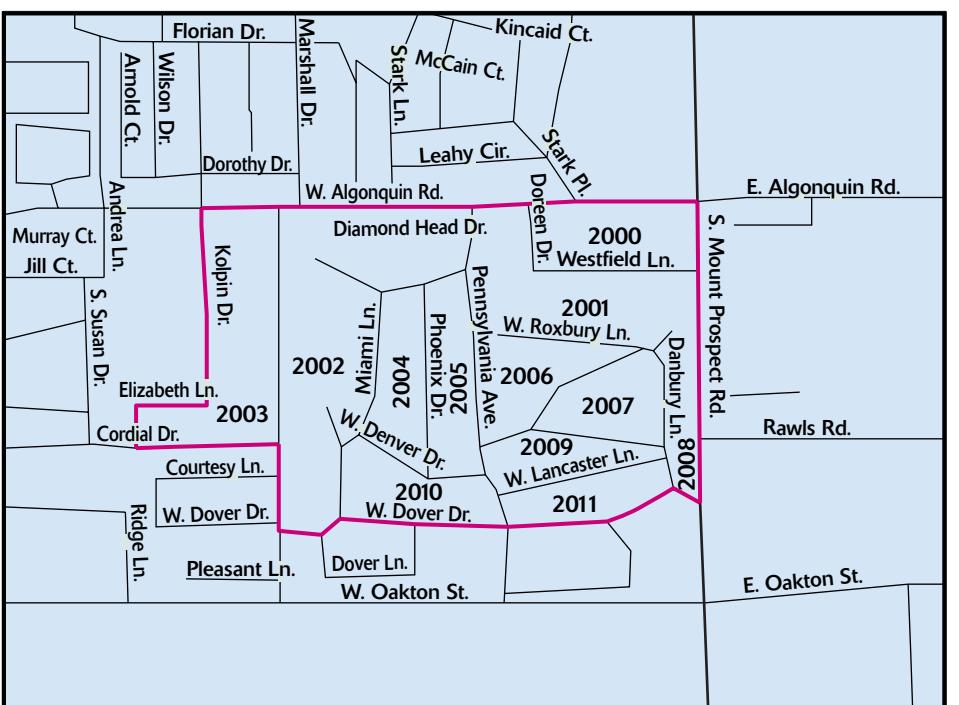


FIGURE 3.2 Census blocks in Cook County, Illinois. The outlined area is a block group. (From factfinder.census.gov.)

systematic random sample

- 3.22 **Random digits.** Which of the following statements are true of a table of random digits, and which are false? Briefly explain your answers.
- There are exactly four 0s in each row of 40 digits.
 - Each pair of digits has chance 1/100 of being 00.
 - The digits 0000 can never appear as a group, because this pattern is not random.
- 3.23 **Systematic random samples.** The last stage of the Current Population Survey chooses clusters of households within small areas called blocks. The method used is **systematic random sampling**. An example will illustrate the idea of a systematic sample. Suppose that we must choose 4 clusters out of 100. Because $100/4 = 25$, we can think of the list as four lists of 25 clusters. Choose 1 of the first 25 at random, using Table B. The sample will contain this cluster and the clusters 25, 50, and 75 places down the list from it. If 13 is chosen, for example, then the systematic random sample consists of the clusters numbered 13, 38, 63, and 88.
- Use Table B to choose a systematic random sample of 5 clusters from a list of 200. Enter the table at line 120.
 - Like an SRS, a systematic sample gives all individuals the same chance to be chosen. Explain why this is true, then explain carefully why a systematic sample is nonetheless *not* an SRS.
- 3.24 **Is this an SRS?** A company employs 2000 male and 500 female engineers. A stratified random sample of 50 female and 200 male engineers gives each engineer one chance in 10 to be chosen. This sample design gives every individual in the population the same chance to be chosen for the sample. Is it an SRS? Explain your answer.
- 3.25 **A stratified sample.** A company employs 2000 male and 500 female engineers. The human resources department wants to poll the opinions of a random sample of engineers about the company's performance review system. To give adequate attention to female opinion, you will choose a stratified random sample of 200 males and 200 females. You have alphabetized lists of female and male engineers. Explain how you would assign labels and use random digits to choose the desired sample. Enter Table B at line 122 and give the labels of the first 5 females and the first 5 males in the sample.
- 3.26 **Wording survey questions.** Comment on each of the following as a potential sample survey question. Is the question clear? Is it slanted toward a desired response?
- Some cell phone users have developed brain cancer. Should all cell phones come with a warning label explaining the danger of using cell phones?
 - In view of escalating environmental degradation and incipient resource depletion, would you favor economic incentives for recycling of resource-intensive consumer goods?
- 3.27 **Bad survey questions.** Write your own examples of bad sample survey questions.
- Write a biased question designed to get one answer rather than another.
 - Write a question that is confusing, so that it is hard to answer.

- 3.28 **Do the people want a tax cut?** During the 2000 presidential campaign, the candidates debated what to do with the large government surplus. The Pew Research Center asked two questions of random samples of adults. Both said that Social Security would be “fixed.” Here are the uses suggested for the remaining surplus:

Should the money be used for a tax cut, or should it be used to fund new government programs?

Should the money be used for a tax cut, or should it be spent on programs for education, the environment, health care, crime-fighting and military defense?

One of these questions drew 60% favoring a tax cut; the other, only 22%. Which wording pulls respondents toward a tax cut? Why?

3.2 Designing Experiments

A study is an experiment when we actually do something to people, animals, or objects to observe the response. Experiments are important tools for product development and improvement of processes and services. Our short introduction here concentrates on basic ideas at a level suitable for managers. Statistically designed experiments are at the core of some large industries, especially pharmaceuticals, where regulations require that new drugs be shown by experiments to be safe and reliable before they can be sold.

Because the purpose of an experiment is to reveal the response of one variable to changes in other variables, the distinction between explanatory and response variables is essential. Here is some basic vocabulary for experiments.

SUBJECTS, FACTORS, TREATMENTS

The individuals studied in an experiment are often called **subjects**, especially if they are people.

The explanatory variables in an experiment are often called **factors**.

A **treatment** is any specific experimental condition applied to the subjects. If an experiment has several factors, a treatment is a combination of a specific value (often called a **level**) of each of the factors.

EXAMPLE 3.11

Absorption of a drug

Researchers at a pharmaceutical company studying the absorption of a drug into the bloodstream inject the drug (the treatment) into 25 people (the subjects). The response variable is the concentration of the drug in a subject’s blood, measured 30 minutes after the injection. This experiment has a single factor with only one level. If three different doses of the drug are injected, there is still a single factor (the dosage of the drug), now with three levels. The three levels of the single factor are the treatments that the experiment compares.

		Factor B Repetitions		
		1 time	3 times	5 times
Factor A Length	30 seconds	1	2	3
	90 seconds	4	5	6

FIGURE 3.3 The treatments in the experimental design of Example 3.12. Combinations of levels of the two factors form six treatments.

EXAMPLE 3.12

Effects of TV advertising

What are the effects of repeated exposure to an advertising message? The answer may depend both on the length of the ad and on how often it is repeated. An experiment investigates this question using undergraduate students as subjects. All subjects view a 40-minute television program that includes ads for a digital camera. Some subjects see a 30-second commercial; others, a 90-second version. The same commercial is repeated either 1, 3, or 5 times during the program. After viewing, all of the subjects answer questions about their recall of the ad, their attitude toward the camera, and their intention to purchase it. These are the response variables.¹⁵

This experiment has two factors: length of the commercial, with 2 levels; and repetitions, with 3 levels. The 6 combinations of one level of each factor form 6 treatments. Figure 3.3 shows the layout of the treatments.

Examples 3.11 and 3.12 illustrate the advantages of experiments over observational studies. Experimentation allows us to study the effects of the specific treatments we are interested in. Moreover, we can control the environment of the subjects to hold constant the factors that are of no interest to us, such as the specific product advertised in Example 3.12. The ideal case is a laboratory experiment in which we control all lurking variables and so see only the effect of the treatments on the response. Like most ideals, such control is not always realized in practice.

Another advantage of experiments is that we can study the combined effects of several factors simultaneously. The **interaction** of several factors can produce effects that could not be predicted from looking at the effect of each factor alone. Perhaps longer commercials increase interest in a product, and more commercials also increase interest, but if we both make a commercial longer and show it more often, viewers get annoyed and their interest in the product drops. The two-factor experiment in Example 3.12 will help us find out.

- 3.29 **Sickle cell disease.** Sickle cell disease is an inherited disorder of the red blood cells that in the United States affects mostly blacks. It can cause severe

APPLY YOUR KNOWLEDGE

pain and many complications. Bristol-Myers Squibb markets Hydrea—a brand name for the drug hydroxyurea—to treat sickle cell disease. Federal regulations allow the sale of a new drug only after statistically designed experiments (called clinical trials in medical language) show that the drug is safe and effective. A clinical trial at the National Institutes of Health gave hydroxyurea to 150 sickle cell sufferers and a placebo (a dummy medication) to another 150. The researchers then counted the episodes of pain reported by each subject. What are the subjects, the factors, the treatments, and the response variables?

- 3.30 **Sealing food packages.** A manufacturer of food products uses package liners that are sealed at the top by applying heated jaws after the package is filled. The customer peels the sealed pieces apart to open the package. What effect does the temperature of the jaws have on the force needed to peel the liner? To answer this question, engineers prepare 20 pairs of pieces of package liner. They seal five pairs at each of 250° F, 275° F, 300° F, and 325° F. Then they measure the force needed to peel each seal.
- What are the individuals studied?
 - There is one factor (explanatory variable). What is it, and what are its levels?
 - What is the response variable?
- 3.31 **An industrial experiment.** A chemical engineer is designing the production process for a new product. The chemical reaction that produces the product may have higher or lower yield, depending on the temperature and the stirring rate in the vessel in which the reaction takes place. The engineer decides to investigate the effects of combinations of two temperatures (50° C and 60° C) and three stirring rates (60 rpm, 90 rpm, and 120 rpm) on the yield of the process. She will process two batches of the product at each combination of temperature and stirring rate.
- What are the individuals and the response variable in this experiment?
 - How many factors are there? How many treatments? Use a diagram like that in Figure 3.3 to lay out the treatments.
 - How many individuals are required for the experiment?

Comparative experiments

Experiments in the laboratory often have a simple design: impose the treatment and see what happens. We can outline that design like this:

Subjects → Treatment → Response

In the laboratory, we try to avoid confounding by rigorously controlling the environment of the experiment so that nothing except the experimental treatment influences the response. Once we get out of the laboratory, however, there are almost always lurking variables waiting to confound us. When our subjects are people or animals rather than electrons or chemical compounds, confounding can happen even in the controlled environment of a laboratory or medical clinic. Here is an example that helps explain why careful experimental design is a key issue for pharmaceutical companies and other makers of medical products.

EXAMPLE 3.13**Gastric freezing to treat ulcers**

“Gastric freezing” is a clever treatment for ulcers. The patient swallows a deflated balloon with tubes attached, then a refrigerated liquid is pumped through the balloon for an hour. The idea is that cooling the stomach will reduce its production of acid and so relieve ulcers. An experiment reported in the *Journal of the American Medical Association* showed that gastric freezing did reduce acid production and relieve ulcer pain. The treatment was widely used for several years. The design of the experiment was

Subjects → Gastric freezing → Observe pain relief

placebo effect

This experiment is poorly designed. The patients’ response may be due to the placebo effect. A placebo is a dummy treatment. Many patients respond favorably to *any* treatment, even a placebo, presumably because of trust in the doctor and expectations of a cure. This response to a dummy treatment is the placebo effect.

A later experiment divided ulcer patients into two groups. One group was treated by gastric freezing as before. The other group received a placebo treatment in which the liquid in the balloon was at body temperature rather than freezing. The results: 34% of the 82 patients in the treatment group improved, but so did 38% of the 78 patients in the placebo group. This and other properly designed experiments showed that gastric freezing was no better than a placebo, and its use was abandoned.¹⁶

**control group**

The first gastric-freezing experiment was *biased*. It systematically favored gastric freezing because the placebo effect was confounded with the effect of the treatment. Fortunately, the remedy is simple. Experiments should *compare* treatments rather than attempt to assess a single treatment in isolation. When we compare the two groups of patients in the second gastric-freezing experiment, the placebo effect and other lurking variables operate on both groups. The only difference between the groups is the actual effect of gastric freezing. The group of patients who receive a sham treatment is called a **control group**, because it enables us to control the effects of lurking variables on the outcome.

Randomized comparative experiments

The design of an experiment first describes the response variables, the factors (explanatory variables), and the layout of the treatments, with *comparison* as the leading principle. The second aspect of design is the rule used to assign the subjects to the treatments. Comparison of the effects of several treatments is valid only when all treatments are applied to similar groups of subjects. If one corn variety is planted on more fertile ground, or if one cancer drug is given to less seriously ill patients, comparisons among treatments are biased. How can we assign individuals to treatments in a way that is fair to all the treatments?

randomization

Our answer is the same as in sampling: let impersonal chance make the assignment. The use of chance to divide subjects into groups is called randomization. Groups formed by randomization don’t depend on any

*randomized
comparative
experiment*

characteristic of the subjects or on the judgment of the experimenter. An experiment that uses both comparison and randomization is a **randomized comparative experiment**. Here is an example.

EXAMPLE 3.14

Testing a breakfast food

A food company assesses the nutritional quality of a new “instant breakfast” product by feeding it to newly weaned male white rats. The response variable is a rat’s weight gain over a 28-day period. A control group of rats eats a standard diet but otherwise receives exactly the same treatment as the experimental group.

This experiment has one factor (the diet) with two levels. The researchers use 30 rats for the experiment and so must divide them into two groups of 15. To do this in an unbiased fashion, put the cage numbers of the 30 rats in a hat, mix them up, and draw 15. These rats form the experimental group and the remaining 15 make up the control group. *Each group is an SRS of the available rats.* Figure 3.4 outlines the design of this experiment.

We can use software or the table of random digits to randomize. Label the rats 01 to 30. Enter Table B at (say) line 130. Run your finger along this line (and continue to lines 131 and 132 as needed) until 15 rats are chosen. They are the rats labeled

05 16 17 20 19 04 25 29 18 07 13 02 23 27 21

These rats form the experimental group; the remaining 15 are the control group.

Completely randomized designs

The design in Figure 3.4 combines comparison and randomization to arrive at the simplest statistical design for an experiment. This “flowchart” outline presents all the essentials: randomization, the sizes of the groups and which treatment they receive, and the response variable. There are, as we will see later, statistical reasons for generally using treatment groups that are about equal in size. We call designs like that in Figure 3.4 *completely randomized*.

COMPLETELY RANDOMIZED DESIGN

In a completely randomized experimental design, all the subjects are allocated at random among all the treatments.



FIGURE 3.4 Outline of a randomized comparative experiment, for Example 3.14.

Completely randomized designs can compare any number of treatments. Here is an example that compares three treatments.

EXAMPLE 3.15

Conserving energy

Many utility companies have introduced programs to encourage energy conservation among their customers. An electric company considers placing electronic indicators in households to show what the cost would be if the electricity use at that moment continued for a month. Will indicators reduce electricity use? Would cheaper methods work almost as well? The company decides to design an experiment.

One cheaper approach is to give customers a chart and information about monitoring their electricity use. The experiment compares these two approaches (indicator, chart) and also a control. The control group of customers receives information about energy conservation but no help in monitoring electricity use. The response variable is total electricity used in a year. The company finds 60 single-family residences in the same city willing to participate, so it assigns 20 residences at random to each of the 3 treatments. Figure 3.5 outlines the design.

To carry out the random assignment, label the 60 households 01 to 60. Enter Table B (or use software) to select an SRS of 20 to receive the indicators. Continue in Table B, selecting 20 more to receive charts. The remaining 20 form the control group.

Examples 3.14 and 3.15 describe completely randomized designs that compare levels of a single factor. In Example 3.14, the factor is the diet fed to the rats. In Example 3.15, it is the method used to encourage energy conservation. Completely randomized designs can have more than one factor. The advertising experiment of Example 3.12 has two factors: the length and the number of repetitions of a television commercial. Their combinations form the six treatments outlined in Figure 3.3 (page 190). A completely randomized design assigns subjects at random to these six treatments. Once the layout of treatments is set, the randomization needed for a completely randomized design is tedious but straightforward.

- 3.32 Gastric freezing.** Example 3.13 describes an experiment that helped end the use of gastric freezing to treat ulcers. The subjects were 160 ulcer patients.
 (a) Use a diagram to outline the design of this experiment, following the information in Example 3.13. (Show the size of the groups, the treatment each group receives, and the response variable. Figures 3.4 and 3.5 are models to follow.)

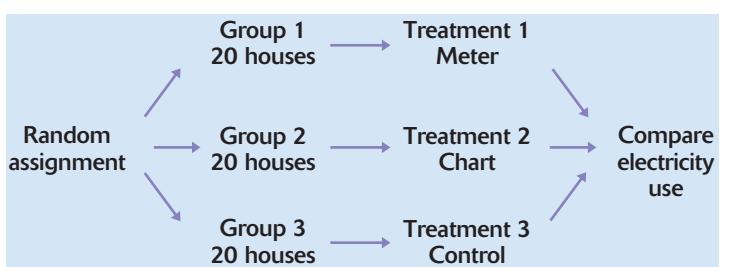


FIGURE 3.5 Outline of a completely randomized design comparing three treatments, for Example 3.15.

- (b) The 82 patients in the gastric-freezing group are an SRS of the 160 subjects. Label the subjects and use Table B, starting at line 131, to choose the first 5 members of the gastric-freezing group.
- 3.33 **Sealing food packages.** Use a diagram to describe a completely randomized experimental design for the package liner experiment of Exercise 3.30. (Show the size of the groups, the treatment each group receives, and the response variable. Figures 3.4 and 3.5 are models to follow.) Use software or Table B, starting at line 120, to do the randomization required by your design.
- 3.34 **Does child care help recruit employees?** Will providing child care for employees make a company more attractive to women, even those who are unmarried? You are designing an experiment to answer this question. You prepare recruiting material for two fictitious companies, both in similar businesses in the same location. Company A's brochure does not mention child care. There are two versions of Company B's material, identical except that one describes the company's on-site child-care facility. Your subjects are 40 unmarried women who are college seniors seeking employment. Each subject will read recruiting material for both companies and choose the one she would prefer to work for. You will give each version of Company B's brochure to half the women. You expect that a higher percentage of those who read the description that includes child care will choose Company B.
- Outline an appropriate design for the experiment.
 - The names of the subjects appear below. Use Table B, beginning at line 131, to do the randomization required by your design. List the subjects who will read the version that mentions child care.

Abrams	Danielson	Gutierrez	Lippman	Rosen
Adamson	Durr	Howard	Martinez	Sugiyara
Affi	Edwards	Hwang	McNeill	Thompson
Brown	Fluharty	Iselin	Morse	Travers
Cansico	Garcia	Janle	Ng	Turing
Chen	Gerson	Kaplan	Quinones	Ullmann
Cortez	Green	Kim	Rivera	Williams
Curzakis	Gupta	Lattimore	Roberts	Wong

The logic of randomized comparative experiments

Randomized comparative experiments are designed to give good evidence that differences in the treatments actually *cause* the differences we see in the response. The logic is as follows:

- Random assignment of subjects forms groups that should be similar in all respects before the treatments are applied.
- Comparative design ensures that influences other than the experimental treatments operate equally on all groups.
- Therefore, differences in average response must be due either to the treatments or to the play of chance in the random assignment of subjects to the treatments.

That “either-or” deserves more thought. In Example 3.14, we cannot say that *any* difference in the average weight gains of rats fed the two diets must be caused by a difference between the diets. There would be some difference even if both groups received the same diet, because the natural variability among rats means that some grow faster than others. Chance assigns the faster-growing rats to one group or the other, and this creates a chance difference between the groups. We would not trust an experiment with just one rat in each group, for example. The results would depend on which group got lucky and received the faster-growing rat. If we assign many rats to each diet, however, the effects of chance will average out and there will be little difference in the average weight gains in the two groups unless the diets themselves cause a difference. “Use enough subjects to reduce chance variation” is the third big idea of statistical design of experiments.

PRINCIPLES OF EXPERIMENTAL DESIGN

1. Control the effects of lurking variables on the response, most simply by comparing two or more treatments.
2. Randomize—use impersonal chance to assign subjects to treatments.
3. Replicate each treatment on enough subjects to reduce chance variation in the results.

We hope to see a difference in the responses so large that it is unlikely to happen just because of chance variation. We can use the laws of probability, which give a mathematical description of chance behavior, to learn if the treatment effects are larger than we would expect to see if only chance were operating. If they are, we call them *statistically significant*.

STATISTICAL SIGNIFICANCE

An observed effect so large that it would rarely occur by chance is called **statistically significant**.

If we observe statistically significant differences among the groups in a comparative randomized experiment, we have good evidence that the treatments actually caused these differences. You will often see the phrase “statistically significant” in reports of investigations in many fields of study. The great advantage of randomized comparative experiments is that they can produce data that give good evidence for a cause-and-effect relationship between the explanatory and response variables. We know that in general a strong association does not imply causation. A statistically significant association in data from a well-designed experiment does imply causation.

- 3.35 **Conserving energy.** Example 3.15 describes an experiment to learn whether providing households with electronic indicators or charts will reduce their electricity consumption. An executive of the electric company objects to

including a control group. He says, “It would be simpler to just compare electricity use last year (before the indicator or chart was provided) with consumption in the same period this year. If households use less electricity this year, the indicator or chart must be working.” Explain clearly why this design is inferior to that in Example 3.15.

- 3.36 **Exercise and heart attacks.** Does regular exercise reduce the risk of a heart attack? Here are two ways to study this question. Explain clearly why the second design will produce more trustworthy data.
1. A researcher finds 2000 men over 40 who exercise regularly and have not had heart attacks. She matches each with a similar man who does not exercise regularly, and she follows both groups for 5 years.
 2. Another researcher finds 4000 men over 40 who have not had heart attacks and are willing to participate in a study. She assigns 2000 of the men to a regular program of supervised exercise. The other 2000 continue their usual habits. The researcher follows both groups for 5 years.
- 3.37 **Statistical significance.** The financial aid office of a university asks a sample of students about their employment and earnings. The report says that “for academic year earnings, a significant difference was found between the sexes, with men earning more on the average. No significant difference was found between the earnings of black and white students.” Explain the meaning of “a significant difference” and “no significant difference” in plain language.

Cautions about experimentation

The logic of a randomized comparative experiment depends on our ability to treat all the subjects identically in every way except for the actual treatments being compared. Good experiments therefore require careful attention to details. For example, the subjects in both groups of the second gastric freezing experiment (Example 3.13, page 192) all got the same medical attention over the several years of the study. The researchers paid attention to such details as ensuring that the tube in the mouth of each subject was cold, whether or not the fluid in the balloon was refrigerated. Moreover, the study was **double-blind**—neither the subjects themselves nor the medical personnel who worked with them knew which treatment any subject had received. The double-blind method avoids unconscious bias by, for example, a doctor who doesn’t think that “just a placebo” can benefit a patient.

double-blind

lack of realism

The most serious potential weakness of experiments is **lack of realism**. The subjects or treatments or setting of an experiment may not realistically duplicate the conditions we really want to study. Here are two examples.

EXAMPLE 3.16

Response to advertising

The study of television advertising in Example 3.12 showed a 40-minute videotape to students who knew an experiment was going on. We can’t be sure that the results apply to everyday television viewers. The student subjects described their reactions but did not actually decide whether to buy the camera. Many experiments in marketing and decision making use as subjects students who know they are taking part in an experiment. That’s not a realistic setting.

EXAMPLE 3.17**Center brake lights**

Do those high center brake lights, required on all cars sold in the United States since 1986, really reduce rear-end collisions? Randomized comparative experiments with fleets of rental and business cars, done before the lights were required, showed that the third brake light reduced rear-end collisions by as much as 50%. Alas, requiring the third light in all cars led to only a 5% drop.

What happened? Most cars did not have the extra brake light when the experiments were carried out, so it caught the eye of following drivers. Now that almost all cars have the third light, they no longer capture attention.

Lack of realism can limit our ability to apply the conclusions of an experiment to the settings of greatest interest. Most experimenters want to generalize their conclusions to some setting wider than that of the actual experiment. Statistical analysis of the original experiment cannot tell us how far the results will generalize. Nonetheless, the randomized comparative experiment, because of its ability to give convincing evidence for causation, is one of the most important ideas in statistics.

APPLY YOUR KNOWLEDGE

- 3.38 **Does meditation reduce anxiety?** Some companies employ consultants to train their managers in meditation in the hope that this practice will relieve stress and make the managers more effective on the job. An experiment that claimed to show that meditation reduces anxiety proceeded as follows. The experimenter interviewed the subjects and rated their level of anxiety. Then the subjects were randomly assigned to two groups. The experimenter taught one group how to meditate and they meditated daily for a month. The other group was simply told to relax more. At the end of the month, the experimenter interviewed all the subjects again and rated their anxiety level. The meditation group now had less anxiety. Psychologists said that the results were suspect because the ratings were not blind. Explain what this means and how lack of blindness could bias the reported results.
- 3.39 **Frustration and teamwork.** A psychologist wants to study the effects of failure and frustration on the relationships among members of a work team. She forms a team of students, brings them to the psychology laboratory, and has them play a game that requires teamwork. The game is rigged so that they lose regularly. The psychologist observes the students through a one-way window and notes the changes in their behavior during an evening of game playing. Why is it doubtful that the findings of this study tell us much about the effect of working for months developing a new product that never works right and is finally abandoned by your company?

Matched pairs designs

Completely randomized designs are the simplest statistical designs for experiments. They illustrate clearly the principles of control, randomization, and replication of treatments on a number of subjects. However, completely randomized designs are often inferior to more elaborate statistical designs. In particular, matching the subjects in various ways can produce more precise results than simple randomization.

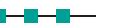
matched pairs design

One common design that combines matching with randomization is the **matched pairs design**. A matched pairs design compares just two treatments. Choose pairs of subjects that are as closely matched as possible. Assign one of the treatments to each subject in a pair by tossing a coin or reading odd and even digits from Table B. Sometimes each “pair” in a matched pairs design consists of just one subject, who gets both treatments one after the other. Each subject serves as his or her own control. The *order* of the treatments can influence the subject’s response, so we randomize the order for each subject, again by a coin toss.

EXAMPLE 3.18**Coke versus Pepsi**

Pepsi wanted to demonstrate that Coke drinkers prefer Pepsi when they taste both colas blind. The subjects, all people who said they were Coke drinkers, tasted both colas from glasses without brand markings and said which they liked better. This is a matched pairs design in which each subject compares the two colas. Because responses may depend on which cola is tasted first, the order of tasting should be chosen at random for each subject.

When more than half the Coke drinkers chose Pepsi, Coke claimed that the experiment was biased. The Pepsi glasses were marked *M* and the Coke glasses were marked *Q*. Aha, said Coke, this just shows that people like the letter *M* better than the letter *Q*. A careful experiment would in fact take care to avoid any distinction other than the actual treatments.¹⁷

**Block designs**

Matched pairs designs apply the principles of comparison of treatments, randomization, and replication. However, the randomization is not complete—we do not randomly assign all the subjects at once to the two treatments. Instead, we only randomize within each matched pair. This allows matching to reduce the effect of variation among the subjects. Matched pairs are an example of *block designs*.

BLOCK DESIGN

A **block** is a group of subjects that are known before the experiment to be similar in some way expected to affect the response to the treatments. In a **block design**, the random assignment of individuals to treatments is carried out separately within each block.

A block design combines the idea of creating equivalent treatment groups by matching with the principle of forming treatment groups at random. Blocks are another form of *control*. They control the effects of some outside variables by bringing those variables into the experiment to form the blocks. The following is a typical example of a block design.

EXAMPLE 3.19**Men, women, and advertising**

Women and men respond differently to advertising. An experiment to compare the effectiveness of three television commercials for the same product will want to look separately at the reactions of men and women, as well as assess the overall response to the ads.

A completely randomized design considers all subjects, both men and women, as a single pool. The randomization assigns subjects to three treatment groups without regard to their sex. This ignores the differences between men and women. A better design considers women and men separately. Randomly assign the women to three groups, one to view each commercial. Then separately assign the men at random to three groups. Figure 3.6 outlines this improved design.

A block is a group of subjects formed before an experiment starts. We reserve the word “treatment” for a condition that we impose on the subjects. We don’t speak of 6 treatments in Example 3.19 even though we can compare the responses of 6 groups of subjects formed by the 2 blocks (men, women) and the 3 commercials. Block designs are similar to stratified samples. Blocks and strata both group similar individuals together. We use two different names only because the idea developed separately for sampling and experiments. The advantages of block designs are the same as the advantages of stratified samples. Blocks allow us to draw separate conclusions about each block—for example, about men and women in the advertising study in Example 3.19. Blocking also allows more precise overall conclusions because the systematic differences between men and women can be removed when we study the overall effects of the three commercials. The idea of blocking is an important additional principle of statistical design of experiments. A wise experimenter will form blocks based on the most important unavoidable sources of variability among the experimental

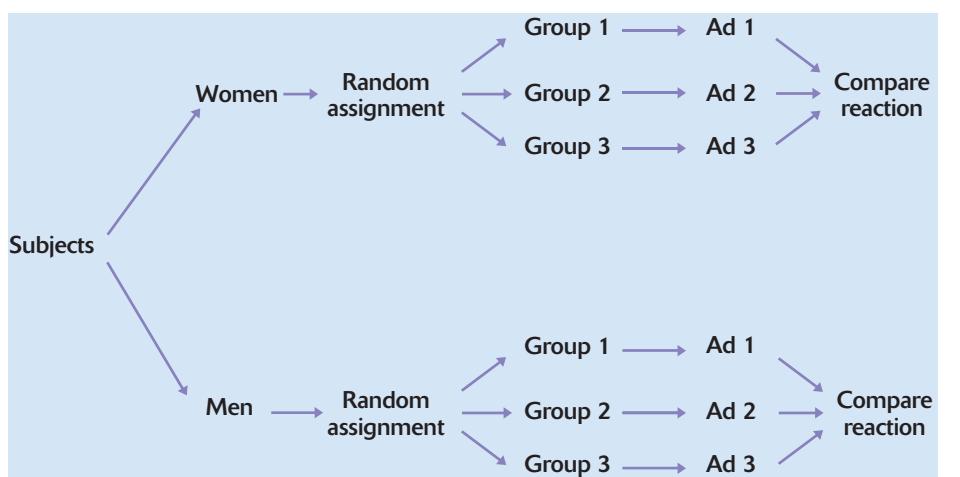


FIGURE 3.6 Outline of a block design, for Example 3.19. The blocks consist of male and female subjects. The treatments are three television commercials.

subjects. Randomization will then average out the effects of the remaining variation and allow an unbiased comparison of the treatments.

Like the design of samples, the design of complex experiments is a job for experts. Now that we have seen a bit of what is involved, we will usually just act as if most experiments were completely randomized.

APPLY YOUR KNOWLEDGE

- 3.40 **Does charting help investors?** Some investment advisors believe that charts of past trends in the prices of securities can help predict future prices. Most economists disagree. In an experiment to examine the effects of using charts, business students trade (hypothetically) a foreign currency at computer screens. There are 20 student subjects available, named for convenience A, B, C, ..., T. Their goal is to make as much money as possible, and the best performances are rewarded with small prizes. The student traders have the price history of the foreign currency in dollars in their computers. They may or may not also have software that highlights trends. Describe two designs for this experiment, a completely randomized design and a matched pairs design in which each student serves as his or her own control. In both cases, carry out the randomization required by the design.
- 3.41 **Comparing weight-loss treatments.** Twenty overweight females have agreed to participate in a study of the effectiveness of 4 weight-loss treatments: A, B, C, and D. The company researcher first calculates how overweight each subject is by comparing the subject's actual weight with her "ideal" weight. The subjects and their excess weights in pounds are

Birnbaum	35	Hernandez	25	Moses	25	Smith	29
Brown	34	Jackson	33	Nevesky	39	Stall	33
Brunk	30	Kendall	28	Obrach	30	Tran	35
Cruz	34	Loren	32	Rodriguez	30	Wilansky	42
Deng	24	Mann	28	Santiago	27	Williams	22

The response variable is the weight lost after 8 weeks of treatment. Because a subject's excess weight will influence the response, a block design is appropriate.

- Arrange the subjects in order of increasing excess weight. Form 5 blocks of 4 subjects each by grouping the 4 least overweight, then the next 4, and so on.
- Use Table B to randomly assign the 4 subjects in each block to the 4 weight-loss treatments. Be sure to explain exactly how you used the table.

SECTION 3.2 SUMMARY

- In an experiment, we impose one or more **treatments** on the **subjects**. Each treatment is a combination of levels of the explanatory variables, which we call **factors**.
- The **design** of an experiment describes the choice of treatments and the manner in which the subjects are assigned to the treatments.

- The basic principles of statistical design of experiments are **control**, **randomization**, and **replication**.
- The simplest form of control is **comparison**. Experiments should compare two or more treatments in order to avoid **confounding** of the effect of a treatment with other influences, such as lurking variables.
- **Randomization** uses chance to assign subjects to the treatments. Randomization creates treatment groups that are similar (except for chance variation) before the treatments are applied. Randomization and comparison together prevent **bias**, or systematic favoritism, in experiments.
- You can carry out randomization by giving numerical labels to the subjects and using a **table of random digits** to choose treatment groups.
- **Replication** of each treatment on many subjects reduces the role of chance variation and makes the experiment more sensitive to differences among the treatments.
- Good experiments require attention to detail as well as good statistical design. Many behavioral and medical experiments are **double-blind**. Lack of **realism** in an experiment can prevent us from generalizing its results.
- In addition to comparison, a second form of control is to restrict randomization by forming **blocks** of subjects that are similar in some way that is important to the response. Randomization is then carried out separately within each block.
- **Matched pairs** are a common form of blocking for comparing just two treatments. In some matched pairs designs, each subject receives both treatments in a random order. In others, the subjects are matched in pairs as closely as possible, and one subject in each pair receives each treatment.

SECTION 3.2 EXERCISES

- 3.42 **Public housing.** A study of the effect of living in public housing on the income and other variables in poverty-level households was carried out as follows. The researchers obtained a list of all applicants for public housing during the previous year. Some applicants had been accepted, while others had been turned down by the housing authority. Both groups were interviewed and compared. Is this study an experiment or an observational study? Why? What are the explanatory and response variables? Why will confounding make it difficult to see the effect of the explanatory variable on the response variables?
- 3.43 **Effects of price promotions.** A researcher studying the effect of price promotions on consumers' expectations makes up a history of the store price of a hypothetical brand of laundry detergent for the past year. Students in a marketing course view the price history on a computer. Some students see a steady price, while others see regular promotions that temporarily cut the price. Then the students are asked what price they would expect to pay for the detergent. Is this study an experiment? Why? What are the explanatory and response variables?

- 3.44 **Computer telephone calls.** You can use your computer to make telephone calls over the Internet. How will the cost affect the behavior of users of this service? You will offer the service to all 200 rooms in a college dormitory. Some rooms will pay a low flat rate. Others will pay higher rates at peak periods and very low rates off-peak. You are interested in the amount and time of use and in the effect on the congestion of the network. Outline the design of an experiment to study the effect of rate structure.
- 3.45 **Marketing to children.** If children are given more choices within a class of products, will they tend to prefer that product to a competing product that offers fewer choices? Marketers want to know. An experiment prepared three sets of beverages. Set 1 contained two milk drinks and two fruit drinks. Set 2 had two fruit drinks and four milk drinks. Set 3 contained four fruit drinks but only two milk drinks. The researchers divided 210 children aged 4 to 12 years into three groups at random. They offered each group one of the sets. As each child chose a beverage to drink from the set presented, the researchers noted whether the choice was a milk drink or a fruit drink.
- What are the experimental subjects?
 - What is the factor and what are its levels? What is the response variable?
 - Use a diagram to outline a completely randomized design for the study.
 - Explain how you would assign labels to the subjects. Use Table B at line 125 to choose the first 5 subjects assigned to the first treatment.
- 3.46 **Clinical trial basics.** Fizz Laboratories, a pharmaceutical company, has developed a new pain-relief medication. Three hundred patients suffering from arthritis and needing pain relief are available. Each patient will be treated and asked an hour later, “About what percentage of pain relief did you experience?”
- Why should Fizz not simply administer the new drug and record the patients’ responses?
 - Outline the design of an experiment to compare the drug’s effectiveness with that of aspirin and of a placebo.
 - Should patients be told which drug they are receiving? How would this knowledge probably affect their reactions?
 - If patients are not told which treatment they are receiving, the experiment is single-blind. Should this experiment be double-blind also? Explain.
- 3.47 **Treating prostate disease.** A large study used records from Canada’s national health care system to compare the effectiveness of two ways to treat prostate disease. The two treatments are traditional surgery and a new method that does not require surgery. The records described many patients whose doctors had chosen each method. The study found that patients treated by the new method were significantly more likely to die within 8 years.¹⁸
- Further study of the data showed that this conclusion was wrong. The extra deaths among patients who got the new method could be explained by lurking variables. What lurking variables might be confounded with a doctor’s choice of surgical or nonsurgical treatment?
 - You have 300 prostate patients who are willing to serve as subjects in an experiment to compare the two methods. Use a diagram to outline the design of a randomized comparative experiment.

- 3.48 **Aspirin and heart attacks.** “Nearly five decades of research now link aspirin to the prevention of stroke and heart attacks.” So says the Bayer Aspirin Web site, www.bayeraspirin.com. The most important evidence for this claim comes from the Physicians’ Health Study, a large medical experiment involving 22,000 male physicians. One group of about 11,000 physicians took an aspirin every second day, while the rest took a placebo. After several years the study found that subjects in the aspirin group had significantly fewer heart attacks than subjects in the placebo group.
- Identify the experimental subjects, the factor and its levels, and the response variable in the Physicians’ Health Study.
 - Use a diagram to outline a completely randomized design for the Physicians’ Health Study.
 - What does it mean to say that the aspirin group had “significantly fewer heart attacks?”
- 3.49 **Prayer and meditation.** Not all effective medical treatments come from pharmaceutical companies. You read in a magazine that “nonphysical treatments such as meditation and prayer have been shown to be effective in controlled scientific studies for such ailments as high blood pressure, insomnia, ulcers, and asthma.” Explain in simple language what the article means by “controlled scientific studies” and why such studies can show that meditation and prayer are effective treatments for some medical problems.
- 3.50 **Sickle cell disease.** Exercise 3.29 (page 190) describes a medical study of a new treatment for sickle cell disease.
- Use a diagram to outline the design of this experiment.
 - Use of a placebo is considered ethical if there is no effective standard treatment to give the control group. It might seem humane to give all the subjects hydroxyurea in the hope that it will help them. Explain clearly why this would not provide information about the effectiveness of the drug. (In fact, the experiment was stopped ahead of schedule because the hydroxyurea group had only half as many pain episodes as the control group. Ethical standards required stopping the experiment as soon as significant evidence became available.)
- 3.51 **Reducing health care spending.** Will people spend less on health care if their health insurance requires them to pay some part of the cost themselves? An experiment on this issue asked if the percent of medical costs that are paid by health insurance has an effect either on the amount of medical care that people use or on their health. The treatments were four insurance plans. Each plan paid all medical costs above a ceiling. Below the ceiling, the plans paid 100%, 75%, 50%, or 0% of costs incurred.
- Outline the design of a randomized comparative experiment suitable for this study.
 - Describe briefly the practical and ethical difficulties that might arise in such an experiment.
- 3.52 **Effects of TV advertising.** You decide to use a completely randomized design in the two-factor experiment on response to advertising described in Example 3.12 (page 190). The 36 students named below will serve as subjects. Outline the design. Then use Table B at line 130 to randomly assign the subjects to the 6 treatments.

Alomar	Denman	Han	Liang	Padilla	Valasco
Ashiro	Durr	Howard	Maldonado	Plochman	Vaughn
Bennett	Edwards	Hruska	Marsden	Rosen	Wei
Bikalis	Farouk	Imrani	Montoya	Solomon	Wilder
Chao	Fleming	James	O'Brian	Trujillo	Willis
Clemente	George	Kaplan	Ogle	Tullock	Zhang

- 3.53 **Temperature and work performance.** An expert on worker performance is interested in the effect of room temperature on the performance of tasks requiring manual dexterity. She chooses temperatures of 20° C (68° F) and 30° C (86° F) as treatments. The response variable is the number of correct insertions, during a 30-minute period, in a peg-and-hole apparatus that requires the use of both hands simultaneously. Each subject is trained on the apparatus and then asked to make as many insertions as possible in 30 minutes of continuous effort.
- Outline a completely randomized design to compare dexterity at 20° and 30° . Twenty subjects are available.
 - Because individuals differ greatly in dexterity, the wide variation in individual scores may hide the systematic effect of temperature unless there are many subjects in each group. Describe in detail the design of a matched pairs experiment in which each subject serves as his or her own control.
- 3.54 **Reaching Mexican Americans.** Advertising that hopes to attract Mexican Americans must keep in mind the cultural orientation of these consumers. There are several psychological tests available to measure the extent to which Mexican Americans are oriented toward Mexican/Spanish or Anglo/English culture. Two such tests are the Bicultural Inventory (BI) and the Acculturation Rating Scale for Mexican Americans (ARSMA). To study the relationship between the scores on these two tests, researchers will give both tests to a group of 22 Mexican Americans.
- Briefly describe a matched pairs design for this study. In particular, how will you use randomization in your design?
 - You have an alphabetized list of the subjects (numbered 1 to 22). Carry out the randomization required by your design and report the result.
- 3.55 **Absorption of a drug.** Example 3.11 (page 189) describes a study in which a maker of pharmaceuticals compares the concentration of a drug in the blood of patients 30 minutes after injecting one of three doses. Use a diagram to outline a completely randomized design for this experiment.
- 3.56 **Absorption of a drug: another step.** The drug studied in Exercise 3.55 can be administered by injection, by a skin patch, or by intravenous drip. Concentration in the blood may depend both on the dose and on the method of administration. Make a sketch that describes the treatments formed by combining dosage and method. Then use a diagram to outline a completely randomized design for this two-factor experiment.
- 3.57 **Potatoes.** A horticulturist is comparing two methods (call them A and B) of growing potatoes. Standard potato cuttings will be planted in small plots of ground. The response variables are number of tubers per plant and fresh weight (weight when just harvested) of vegetable growth per plant. There

are 20 plots available for the experiment. Sketch a rectangular field divided into 5 rows of 4 plots each. Then diagram the experimental design and do the required randomization. (If you use Table B, start at line 145.) Mark on your sketch which growing method you will use in each plot.

3.3 Toward Statistical Inference

statistical inference

A market research firm interviews a random sample of 2500 adults. Result: 66% find shopping for clothes frustrating and time-consuming. That's the truth about the 2500 people in the sample. What is the truth about the almost 210 million American adults who make up the population? Because the sample was chosen at random, it's reasonable to think that these 2500 people represent the entire population pretty well. So the market researchers turn the *fact* that 66% of the *sample* find shopping frustrating into an *estimate* that about 66% of *all adults* feel this way. That's a basic move in statistics: use a fact about a sample to estimate the truth about the whole population. We call this **statistical inference** because we infer conclusions about the wider population from data on selected individuals. To think about inference, we must keep straight whether a number describes a sample or a population. Here is the vocabulary we use.

PARAMETERS AND STATISTICS

A **parameter** is a number that describes the **population**. A parameter is a fixed number, but in practice we do not know its value.

A **statistic** is a number that describes a **sample**. The value of a statistic is known when we have taken a sample, but it can change from sample to sample. We often use a statistic to estimate an unknown parameter.



CASE 3.1 IS CLOTHES SHOPPING FRUSTRATING?

Changing consumer attitudes toward shopping are of great interest to retailers and makers of consumer goods. The Yankelovich market research firm specializes in the study of “consumer trends” that go beyond preferences for one or another product. The firm promises its clients that its work “reveals the reasons why consumers are changing their behavior to create the lifestyle they want. We also show you how smart businesses are using that knowledge to make smarter marketing decisions.”¹⁹

One trend of concern to marketers is that fewer people enjoy shopping than in the past. Yankelovich conducts an annual survey of consumer attitudes. The population is all U.S. residents aged 18 and over. A recent survey asked a nationwide random sample of 2500 adults if they agreed or disagreed that “I like buying new clothes, but shopping is often frustrating and time-consuming.” Of the respondents, 1650 said they agreed. That’s

66%, a percent high enough to encourage development of alternatives such as online sales.



The proportion of the Yankelovich sample who agreed that clothes shopping is often frustrating is

$$\hat{p} = \frac{1650}{2500} = 0.66 = 66\%$$

The number $\hat{p} = 0.66$ is a *statistic*. The corresponding *parameter* is the proportion (call it p) of all adult U.S. residents who would have said “Agree” if asked the same question. We don’t know the value of the parameter p , so we use the statistic \hat{p} as an estimate.

APPLY YOUR KNOWLEDGE

- 3.58 **Unlisted telephone numbers.** A telemarketing firm in Los Angeles uses a device that dials residential telephone numbers in that city at random. Of the first 100 numbers dialed, 43 are unlisted. This is not surprising, because 52% of all Los Angeles residential phones are unlisted. Which of the bold numbers is a parameter and which is a statistic?
- 3.59 **Indianapolis voters.** Voter registration records show that 68% of all voters in Indianapolis are registered as Republicans. To test a random-digit dialing device, you use the device to call 150 randomly chosen residential telephones in Indianapolis. Of the registered voters contacted, 73% are registered Republicans. Which of the bold numbers is a parameter and which is a statistic?

Sampling variability, sampling distributions

If the Yankelovich firm took a second random sample of 2500 adults, the new sample would have different people in it. It is almost certain that there would not be exactly 1650 positive responses. That is, the value of the statistic \hat{p} will *vary* from sample to sample. Could it happen that one random sample finds that 66% of adults find clothes shopping frustrating and a second random sample finds that only 42% feel this way? Random samples eliminate *bias* from the act of choosing a sample, but they can still be wrong because of the *variability* that results when we choose at random. If the variation when we take repeat samples from the same population is too great, we can’t trust the results of any one sample.

We are saved by the second great advantage of random samples. The first advantage is that choosing at random eliminates favoritism. That is, random sampling attacks bias. The second advantage is that if we take lots of random samples of the same size from the same population, the variation from sample to sample will follow a predictable pattern. **All of statistical inference is based on one idea: to see how trustworthy a procedure is, ask what would happen if we repeated it many times.**

To understand why sampling variability is not fatal, we therefore ask, “What would happen if we took many samples?” Here’s how to answer that question:

- Take a large number of samples from the same population.
- Calculate the sample proportion \hat{p} for each sample.

- Make a histogram of the values of \hat{p} .
- Examine the distribution displayed in the histogram for shape, center, and spread, as well as outliers or other deviations.

We can't afford to actually take many samples from a large population such as all adult U.S. residents. But we can imitate many samples by using random digits. Using random digits from a table or computer software to imitate chance behavior is called **simulation**.

EXAMPLE 3.20

Lots of samples



Suppose that in fact (unknown to Yankelovich), exactly 60% of all adults find shopping for clothes frustrating and time-consuming. That is, the truth about the population is that $p = 0.6$. What if we select an SRS of size 100 from this population and use the sample proportion \hat{p} to estimate the unknown value of the population proportion p ? Using software, we simulated doing this 1000 times.

Figure 3.7 illustrates the process of choosing many samples and finding \hat{p} for each one. In the first sample, 56 of the 100 people say they find shopping frustrating, so $\hat{p} = 56/100 = 0.56$. Only 46 in the next sample feel this way, so for that sample $\hat{p} = 0.46$. Choose 1000 samples and make a histogram of the 1000 values of \hat{p} . That's the graph at the right of Figure 3.7.

Of course, Yankelovich interviewed 2500 people, not just 100. Figure 3.8 shows the process of choosing 1000 SRSs, each of size 2500, from a population in which the true sample proportion is $p = 0.6$. The 1000 values of \hat{p} from these samples form the histogram at the right of the figure. Figures 3.7 and 3.8 are drawn on the same scale. Comparing them shows what happens when we increase the size of our samples from 100 to 2500. These histograms display the *sampling distribution* of the statistic \hat{p} for two sample sizes.

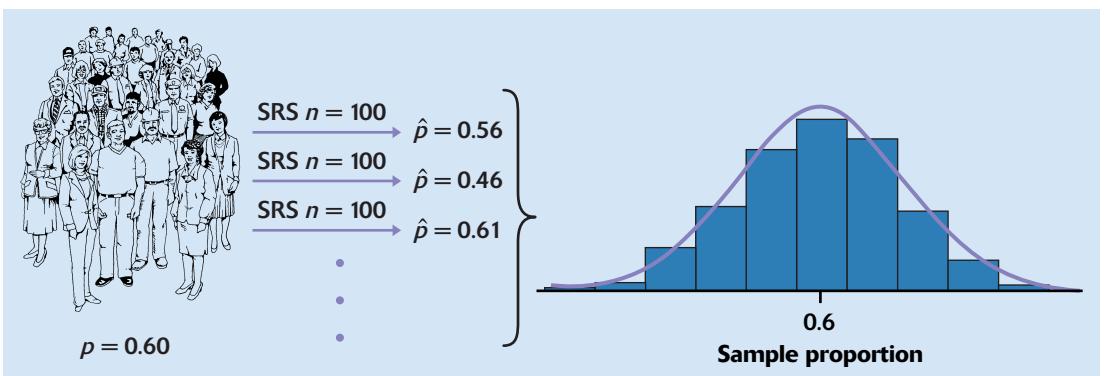


FIGURE 3.7 The results of many SRSs have a regular pattern. Here, we draw 1000 SRSs of size 100 from the same population. The population proportion is $p = 0.60$. The histogram shows the distribution of the 1000 sample proportions \hat{p} .

SAMPLING DISTRIBUTION

The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Strictly speaking, the sampling distribution is the ideal pattern that would emerge if we looked at all possible samples of the same size from our population. A distribution obtained from a fixed number of samples, like the 1000 samples in Figures 3.7 and 3.8, is only an approximation to the sampling distribution. One of the uses of probability theory in statistics is to obtain sampling distributions without simulation. The interpretation of a sampling distribution is the same, however, whether we obtain it by simulation or by the mathematics of probability.

We can use the tools of data analysis to describe any distribution. Let's apply those tools to Figures 3.7 and 3.8.

- **Shape:** The histograms look Normal. Figure 3.9 is a Normal quantile plot of the values of \hat{p} for our samples of size 100. It confirms that the distribution in Figure 3.7 is close to Normal. The 1000 values for samples of size 2500 in Figure 3.8 are even closer to Normal. The Normal curves drawn through the histograms describe the overall shape quite well.
- **Center:** In both cases, the values of the sample proportion \hat{p} vary from sample to sample, but the values are centered at 0.6. Recall that $p = 0.6$ is the true population parameter. Some samples have a \hat{p} less than 0.6 and some greater, but there is no tendency to be always low or always high. That is, \hat{p} has no *bias* as an estimator of p . This is true for both large and small samples. (Want the details? The mean of the 1000 values of \hat{p} is 0.598 for samples of size 100 and 0.6002 for samples of size 2500. The median value of \hat{p} is exactly 0.6 for samples of both sizes.)
- **Spread:** The values of \hat{p} from samples of size 2500 are much less spread out than the values from samples of size 100. In fact, the standard deviations are 0.051 for Figure 3.7 and 0.0097, or about 0.01, for Figure 3.8.

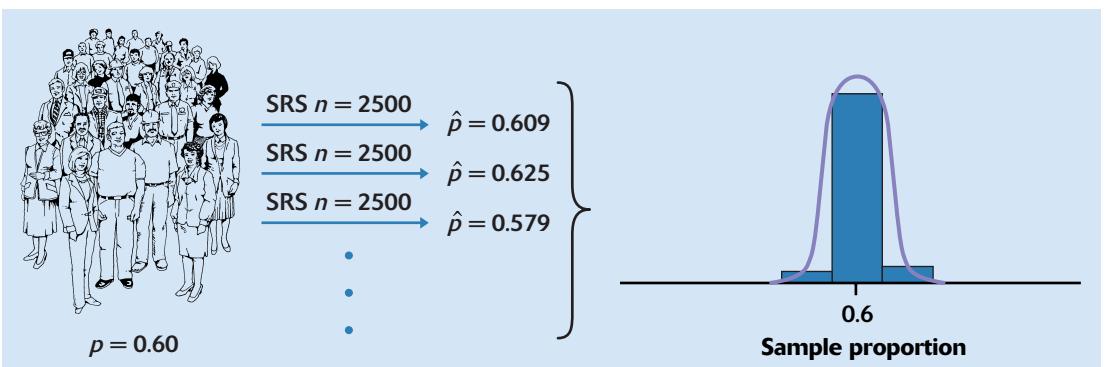


FIGURE 3.8 The distribution of sample proportions \hat{p} for 1000 SRSs of size 2500 drawn from the same population as in Figure 3.7. The two histograms have the same scale. The statistic from the larger sample is less variable.

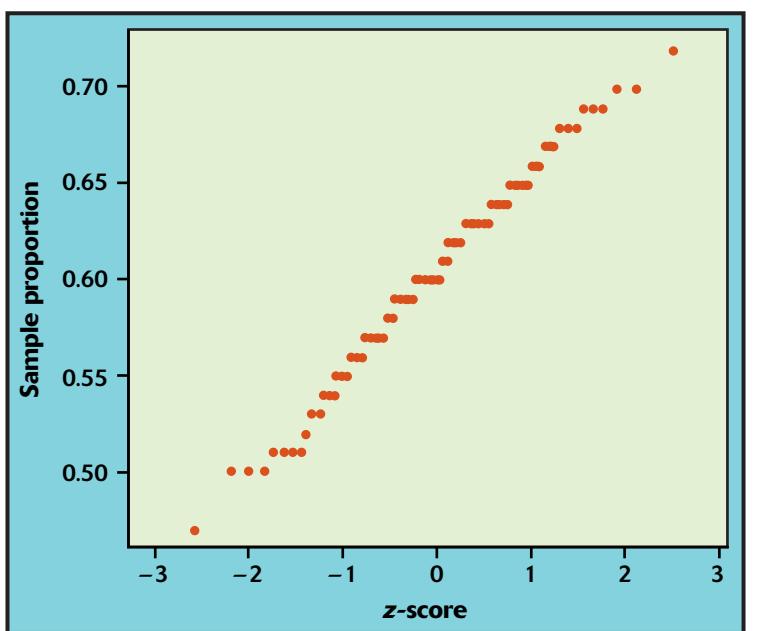


FIGURE 3.9 Normal quantile plot of the sample proportions in Figure 3.7. The distribution is close to Normal except for a staircase pattern due to the fact that sample proportions from a sample of size 100 can take only values that are multiples of 0.01. Because a plot of 1000 points is hard to read, this plot presents only every 10th value.

Although these results describe just two sets of simulations, they reflect facts that are true whenever we use random sampling.

APPLY YOUR KNOWLEDGE

- 3.60 **Simulation by hand.** You can use a table of random digits to simulate sampling from a population. Suppose that 60% of the population find shopping frustrating. That is, the population proportion is $p = 0.6$. CASE 3.1
- Let each digit in the table stand for one person in this population. Digits 0 to 5 stand for people who find shopping frustrating, and 6 to 9 stand for people who do not. Why does looking at one digit from Table B simulate drawing one person at random from a population with $p = 0.6$?
 - The first 100 entries in Table B contain 63 digits between 0 and 5, so $\hat{p} = 63/100 = 0.63$ for this sample. Why do the first 100 digits simulate drawing 100 people at random?
 - Simulate a second SRS from this population, using 100 entries in Table B, starting at line 104. What is \hat{p} for this sample? You see that the two sample results are different, and neither is equal to the true population value $p = 0.6$. That's the issue we face in this section.
- 3.61 **Making movies, making money.** During the 1990s, a total of 1986 movies were released in the United States. We used statistical software to choose several SRSs of size 25 from this population. Our first sample of 25 movies had mean domestic gross sales 33.916 million dollars, standard deviation

TABLE 3.1 Sample of 25 movies released in the 1990s

Movie name (year)	Domestic gross (\$ millions)
<i>Aces: Iron Eagle III</i> (1992)	2.5
<i>BASEketball</i> (1998)	7.0
<i>Body of Evidence</i> (1993)	13.3
<i>Car 54, Where Are You?</i> (1994)	1.2
<i>City of Angels</i> (1998)	78.9
<i>Coneheads</i> (1993)	21.3
<i>Days of Thunder</i> (1990)	82.7
<i>Death Warrant</i> (1990)	16.9
<i>Desperate Hours</i> (1990)	2.7
<i>Ernest Scared Stupid</i> (1991)	14.1
<i>Executive Decision</i> (1996)	68.8
<i>For Love of the Game</i> (1999)	35.2
<i>I Come in Peace</i> (1990)	4.3
<i>Jumanji</i> (1995)	100.2
<i>Kika</i> (1993)	2.1
<i>Miami Rhapsody</i> (1995)	5.2
<i>Mighty, The</i> (1998)	2.6
<i>Perez Family, The</i> (1995)	2.8
<i>Revenge</i> (1990)	15.7
<i>Shine</i> (1996)	35.8
<i>So I Married an Axe Murderer</i> (1993)	11.6
<i>Thinner</i> (1996)	15.2
<i>Wedding Singer, The</i> (1998)	80.2
<i>Wing Commander</i> (1999)	11.6
<i>Xizao</i> (1999)	1.2

50.2697 million dollars, and median 16.1 million dollars. A second random sample of 25 movies had mean 38.712 million dollars, standard deviation 58 million dollars, and median 14.1 million dollars. Table 3.1 lists the members of a third SRS.²⁰

- (a) Calculate the mean domestic gross \bar{x} for the sample in Table 3.1. How does it compare with the means of the first two samples?
- (b) Calculate the standard deviation s . How does it compare with that of the first two samples?
- (c) Calculate the median M . How does it compare with the previous two medians?

- 3.62 Making movies, making money, continued. There are over 1.5×10^{57} different samples of size 25 possible from the 1986 movies released in the

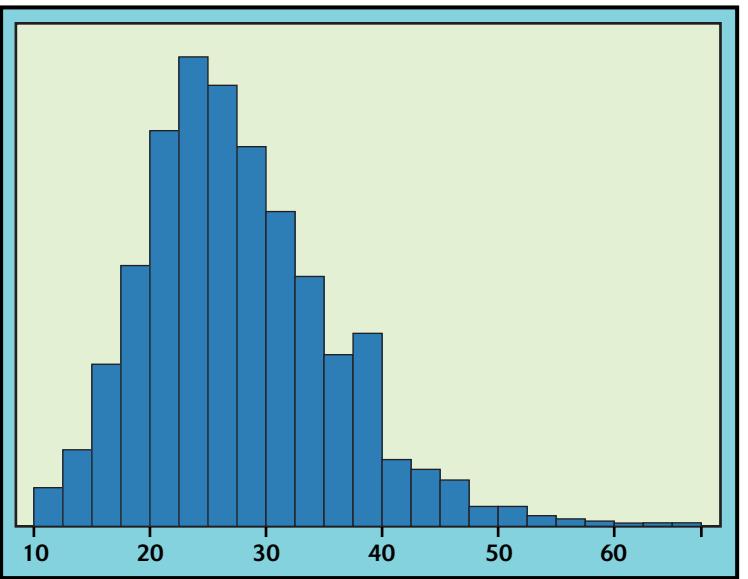


FIGURE 3.10 The distribution of sample means \bar{x} for 1000 SRSs of size 25 from the domestic gross sales (millions of dollars) of all movies released in the United States in the 1990s (for Exercises 3.62 and 3.63).

1990s. Table 3.1 displays just one of these samples. The distribution of the 1.5×10^{57} values of \bar{x} from all these samples is the sampling distribution of the sample mean. Let's look at just 1000 samples to get a rough picture of this sampling distribution.

- (a) Figure 3.10 is a histogram of 1000 sample means based on 1000 different samples of size 25 from the population of movies. Describe the shape, center, and spread of this distribution.
- (b) In practice, we take a sample because we don't have access to the entire population. In this case, however, we know the mean domestic gross for the population of movies released in the 1990s. It is 27.77 million dollars. Given this information and the histogram in Figure 3.10, what can you say about how far a particular sample mean might be from the target of 27.77 million dollars?

Bias and variability

Our simulations show that a sample of size 2500 will almost always give an estimate \hat{p} that is close to the truth about the population. Figure 3.8 illustrates this fact for just one value of the population proportion, but it is true for any population. Samples of size 100, on the other hand, might give an estimate of 50% or 70% when the truth is 60%.

Thinking about Figures 3.7 and 3.8 helps us restate the idea of bias when we use a statistic like \hat{p} to estimate a parameter like p . It also reminds us that variability matters as much as bias.

BIAS AND VARIABILITY

Bias concerns the center of the sampling distribution. A statistic used to estimate a parameter is **unbiased** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

The **variability** of a statistic is described by the spread of its sampling distribution. This spread is determined by the sampling design and the sample size n . Statistics from larger samples have smaller spreads.

We can think of the true value of the population parameter as the bull's-eye on a target, and of the sample statistic as an arrow fired at the bull's-eye. Bias and variability describe what happens when an archer fires many arrows at the target. Bias means that the aim is off, and the arrows land consistently off the bull's-eye in the same direction. The sample values do not center about the population value. Large *variability* means that repeated shots are widely scattered on the target. Repeated samples do not give similar results but differ widely among themselves. Figure 3.11 shows this target illustration of the two types of error.

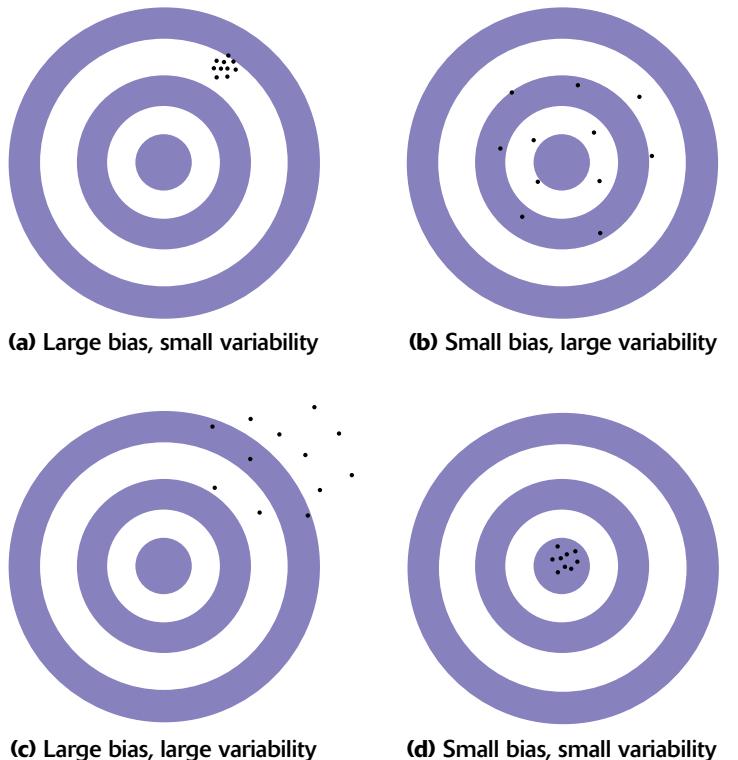


FIGURE 3.11 Bias and variability in shooting arrows at a target. Bias means the archer systematically misses in the same direction. Variability means that the arrows are scattered.

Notice that small variability (repeated shots are close together) can accompany large bias (the arrows are consistently away from the bull's-eye in one direction). And small bias (the arrows center on the bull's-eye) can accompany large variability (repeated shots are widely scattered). A good sampling scheme, like a good archer, must have both small bias and small variability. Here's how we do this.

MANAGING BIAS AND VARIABILITY

To reduce bias, use random sampling. When we start with a list of the entire population, simple random sampling produces **unbiased estimates**—the values of a statistic computed from an SRS neither consistently overestimate nor consistently underestimate the value of the population parameter.

To reduce the variability of a statistic from an SRS, use a larger sample. You can make the variability as small as you want by taking a large enough sample.

In practice, Yankelovich takes only one sample. We don't know how close to the truth an estimate from this one sample is, because we don't know what the truth about the population is. But *large random samples almost always give an estimate that is close to the truth*. Looking at the pattern of many samples shows that we can trust the result of one sample. The Current Population Survey's sample of 55,000 households estimates the national unemployment rate very accurately. Of course, only probability samples carry this guarantee. *Nightline's voluntary response sample* (Example 3.3) is worthless even though 186,000 people called in. Using a probability sampling design and taking care to deal with practical difficulties reduce bias in a sample. The size of the sample then determines how close to the population truth the sample result is likely to fall. Results from a sample survey usually come with a *margin of error* that sets bounds on the size of the likely error. How to do this is part of the detail of statistical inference. We will describe the reasoning in Chapter 6.

APPLY YOUR KNOWLEDGE

- 3.63 **Making movies, making money, continued.** During the 1990s, a total of 1986 movies were released in the United States. We want to estimate the mean domestic gross sales of this population. We will take an SRS and use the sample mean \bar{x} as our estimate.
- Figure 3.10 is a histogram of 1000 sample means from 1000 different samples of size 25. What is the approximate range (smallest to largest) of these sample means?
 - Figure 3.12 is a histogram of 1000 sample means from 1000 different samples of size 10. What is the approximate range of these sample means? How does this compare with the spread in part (a)?
 - Figure 3.13 is a histogram of 1000 sample means from 1000 different samples of size 100. What is the approximate range? How does this compare with the spread in part (a)? In part (b)?
 - What important fact do these three sampling distributions illustrate?

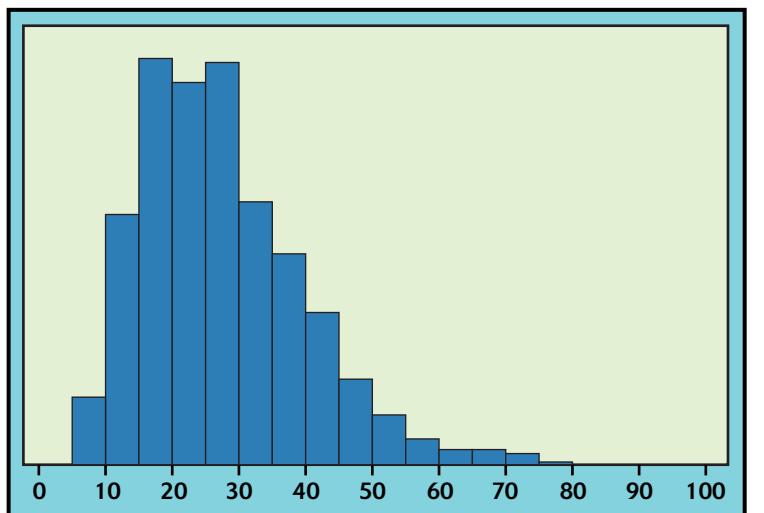


FIGURE 3.12 The distribution of sample means \bar{x} for 1000 SRSs of size 10 from the domestic gross sales (millions of dollars) of all movies released in the United States in the 1990s (for Exercise 3.63).

- 3.64 Ask more people. Just before a presidential election, a national opinion-polling firm increases the size of its weekly sample from the usual 1500 people to 4000 people. Why do you think the firm does this?
- 3.65 Canada's national health care. The Ministry of Health in the Canadian province of Ontario wants to know whether the national health care system is achieving its goals in the province. Much information about health

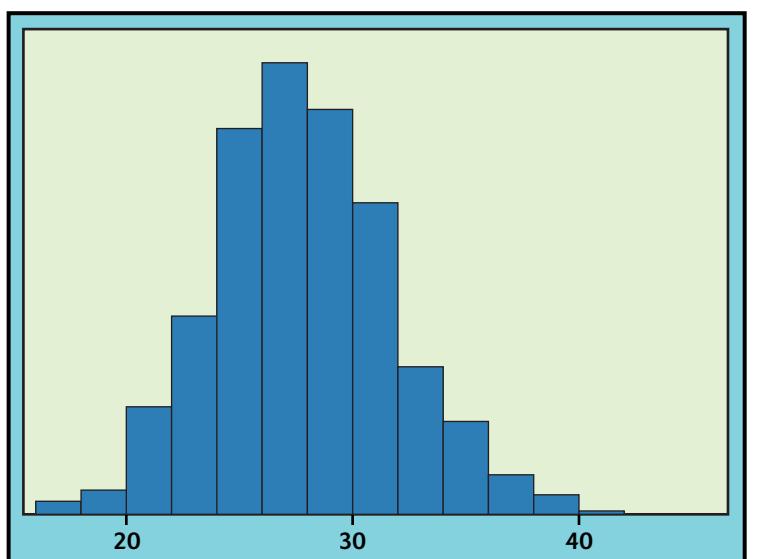


FIGURE 3.13 The distribution of sample means \bar{x} for 1000 SRSs of size 100 from the domestic gross sales (millions of dollars) of all movies released in the United States in the 1990s (for Exercise 3.63).

care comes from patient records, but that source doesn't allow us to compare people who use health services with those who don't. So the Ministry of Health conducted the Ontario Health Survey, which interviewed a probability sample of 61,239 people who live in Ontario.²¹

- (a) What is the population for this sample survey? What is the sample?
- (b) The survey found that 76% of males and 86% of females in the sample had visited a general practitioner at least once in the past year. Do you think these estimates are close to the truth about the entire population? Why?

Sampling from large populations

Yankelovich's sample of 2500 adults is only about 1 out of every 85,000 adults in the United States. Does it matter whether we sample 1 in 100 individuals in the population or 1 in 85,000?

POPULATION SIZE DOESN'T MATTER

The variability of a statistic from a random sample does not depend on the size of the population, as long as the population is at least 100 times larger than the sample.

Why does the size of the population have little influence on the behavior of statistics from random samples? To see why this is plausible, imagine sampling harvested corn by thrusting a scoop into a lot of corn kernels. The scoop doesn't know whether it is surrounded by a bag of corn or by an entire truckload. As long as the corn is well mixed (so that the scoop selects a random sample), the variability of the result depends only on the size of the scoop.

The fact that the variability of sample results is controlled by the size of the sample has important consequences for sampling design. An SRS of size 2500 from the 280 million residents of the United States gives results as precise as an SRS of size 2500 from the 740,000 inhabitants of San Francisco. This is good news for designers of national samples but bad news for those who want accurate information about the citizens of San Francisco. If both use an SRS, both must use the same size sample to obtain equally trustworthy results.

Why randomize?

Why randomize? The act of randomizing guarantees that our data are subject to the laws of probability. The behavior of statistics is described by a sampling distribution. The form of the distribution is known, and in many cases is approximately Normal. Often, the center of the distribution lies at the true parameter value, so the notion that randomization eliminates bias is made more precise. The spread of the distribution describes the variability of the statistic and can be made as small as we wish by choosing a large enough sample. Randomized experiments behave similarly: we can reduce variability by choosing larger groups of subjects for each treatment.

These facts are at the heart of formal statistical inference. Later chapters will have much to say in more technical language about sampling distributions and the way statistical conclusions are based on them. What any user of statistics must understand is that all the technical talk has its basis in a simple question: What would happen if the sample or the experiment were repeated many times? The reasoning applies not only to an SRS but also to the complex sampling designs actually used by opinion polls and other national sample surveys. The same conclusions hold as well for randomized experimental designs. The details vary with the design but the basic facts are true whenever randomization is used to produce data.

Remember that proper statistical design is not the only aspect of a good sample or experiment. The sampling distribution shows only how a statistic varies due to the operation of chance in randomization. It reveals nothing about possible bias due to undercoverage or nonresponse in a sample, or to lack of realism in an experiment. The true distance of a statistic from the parameter it is estimating can be much larger than the sampling distribution suggests. What is worse, there is no way to say how large the added error is. The real world is less orderly than statistics textbooks imply.

SECTION 3.3 SUMMARY

- A number that describes a population is a **parameter**. A number that can be computed from the data is a **statistic**. The purpose of sampling or experimentation is usually to use statistics to make statements about unknown parameters.
- A statistic from a probability sample or randomized experiment has a **sampling distribution** that describes how the statistic varies in repeated data production. The sampling distribution answers the question, “What would happen if we repeated the sample or experiment many times?” Formal statistical inference is based on the sampling distributions of statistics.
- A statistic as an estimator of a parameter may suffer from **bias** or from high **variability**. Bias means that the center of the sampling distribution is not equal to the true value of the parameter. The variability of the statistic is described by the spread of its sampling distribution.
- Properly chosen statistics from randomized data production designs have no bias resulting from the way the sample is selected or the way the subjects are assigned to treatments. We can reduce the variability of the statistic by increasing the size of the sample or the size of the experimental groups.

SECTION 3.3 EXERCISES

- 3.66 **Unemployment.** The Bureau of Labor Statistics announces that last month it interviewed all members of the labor force in a sample of 55,000 households; 6.2% of the people interviewed were unemployed. Is the bold number a parameter or a statistic? Why?

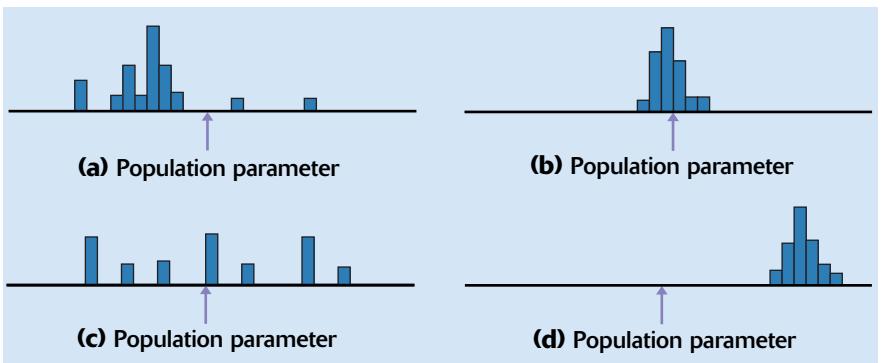


FIGURE 3.14 Which of these sampling distributions displays high or low bias and high or low variability? (For Exercise 3.68.)

- 3.67 **Acceptance sampling.** A carload lot of ball bearings has a mean diameter of 2.503 centimeters (cm). This is within the specifications for acceptance of the lot by the purchaser. The inspector happens to inspect 100 bearings from the lot with a mean diameter of 2.515 cm. This is outside the specified limits, so the lot is mistakenly rejected. Is each of the bold numbers a parameter or a statistic? Explain your answers.
- 3.68 **Bias and variability.** Figure 3.14 shows histograms of four sampling distributions of statistics intended to estimate the same parameter. Label each distribution relative to the others as high or low bias and as high or low variability.
- 3.69 **Sampling students.** A management student is planning to take a survey of student attitudes toward part-time work while attending college. He develops a questionnaire and plans to ask 25 randomly selected students to fill it out. His faculty advisor approves the questionnaire but urges that the sample size be increased to at least 100 students. Why is the larger sample helpful?
- 3.70 **Sampling in the states.** The Internal Revenue Service plans to examine an SRS of individual federal income tax returns from each state. One variable of interest is the proportion of returns claiming itemized deductions. The total number of individual tax returns in a state varies from 14 million in California to 227,000 in Wyoming.
- Will the variability of the sample proportion vary from state to state if an SRS of size 2000 is taken in each state? Explain your answer.
 - Will the variability of the sample proportion change from state to state if an SRS of 1/10 of 1% (0.001) of the state's population is taken in each state? Explain your answer.
- 3.71 **Margin of error.** A *New York Times* opinion poll on women's issues contacted a sample of 1025 women and 472 men by randomly selecting telephone numbers. The *Times* publishes descriptions of its polling methods. Here is part of the description for this poll:

In theory, in 19 cases out of 20 the results based on the entire sample will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all adult Americans.

The potential sampling error for smaller subgroups is larger. For example, for men it is plus or minus five percentage points.²²

Explain why the margin of error is larger for conclusions about men alone than for conclusions about all adults.

- 3.72 Coin tossing.** Coin tossing can illustrate the idea of a sampling distribution. The population is all outcomes (heads or tails) we would get if we tossed a coin forever. The parameter p is the proportion of heads in this population. We suspect that p is close to 0.5. That is, we think the coin will show about one-half heads in the long run. The sample is the outcomes of 20 tosses, and the statistic \hat{p} is the proportion of heads in these 20 tosses.

- Toss a coin 20 times and record the value of \hat{p} .
- Repeat this sampling process 10 times. Make a stemplot of the 10 values of \hat{p} . Is the center of this distribution close to 0.5? (Ten repetitions give only a crude approximation to the sampling distribution. If possible, pool your work with that of other students to obtain at least 100 repetitions and make a histogram of the values of \hat{p} .)

- 3.73 Sampling invoices.** We will illustrate the idea of a sampling distribution in the case of a very small sample from a very small population. The population contains 10 past due invoices. Here are the number of days each invoice is past due:

Invoice	0	1	2	3	4	5	6	7	8	9
Days past due	8	12	10	5	7	3	15	9	7	6

The parameter of interest is the mean of this population, which is 8.2. The sample is an SRS of $n = 4$ invoices drawn from the population. Because the invoices are labeled 0 to 9, a single random digit from Table B chooses one invoice for the sample.

- Use Table B to draw an SRS of size 4 from this population. Write the past due values for the days in your sample and calculate their mean \bar{x} . This statistic is an estimate of the population parameter.
- Repeat this process 10 times. Make a histogram of the 10 values of \bar{x} . You are constructing the sampling distribution of \bar{x} . Is the center of your histogram close to 8.2? (Ten repetitions give only a crude approximation to the sampling distribution. If possible, pool your work with that of other students—using different parts of Table B—to obtain at least 100 repetitions. A histogram of these values of \bar{x} is a better approximation to the sampling distribution.)

- 3.74 A sampling applet experiment.** The *Simple Random Sampling* applet available at www.whfreeman.com/pbs can animate the idea of a sampling distribution. Form a population labeled 1 to 100. We will choose an SRS of 10 of these numbers. That is, in this exercise the numbers themselves are the population, not just labels for 100 individuals. The proportion of the whole numbers 1 to 100 that are equal to or less than 60 is $p = 0.6$. This is the population proportion.

- Use the applet to choose an SRS of size 25. Which 25 numbers were chosen? Count the numbers ≤ 60 in your sample and divide this count by 25. This is the sample proportion \hat{p} .
- Although the population and the parameter $p = 0.6$ remain fixed, the sample proportion changes as we take more samples. Take another SRS



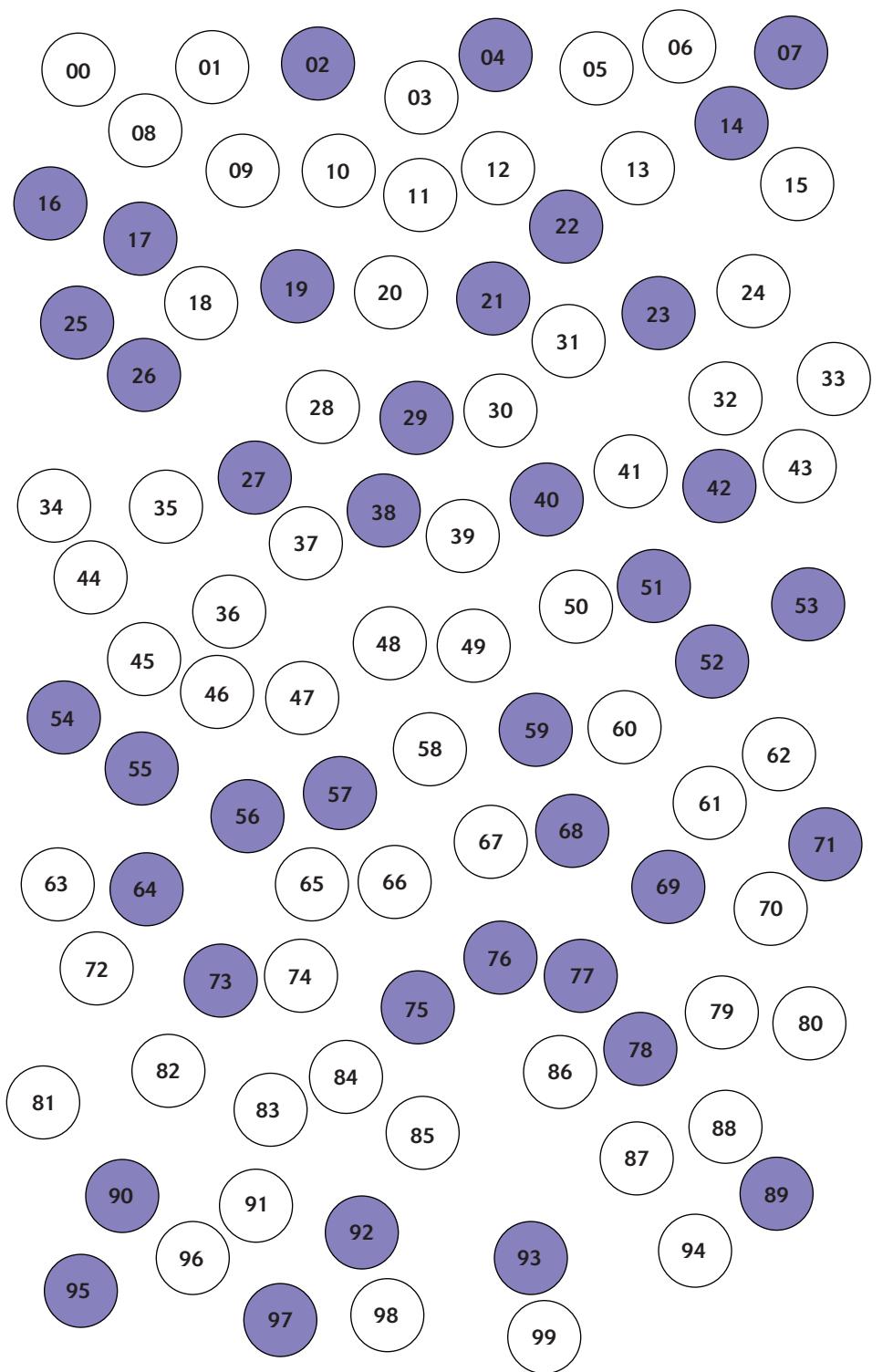


FIGURE 3.15 A population of 100 individuals for Exercise 3.75. Some individuals (white circles) find clothes shopping frustrating, and the others do not.

of size 25. (Use the “Reset” button to return to the original population before taking the second sample.) What are the 25 numbers in your sample? What proportion of these is ≤ 60 ? This is another value of \hat{p} .

- (c) Take 8 more SRSs from this same population and record their sample proportions. You now have 10 values of the statistic \hat{p} from 10 SRSs of the same size from the same population. Make a histogram of the 10 values and mark the population parameter $p = 0.6$ on the horizontal axis. Are your 10 sample values roughly centered at the population value p ? (If you kept going forever, your \hat{p} -values would form the sampling distribution of the sample proportion; the population proportion p would indeed be the center of this distribution.)

- 3.75 **A sampling experiment.** Figures 3.7 and 3.8 show how the sample proportion \hat{p} behaves when we take many samples from a population in which the population proportion is $p = 0.6$. You can follow the steps in this process on a small scale.



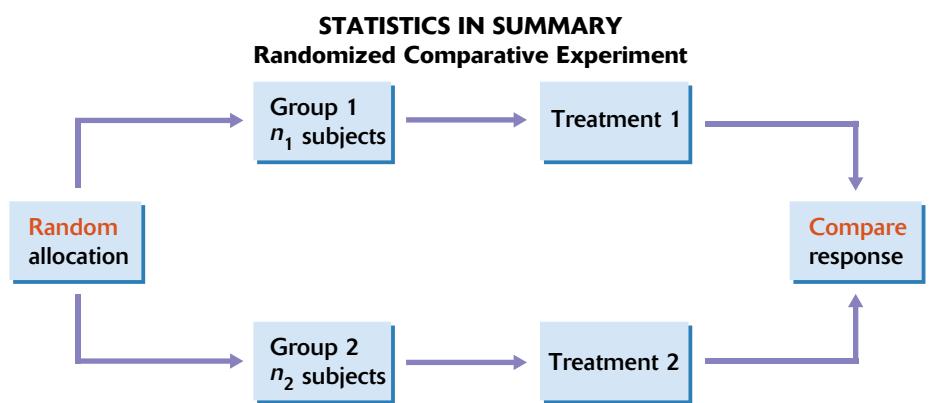
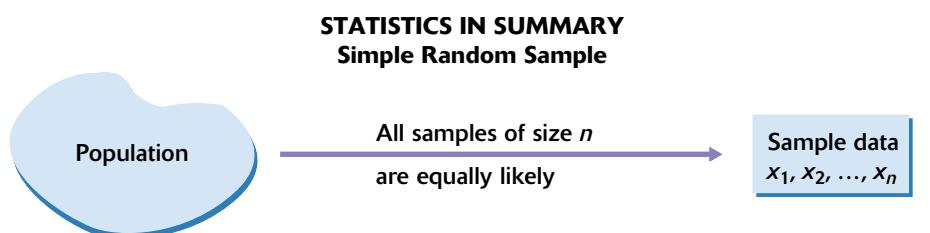
Figure 3.15 is a small population. Each circle represents an adult. The white circles are people who find clothes shopping frustrating, and the colored circles are people who do not. You can check that 60 of the 100 circles are white, so in this population the proportion who find shopping frustrating is $p = 60/100 = 0.6$.

- (a) The circles are labeled 00, 01, ..., 99. Use line 101 of Table B to draw an SRS of size 5. What is the proportion \hat{p} of the people in your sample who find shopping frustrating?
- (b) Take 9 more SRSs of size 5 (10 in all), using lines 102 to 110 of Table B, a different line for each sample. You now have 10 values of the sample proportion \hat{p} .
- (c) Because your samples have only 5 people, the only values \hat{p} can take are $0/5, 1/5, 2/5, 3/5, 4/5$, and $5/5$. That is, \hat{p} is always 0, 0.2, 0.4, 0.6, 0.8, or 1. Mark these numbers on a line and make a histogram of your 10 results by putting a bar above each number to show how many samples had that outcome.
- (d) Taking samples of size 5 from a population of size 100 is not a practical setting, but let’s look at your results anyway. How many of your 10 samples estimated the population proportion $p = 0.6$ exactly correctly? Is the true value 0.6 roughly in the center of your sample values? Explain why 0.6 would be in the center of the sample values if you took a large number of samples.

STATISTICS IN SUMMARY

Designs for producing data are essential parts of statistics in practice. The Statistics in Summary figure (on the next page) displays the big ideas visually. Random sampling and randomized comparative experiments are perhaps the most important statistical inventions of the twentieth century. Both were slow to gain acceptance, and you will still see many voluntary response samples and uncontrolled experiments. This chapter has explained good techniques for producing data and has also explained why bad techniques often produce worthless data.

The deliberate use of chance in producing data is a central idea in statistics. It allows use of the laws of probability to analyze data, as we will



see in the following chapters. Here are the major skills you should have now that you have studied this chapter.

A. SAMPLING

1. Identify the population in a sampling situation.
2. Recognize bias due to voluntary response samples and other inferior sampling methods.
3. Use software or Table B of random digits to select a simple random sample (SRS) from a population.
4. Recognize the presence of undercoverage and nonresponse as sources of error in a sample survey. Recognize the effect of the wording of questions on the responses.
5. Use random digits to select a stratified random sample from a population when the strata are identified.

B. EXPERIMENTS

1. Recognize whether a study is an observational study or an experiment.
2. Recognize bias due to confounding of explanatory variables with lurking variables in either an observational study or an experiment.
3. Identify the factors (explanatory variables), treatments, response variables, and individuals or subjects in an experiment.

4. Outline the design of a completely randomized experiment using a diagram like those in Figures 3.4 and 3.5. The diagram in a specific case should show the sizes of the groups, the specific treatments, and the response variable.
5. Use Table B of random digits to carry out the random assignment of subjects to groups in a completely randomized experiment.
6. Recognize the placebo effect. Recognize when the double-blind technique should be used. Be aware of weaknesses in an experiment, especially lack of realism that makes it hard to generalize conclusions.
7. Explain why a randomized comparative experiment can give good evidence for cause-and-effect relationships.

C. TOWARD INFERENCE

1. Identify parameters and statistics in a sample or experiment.
2. Recognize the fact of sampling variability: a statistic will take different values when you repeat a sample or experiment.
3. Interpret a sampling distribution as describing the values taken by a statistic in all possible repetitions of a sample or experiment under the same conditions.
4. Understand the effect of sample size on the variability of sample statistics.

CHAPTER 3 REVIEW EXERCISES

- 3.76 **Physical fitness and leadership.** A study of the relationship between physical fitness and leadership uses as subjects middle-aged executives who have volunteered for an exercise program. The executives are divided into a low-fitness group and a high-fitness group on the basis of a physical examination. All subjects then take a psychological test designed to measure leadership, and the results for the two groups are compared. Is this an observational study or an experiment? Explain your answer.
- 3.77 **Taste testing.** Before a new variety of frozen muffins is put on the market, it is subjected to extensive taste testing. People are asked to taste the new muffin and a competing brand and to say which they prefer. (Both muffins are unidentified in the test.) Is this an observational study or an experiment? Why?
- 3.78 **A hot fund.** A large mutual-fund group assigns a young securities analyst to manage its small biotechnology stock fund. The fund's share value increases an impressive 43% during the first year under the new manager. Explain why this performance does not necessarily establish the manager's ability.
- 3.79 **Are you sure?** Spain's Centro de Investigaciones Sociológicos carried out a sample survey on the attitudes of Spaniards toward private business and state intervention in the economy.²³ Of the 2496 adults interviewed, 72% agreed that "employees with higher performance must get higher pay." On the other hand, 71% agreed that "everything a society produces should be distributed among its members as equally as possible and there

should be no major differences.” Use these conflicting results as an example in a short explanation of why opinion polls often fail to reveal public attitudes clearly.

- 3.80 **Daytime running lights.** Canada requires that cars be equipped with “daytime running lights,” headlights that automatically come on at a low level when the car is started. Some manufacturers are now equipping cars sold in the United States with running lights. Will running lights reduce accidents by making cars more visible?
- Briefly discuss the design of an experiment to help answer this question. In particular, what response variables will you examine?
 - Example 3.17 (page 198) discusses center brake lights. What cautions do you draw from that example that apply to an experiment on the effects of running lights?
- 3.81 **Learning about markets.** Your economics professor wonders if playing market games online will help students understand how markets set prices. You suggest an experiment: have some students use the online games, while others discuss markets in recitation sections. The course has two lectures, at 8:30 a.m. and 2:30 p.m. There are 10 recitation sections attached to each lecture. The students are already assigned to recitations. For practical reasons, all students in each recitation must follow the same program.
- The professor says, “Let’s just have the 8:30 group do online work in recitation and the 2:30 group do discussion.” Why is this a bad idea?
 - Outline the design of an experiment with the 20 recitation sections as individuals. Carry out your randomization and include in your outline the recitation numbers assigned to each treatment.
- 3.82 **How much do students earn?** A university’s financial aid office wants to know how much it can expect students to earn from summer employment. This information will be used to set the level of financial aid. The population contains 3478 students who have completed at least one year of study but have not yet graduated. The university will send a questionnaire to an SRS of 100 of these students, drawn from an alphabetized list.
- Describe how you will label the students in order to select the sample.
 - Use Table B, beginning at line 105, to select the first 5 students in the sample.
- 3.83 **Sampling entrepreneurs.** A group of business school researchers wanted information on “why some firms survive while other firms with equal economic performance do not.” Here are some bare facts about the study’s data. Write a brief discussion of the difficulty of gathering data from this population, using this study as an example. Include nonresponse rates in percents.
- The researchers sent approximately 13,000 questionnaires to members of the National Federation of Independent Businesses who reported that they had recently become business owners. Responses were obtained from 4814 entrepreneurs, of whom 2294 had become owners during the preceding 17 months. Followup questionnaires were sent to these 2294 people after one year and after two years. After two years, 963 replied that their firms had survived and 171 firms had been sold. From other sources, the authors could identify 611 firms that no longer existed. The remaining*

businesses did not respond and could not be identified as discontinued or sold. The study then compared the 963 surviving firms with the 611 known to have failed.²⁴

- 3.84 **The price of digital cable TV.** A cable television provider plans to introduce “digital cable” to its current market. Digital cable offers more channels and features than standard cable television, but at a higher price. The provider wants to gauge how many consumers would sign up for digital service. You must plan a sample survey of households in the company’s service area. You feel that these groups may respond differently:

- Households that currently subscribe to standard cable television service.
- Households that currently subscribe to some other form of television service such as satellite television.
- Households that do not subscribe to any pay television service.

Briefly discuss the design of your sample.

- 3.85 **Sampling college faculty.** A labor organization wants to study the attitudes of college faculty members toward collective bargaining. These attitudes appear to be different depending on the type of college. The American Association of University Professors classifies colleges as follows:

- Class I. Offer doctorate degrees and award at least 15 per year.
- Class IIA. Award degrees above the bachelor's but are not in Class I.
- Class IIB. Award no degrees beyond the bachelor's.
- Class III. Two-year colleges.

Discuss the design of a sample of faculty from colleges in your state, with total sample size about 200.

- 3.86 **Sampling students.** You want to investigate the attitudes of students at your school about the labor practices of factories that make college-brand apparel. You have a grant that will pay the costs of contacting about 500 students.

- (a) Specify the exact population for your study. For example, will you include part-time students?
- (b) Describe your sample design. Will you use a stratified sample?
- (c) Briefly discuss the practical difficulties that you anticipate. For example, how will you contact the students in your sample?

- 3.87 **Did you vote?** When the Current Population Survey asked the adults in its sample of 55,000 households if they had voted in the 1996 presidential election, 54% said they had. In fact, only 49% of the adult population voted in that election. Why do you think the CPS result missed by much more than its margin of error?

- 3.88 **Treating drunk drivers.** Once a person has been convicted of drunk driving, one purpose of court-mandated treatment or punishment is to prevent future offenses of the same kind. Suggest three different treatments that a court might require. Then outline the design of an experiment to compare their effectiveness. Be sure to specify the response variables you will measure.

- 3.89 **Do antioxidants prevent cancer?** People who eat lots of fruits and vegetables have lower rates of colon cancer than those who eat little of these foods.

Fruits and vegetables are rich in “antioxidants” such as vitamins A, C, and E. Will taking antioxidants help prevent colon cancer? A clinical trial studied this question with 864 people who were at risk of colon cancer. The subjects were divided into four groups: daily beta carotene, daily vitamins C and E, all three vitamins every day, and daily placebo. After four years, the researchers were surprised to find no significant difference in colon cancer among the groups.²⁵

- (a) What are the explanatory and response variables in this experiment?
 - (b) Outline the design of the experiment. Use your judgment in choosing the group sizes.
 - (c) Assign labels to the 864 subjects and use Table B, starting at line 118, to choose the first 5 subjects for the beta carotene group.
 - (d) The study was double-blind. What does this mean?
 - (e) What does “no significant difference” mean in describing the outcome of the study?
 - (f) Suggest some lurking variables that could explain why people who eat lots of fruits and vegetables have lower rates of colon cancer. The experiment suggests that these variables, rather than the antioxidants, may be responsible for the observed benefits of fruits and vegetables.
- 3.90 **Stocks go down on Monday.** Puzzling but true: stocks tend to go down on Mondays, both in the United States and in overseas markets. There is no convincing explanation for this fact. A recent study looked at this “Monday effect” in more detail, using data on the daily returns of stocks on several U.S. exchanges over a 30-year period. Here are some of the findings:

To summarize, our results indicate that the well-known Monday effect is caused largely by the Mondays of the last two weeks of the month. The mean Monday return of the first three weeks of the month is, in general, not significantly different from zero and is generally significantly higher than the mean Monday return of the last two weeks. Our finding seems to make it more difficult to explain the Monday effect.²⁶

A friend thinks that “significantly” in this article has its plain English meaning, roughly “I think this is important.” Explain in simple language what “significantly higher” and “not significantly different from zero” actually tell us here.

- 3.91 **Potatoes for french fries.** Few people want to eat discolored french fries. Potatoes are kept refrigerated before being cut for french fries to prevent spoiling and preserve flavor. But immediate processing of cold potatoes causes discoloring due to complex chemical reactions. The potatoes must therefore be brought to room temperature before processing. Fast-food chains and other sellers of french fries must understand potato behavior. Design an experiment in which tasters will rate the color and flavor of french fries prepared from several groups of potatoes. The potatoes will be freshly harvested or stored for a month at room temperature or stored for a month refrigerated. They will then be sliced and cooked either immediately or after an hour at room temperature.
- (a) What are the factors and their levels, the treatments, and the response variables?
 - (b) Describe and outline the design of this experiment.

- (c) It is efficient to have each taster rate fries from all treatments. How will you use randomization in presenting fries to the tasters?
- 3.92 **Speeding the mail?** Statistical studies can often help service providers assess the quality of their service. The United States Postal Service is one such provider of services. We wonder if the number of days a letter takes to reach another city is affected by the time of day it is mailed and whether or not the zip code is used. Describe briefly the design of a two-factor experiment to investigate this question. Be sure to specify the treatments exactly and to tell how you will handle lurking variables such as the day of the week on which the letter is mailed.
- 3.93 **McDonald's versus Wendy's.** The food industry uses taste tests to improve products and for comparison with competitors. Do consumers prefer the taste of a cheeseburger from McDonald's or from Wendy's in a blind test in which neither burger is identified? Describe briefly the design of a matched pairs experiment to investigate this question. How will you be sure the comparison is "blind"?
- 3.94 **Design your own experiment.** The previous two exercises illustrate the use of statistically designed experiments to answer questions of interest to consumers as well as to businesses. Select a question of interest to you that an experiment might answer and briefly discuss the design of an appropriate experiment.
- 3.95 **Randomization at work.** To demonstrate how randomization reduces confounding, return to the nutrition experiment described in Example 3.14 (page 193). Label the 30 rats 01 to 30. Suppose that, unknown to the experimenter, the 10 rats labeled 01 to 10 have a genetic defect that will cause them to grow more slowly than normal rats. If the experimenter simply puts rats 01 to 15 in the experimental group and rats 16 to 30 in the control group, this lurking variable will bias the experiment against the new food product.
Use Table B to assign 15 rats at random to the experimental group as in Example 3.14. Record how many of the 10 rats with genetic defects are placed in the experimental group and how many are in the control group. Repeat the randomization using different lines in Table B until you have done five random assignments. What is the mean number of genetically defective rats in experimental and control groups in your five repetitions?
- Simulating samples.* Statistical software offers a shortcut to simulate the results of, say, 100 SRSs of size n drawn from a population in which proportion p would say "Yes" to a certain question. We will see in Chapter 5 that "binomial" is the key word to look for in the software menus. Set the sample size n and the sample proportion p . The software will generate the number of "Yes" answers in an SRS. You can divide by n to get the sample proportion \hat{p} . Use simulation to complete Exercises 3.96 and 3.97.
- 3.96 **Changing the population.** Draw 100 samples of size $n = 50$ from populations with $p = 0.1$, $p = 0.3$, and $p = 0.5$. Make a stemplot of the 100 values of \hat{p} obtained in each simulation. Compare your three stemplots. Do they show about the same variability? How does changing the parameter p affect the sampling distribution? If your software permits, make a Normal quantile plot of the \hat{p} -values from the population with $p = 0.5$. Is the sampling distribution approximately Normal?

- 3.97 **Changing the sample size.** Draw 100 samples of each of the sizes $n = 50$, $n = 200$, and $n = 800$ from a population with $p = 0.6$. Make a histogram of the \hat{p} -values for each simulation, using the same horizontal and vertical scales so that the three graphs can be compared easily. How does increasing the size of an SRS affect the sampling distribution of \hat{p} ?



CHAPTER 3 CASE STUDY EXERCISES

CASE STUDY 3.1: Consumer preferences in car colors. The most popular colors for cars and light trucks sold in recent years are white and silver. A survey shows that 17.6% of cars and trucks made in North America in 2000 were silver and 17.2% were white.²⁷ (“White” includes various shades of “pearl” and “cream.”) The preferences of some groups of consumers may of course differ from national patterns. Undertake a study to determine the most popular colors among the vehicles driven by students at your school. You might collect data by questioning a sample of students or by looking at cars in student parking areas. Explain carefully how you attempted to get data that are close to an SRS of student cars, including how you used random selection. Then report your findings on student preferences in motor vehicle colors.

CASE STUDY 3.2: Individual incomes. The Data Appendix describes the file *individuals.dat*, which contains the total incomes of 55,899 people between the ages of 25 and 65. We will illustrate sampling variability and sampling distributions in a case study that considers these people as a population. This is similar to sampling the employees of a major corporation. For example, Microsoft employs roughly as many people as are represented in this data file.

- A. **The population.** Describe the distribution of the 55,899 incomes in the population. Include a histogram and the five-number summary.
- B. **Samples and sampling distributions.** Choose an SRS of 500 members from this population. Make a histogram of the 500 incomes in the sample and find the five-number summary. Briefly compare the shape, center, and spread of the income distributions in the sample and in the population. Then repeat the process of choosing an SRS of size 500 four more times (five in all). Does it seem reasonable to you from this small trial that an SRS of 500 people will usually produce a sample whose shape is generally representative of the population?
- C. **Statistical estimation.** Do the medians and quartiles of the samples provide reasonable estimates of the population median and quartiles? Explain why we expect that the minimum and maximum of a sample will *not* satisfactorily estimate the population minimum and maximum. Now examine estimation of mean income in more detail. Use your software to choose 50 SRSs of size 500 from this population. Find the mean income for each sample and save these 50 sample means in a separate file. Make a histogram of the distribution of the 50 sample means. How do the shape, center, and spread of this distribution compare with the distribution of individual incomes from part A? Does it appear that the sample mean from an SRS of size 500 is usually a reasonable estimator of the population mean? (The sampling distribution of the sample mean for samples of size 500 from this population is the distribution of the means of all possible samples. Your 50 samples give a rough idea of the nature of the sampling distribution.)

Probability and Inference

PART 2

CHAPTER 4

Probability and Sampling Distributions

CHAPTER 5

Probability Theory

CHAPTER 6

Introduction to Inference

CHAPTER 7

Inference for Distributions

CHAPTER 8

Inference for Proportions

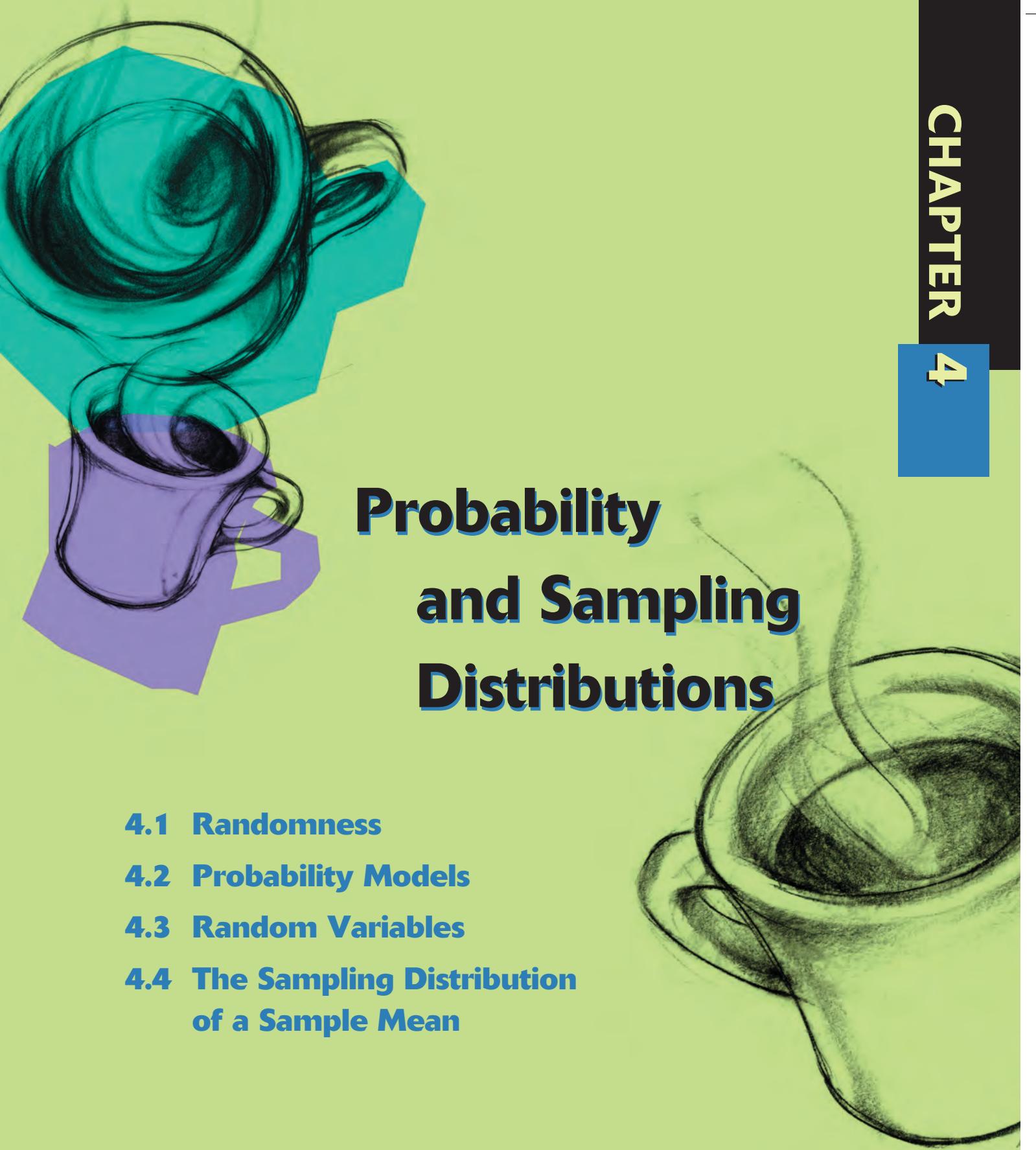
Prelude

Of audits and probabilities

A firm reimburses its employees for business-related expenses like meals with clients. To obtain a reimbursement an employee must file an expense voucher with receipts for expenses of \$25 or more. A look at the vouchers shows suspiciously many expenses in the \$24.00 to \$24.99 range. Are employees padding their vouchers by claiming amounts just a bit too small to require a receipt?

A bank allows individual officers to decide if credit card balances up to \$5000 should be written off as uncollectible. Write-offs above \$5000 must be approved by a supervisor. An auditor finds more write-offs than expected in the \$4500–\$4999 range. An investigation finds that some bank officers have been writing off credit debts of friends who ran up credit card bills to just below the \$5000 limit.

In both scenarios, the auditor must have a legitimate way to establish a cutoff for how many is “too many.” How many \$24 expense vouchers should the firm expect? How many credit card balance write-offs in the \$4500 range should the bank expect? Only if questions like these are answered can a company recognize when something suspicious is taking place. We can use a *probability model* to specify how often different values occur when we sample from a large collection of data. One particularly striking probability model describes the distribution of first digits in large collections of numbers such as claimed expenses and credit card account write-offs. This is *Benford’s Law*, the topic of our first case in this chapter.



CHAPTER

4

Probability and Sampling Distributions

- 4.1 Randomness**
- 4.2 Probability Models**
- 4.3 Random Variables**
- 4.4 The Sampling Distribution
of a Sample Mean**



Introduction

The reasoning of statistical inference rests on asking, “How often would this method give a correct answer if I used it very many times?” Inference is most secure when we produce data by random sampling or randomized comparative experiments. The reason is that when we use chance to choose respondents or assign subjects, the laws of probability answer the question “What would happen if we did this many times?” The purpose of this chapter is to see what the laws of probability tell us, but without going into the mathematics of probability theory.

4.1 Randomness

What is the mean income of households in the United States? The Bureau of Labor Statistics contacted a random sample of 55,000 households in March 2001 for the Current Population Survey. The mean income of the 55,000 households for the year 2000 was $\bar{x} = \$57,045$.¹ That \$57,045 is a *statistic* that describes the CPS sample households. We use it to estimate an unknown *parameter*, the mean income of all 106 million American households. We know that \bar{x} would take several different values if the Bureau of Labor Statistics had taken several samples in March 2001. Fortunately, we also know that this sampling variability follows a regular pattern that can tell us how accurate the sample result is likely to be. That pattern obeys the laws of probability. Our starting point in understanding probability is the phenomenon we observed in the simulations of Section 3.3: **chance behavior is unpredictable in the short run but has a regular and predictable pattern in the long run.**

The idea of probability

Toss a coin, or choose an SRS. The result can't be predicted in advance, because the result will vary when you toss the coin or choose the sample repeatedly. But there is still a regular pattern in the results, a pattern that emerges clearly only after many repetitions. This remarkable fact is the basis for the idea of probability.

EXAMPLE 4.1

Coin tossing

When you toss a coin, there are only two possible outcomes, heads or tails. Figure 4.1 shows the results of tossing a coin 5000 times twice. For each number of tosses from 1 to 5000, we have plotted the proportion of those tosses that gave a head. Trial A (solid line) begins tail, head, tail, tail. You can see that the proportion of heads for Trial A starts at 0 on the first toss, rises to 0.5 when the second toss gives a head, then falls to 0.33 and 0.25 as we get two more tails. Trial B, on the other hand, starts with five straight heads, so the proportion of heads is 1 until the sixth toss.

The proportion of tosses that produce heads is quite variable at first. Trial A starts low and Trial B starts high. As we make more and more tosses, however, the

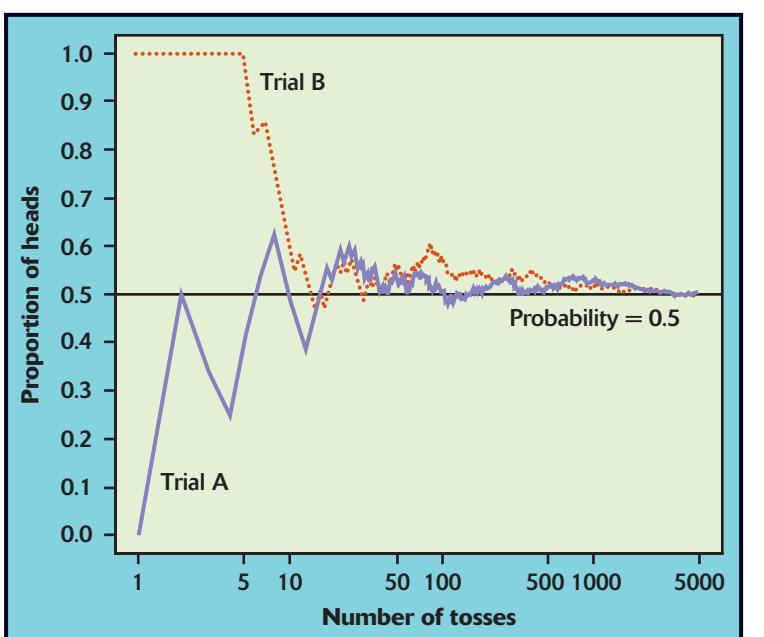


FIGURE 4.1 The proportion of tosses of a coin that give a head changes as we make more tosses. Eventually, however, the proportion approaches 0.5, the probability of a head. This figure shows the results of two trials of 5000 tosses each.

proportion of heads for both trials gets close to 0.5 and stays there. If we made yet a third trial at tossing the coin a great many times, the proportion of heads would again settle down to 0.5 in the long run. We say that 0.5 is the *probability* of a head. The probability 0.5 appears as a horizontal line on the graph.



The *What Is Probability?* applet available on the Web site for this book, www.whfreeman.com/pbs, animates Figure 4.1. It allows you to choose the probability of a head and simulate any number of tosses of a coin with that probability. Experience shows that the proportion of heads gradually settles down close to the probability. Equally important, it also shows that the proportion in a small or moderate number of tosses can be far from the probability. Probability describes *only* what happens in the long run.

“Random” in statistics is not a synonym for “haphazard” but a description of a kind of order that emerges only in the long run. We often encounter the unpredictable side of randomness in our everyday experience, but we rarely see enough repetitions of the same random phenomenon to observe the long-term regularity that probability describes. You can see that regularity emerging in Figure 4.1. In the very long run, the proportion of tosses that give a head is 0.5. This is the intuitive idea of probability. Probability 0.5 means “occurs half the time in a very large number of trials.”

We might suspect that a coin has probability 0.5 of coming up heads just because the coin has two sides. As Exercises 4.1 and 4.2 illustrate, such suspicions are not always correct. The idea of probability is *empirical*. That

is, it is based on observation rather than theorizing. Probability describes what happens in very many trials, and we must actually observe many trials to pin down a probability. In the case of tossing a coin, some diligent people have in fact made thousands of tosses.

EXAMPLE 4.2**Some coin tossers**

The French naturalist Count Buffon (1707–1788) tossed a coin 4040 times. Result: 2048 heads, or proportion $2048/4040 = 0.5069$ for heads.

Around 1900, the English statistician Karl Pearson heroically tossed a coin 24,000 times. Result: 12,012 heads, a proportion of 0.5005.

While imprisoned by the Germans during World War II, the South African mathematician John Kerrich tossed a coin 10,000 times. Result: 5067 heads, a proportion of 0.5067.

**RANDOMNESS AND PROBABILITY**

We call a phenomenon **random** if individual outcomes are uncertain but there is nonetheless a regular distribution of outcomes in a large number of repetitions.

The **probability** of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions.

APPLY YOUR KNOWLEDGE

- 4.1 **Nickels spinning.** Hold a nickel upright on its edge under your forefinger on a hard surface, then snap it with your other forefinger so that it spins for some time before falling. Based on 50 spins, estimate the probability of heads.
- 4.2 **Nickels falling over.** You may feel that it is obvious that the probability of a head in tossing a coin is about 1/2 because the coin has two faces. Such opinions are not always correct. The previous exercise asked you to spin a nickel rather than toss it—that changes the probability of a head. Now try another variation. Stand a nickel on edge on a hard, flat surface. Pound the surface with your hand so that the nickel falls over. What is the probability that it falls with heads upward? Make at least 50 trials to estimate the probability of a head.

Thinking about randomness

That some things are random is an observed fact about the world. The outcome of a coin toss, the time between customers arriving at an ATM machine, and the price of a share of a company's stock at the close of the market are all random. So is the outcome of a random sample or a randomized experiment. Probability theory is the branch of mathematics that describes random behavior. Of course, we can never observe a probability exactly. We could always continue tossing the coin, for example.

Mathematical probability is an idealization based on imagining what would happen in an indefinitely long series of trials.

The best way to understand randomness is to observe random behavior—not only the long-run regularity but the unpredictable results of short runs. You can do this with physical devices, as in Exercises 4.1 to 4.5, but computer simulations (imitations) of random behavior allow faster exploration. Exercises 4.9, 4.10, and 4.11 suggest some simulations of random behavior. As you explore randomness, remember:

independence

- You must have a long series of **independent** trials. That is, the outcome of one trial must not influence the outcome of any other. Imagine a crooked gambling house where the operator of a roulette wheel can stop it where she chooses—she can prevent the proportion of “red” from settling down to a fixed number. These trials are not independent.
- The idea of probability is empirical. Computer simulations start with given probabilities and imitate random behavior, but we can estimate a real-world probability only by actually observing many trials.
- Nonetheless, computer simulations are very useful because we need long runs of trials. In situations such as coin tossing, the proportion of an outcome often requires several hundred trials to settle down to the probability of that outcome. The kinds of physical techniques suggested in the exercises are too slow for this. Short runs give only rough estimates of a probability.

SECTION 4.1 SUMMARY

- A **random phenomenon** has outcomes that we cannot predict but that nonetheless have a regular distribution in very many repetitions.
- The **probability** of an event is the proportion of times the event occurs in many repeated trials of a random phenomenon.

SECTION 4.1 EXERCISES

- 4.3 **Random digits.** The table of random digits (Table B) was produced by a random mechanism that gives each digit probability 0.1 of being a 0. What proportion of the first 200 digits in the table are 0s? This proportion is an estimate, based on 200 repetitions, of the true probability, which in this case is known to be 0.1.
- 4.4 **How many tosses to get a head?** When we toss a penny, experience shows that the probability (long-term proportion) of a head is close to 1/2. Suppose now that we toss the penny repeatedly until we get a head. What is the probability that the first head comes up in an odd number of tosses (1, 3, 5, and so on)? To find out, repeat this experiment 50 times, and keep a record of the number of tosses needed to get a head on each of your 50 trials.
- From your experiment, estimate the probability of a head on the first toss. What value should we expect this probability to have?
 - Use your results to estimate the probability that the first head appears on an odd-numbered toss.

- 4.5 **Tossing a thumbtack.** Toss a thumbtack on a hard surface 100 times. How many times did it land with the point up? What is the approximate probability of landing point up?
- 4.6 **Three of a kind.** You read in a book on poker that the probability of being dealt three of a kind in a five-card poker hand is $1/50$. Explain in simple language what this means.
- 4.7 Probability is a measure of how likely an event is to occur. Match one of the probabilities that follow with each statement of likelihood given. (The probability is usually a more exact measure of likelihood than is the verbal statement.)

0 0.01 0.3 0.6 0.99 1

- (a) This event is impossible. It can never occur.
- (b) This event is certain. It will occur on every trial.
- (c) This event is very unlikely, but it will occur once in a while in a long sequence of trials.
- (d) This event will occur more often than not.



- 4.8 **What probability doesn't say.** The idea of probability is that the *proportion* of heads in many tosses of a balanced coin eventually gets close to 0.5. But does the actual *count* of heads get close to one-half the number of tosses? Let's find out. Set the "Probability of heads" in the *What Is Probability?* applet to 0.5 and the number of tosses to 40. You can extend the number of tosses by clicking "Toss" again to get 40 more. Don't click "Reset" during this exercise.

- (a) After 40 tosses, what is the proportion of heads? What is the count of heads? What is the difference between the count of heads and 20 (one-half the number of tosses)?
- (b) Keep going to 120 tosses. Again record the proportion and count of heads and the difference between the count and 60 (half the number of tosses).
- (c) Keep going. Stop at 240 tosses and again at 480 tosses to record the same facts. Although it may take a long time, the laws of probability say that the proportion of heads will always get close to 0.5 and also that the difference between the count of heads and half the number of tosses will always grow without limit.



- 4.9 **Simulating consumer behavior.** About half of the customers entering a local computer store will purchase a new computer before leaving the store. Use the *What Is Probability?* applet or your statistical software to simulate 100 customers independently entering the store with each having a probability of 0.5 of purchasing a new computer on this visit to the store. (In most software, the key phrase to look for is "Bernoulli trials." This is the technical term for independent trials with Yes/No outcomes. Our outcomes here are "buy" and "not buy.")
- (a) What percent of the 100 simulated customers bought a new computer?
 - (b) Examine the sequence of "buys" and "not buys." How long was the longest run of "buys"? Of "not buys"? (Sequences of random outcomes often show runs longer than our intuition thinks likely.)



- 4.10 **Simulating an opinion poll.** A recent opinion poll showed that about 65% of the American public have a favorable opinion of the software company Microsoft.² Suppose that this is exactly true. Choosing a person at random then has probability 0.65 of getting one who has a favorable opinion of Microsoft. Use the *What Is Probability?* applet or your statistical software to simulate choosing many people independently. (In most software, the key phrase to look for is “Bernoulli trials.” This is the technical term for independent trials with Yes/No outcomes. Our outcomes here are “favorable” or not.)
- Simulate drawing 20 people, then 80 people, then 320 people. What proportion has a favorable opinion of Microsoft in each case? We expect (but because of chance variation we can’t be sure) that the proportion will be closer to 0.65 in longer runs of trials.
 - Simulate drawing 20 people 10 times and record the percents in each trial who have a favorable opinion of Microsoft. Then simulate drawing 320 people 10 times and again record the 10 percents. Which set of 10 results is less variable? We expect the results of 320 trials to be more predictable (less variable) than the results of 20 trials. That is “long-run regularity” showing itself.
- 4.11 **More efficient simulation.** Continue the exploration begun in Exercise 4.10. Software allows you to simulate many independent “Yes/No” trials more quickly if all you want to save is the count of “Yes” outcomes. The key word “Binomial” simulates n independent Bernoulli trials, each with probability p of a “Yes,” and records just the count of “Yes” outcomes.
- Simulate 100 draws of 20 people from the population in Exercise 4.10. Record the number who have a favorable opinion of Microsoft on each draw. What is the approximate probability that out of 20 people drawn at random at least 14 have a favorable opinion of Microsoft?
 - Convert the “favorable” counts into percents of the 20 people in each trial. Make a histogram of these 100 percents. Describe the shape, center, and spread of this distribution.
 - Now simulate drawing 320 people. Do this 100 times and record the percent who respond “favorable” on each of the 100 draws. Make a histogram of the percents and describe the shape, center, and spread of the distribution.
 - In what ways are the distributions in parts (b) and (c) alike? In what ways do they differ? (Because regularity emerges in the long run, we expect the results of drawing 320 individuals to be less variable than the results of drawing 20 individuals.)

4.2 Probability Models



UNCOVERING FRAUD BY DIGITAL ANALYSIS

“Digital analysis” is one of the big new tools among auditors and investigators looking for fraud. Faked numbers in tax returns, payment records, invoices, expense account claims, and many other settings often display patterns that aren’t present in legitimate records. Some patterns, like too many round

numbers, are obvious and easily avoided by a clever crook. Others are more subtle. It is a striking fact that the first digits of numbers in legitimate records often follow a distribution known as *Benford's Law*. Here it is:

First digit	1	2	3	4	5	6	7	8	9
Proportion	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

These proportions are in fact *probabilities*. That is, they are the proportions specified by Benford's Law if we examine a very large number of similar records. Set a computer to work comparing invoices from your company's vendors with this distribution. Aha—here is a vendor whose invoices show a very different pattern. All the invoices were approved by the same manager. Confronted, he admits that he was raising the amounts of the invoices and buying from his wife's company at the inflated amounts. Digital analysis has uncovered another case of fraud.

Of course, not all sets of data follow Benford's Law. Numbers that are assigned, such as Social Security numbers, do not. Nor do data with a fixed maximum, such as deductible contributions to individual retirement accounts (IRAs). Nor of course do random numbers. But a surprising variety of data from natural science, social affairs, and business obeys this distribution.³



Gamblers have known for centuries that the fall of coins, cards, and dice displays clear patterns in the long run. Benford's Law says that first digits also follow a clear pattern. The idea of probability rests on the observed fact that the average result of many thousands of chance outcomes can be known with near certainty. How can we give a mathematical description of long-run regularity?

To see how to proceed, think first about a very simple random phenomenon, tossing a coin once. When we toss a coin, we cannot know the outcome in advance. What do we know? We are willing to say that the outcome will be either heads or tails. We believe that each of these outcomes has probability 1/2. This description of coin tossing has two parts:

- a list of possible outcomes
- a probability for each outcome

This description is the basis for all probability models. Here is the vocabulary we use.

PROBABILITY MODELS

The **sample space** of a random phenomenon is the set of all possible outcomes. S is used to denote sample space.

An **event** is an outcome or a set of outcomes of a random phenomenon. That is, an event is a subset of the sample space.

A **probability model** is a mathematical description of a random phenomenon consisting of two parts: a sample space S and a way of assigning probabilities to events.

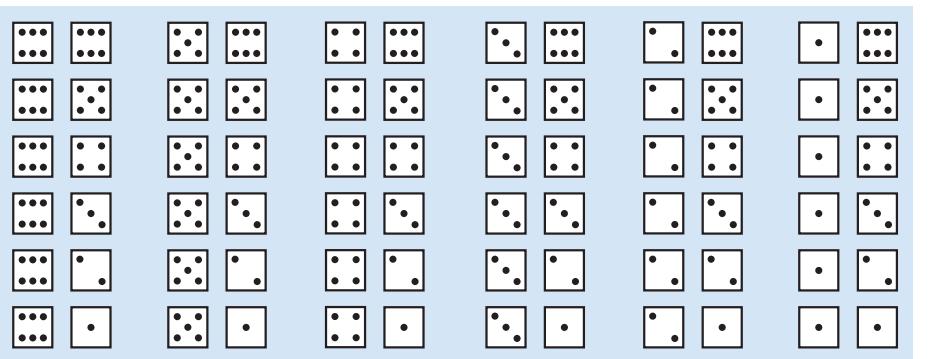


FIGURE 4.2 The 36 possible outcomes in rolling two dice.

The sample space S can be very simple or very complex. When we toss a coin once, there are only two outcomes, heads and tails. The sample space is $S = \{H, T\}$. If we draw a random sample of 55,000 U.S. households, as the Current Population Survey does, the sample space contains all possible choices of 55,000 of the 106 million households in the country. This S is extremely large. Each member of S is a possible sample, which explains the term *sample space*.

EXAMPLE 4.3

Rolling dice

Rolling two dice is a common way to lose money in casinos. There are 36 possible outcomes when we roll two dice and record the up-faces in order (first die, second die). Figure 4.2 displays these outcomes. They make up the sample space S . “Roll a 5” is an event, call it A , that contains four of these 36 outcomes:

$$A = \{ \begin{array}{c} \bullet \\ \square \end{array}, \begin{array}{c} \bullet \\ \bullet \end{array}, \begin{array}{c} \bullet \\ \circ \end{array}, \begin{array}{c} \circ \\ \bullet \end{array}, \begin{array}{c} \circ \\ \circ \end{array}, \begin{array}{c} \bullet \\ \circ \end{array} \}$$

In craps and other games, all that matters is the *sum* of the spots on the up-faces. Let’s change the random outcomes we are interested in: roll two dice and count the spots on the up-faces. Now there are only 11 possible outcomes, from a sum of 2 for rolling a double one through 3, 4, 5, and on up to 12 for rolling a double six. The sample space is now

$$S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

Comparing this S with Figure 4.2 reminds us that we can change S by changing the detailed description of the random phenomenon we are describing.

APPLY YOUR KNOWLEDGE

- 4.12 In each of the following situations, describe a sample space S for the random phenomenon. In some cases, you have some freedom in your choice of S .
- A new business is started. After two years, it is either still in business or it has closed.
 - A rust prevention treatment is applied to a new car. The response variable is the length of time before rust begins to develop on the vehicle.

- (c) A student enrolls in a statistics course and at the end of the semester receives a letter grade.
 - (d) A quality inspector examines four portable CD players and rates each as either “acceptable” or “unacceptable.” You record the sequence of ratings.
 - (e) A quality inspector examines four portable CD players and rates each as either “acceptable” or “unacceptable.” You record the number of units rated “acceptable.”
- 4.13 In each of the following situations, describe a sample space S for the random phenomenon. In some cases you have some freedom in specifying S , especially in setting the largest and the smallest value in S .
- (a) Choose a student in your class at random. Ask how much time that student spent studying during the past 24 hours.
 - (b) The Physicians’ Health Study asked 11,000 physicians to take an aspirin every other day and observed how many of them had a heart attack in a 5-year period.
 - (c) In a test of a new package design, you drop a carton of a dozen eggs from a height of 1 foot and count the number of broken eggs.
 - (d) Choose a Fortune 500 company at random. Look up how much the CEO of the company makes annually.
 - (e) A nutrition researcher feeds a new diet to a young male white rat. The response variable is the weight (in grams) that the rat gains in 8 weeks.

Probability rules

The true probability of any outcome—say, “roll a 5 when we toss two dice”—can be found only by actually tossing two dice many times, and then only approximately. How then can we describe probability mathematically? Rather than try to give “correct” probabilities, we start by laying down facts that must be true for any assignment of probabilities. These facts follow from the idea of probability as “the long-run proportion of repetitions in which an event occurs.”

1. **Any probability is a number between 0 and 1.** Any proportion is a number between 0 and 1, so any probability is also a number between 0 and 1. An event with probability 0 never occurs, and an event with probability 1 occurs on every trial. An event with probability 0.5 occurs in half the trials in the long run.
2. **All possible outcomes together must have probability 1.** Because some outcome must occur on every trial, the sum of the probabilities for all possible outcomes must be exactly 1.
3. **The probability that an event does not occur is 1 minus the probability that the event does occur.** If an event occurs in (say) 70% of all trials, it fails to occur in the other 30%. The probability that an event occurs and the probability that it does not occur always add to 100%, or 1.

4. If two events have no outcomes in common, the probability that one or the other occurs is the sum of their individual probabilities. If one event occurs in 40% of all trials, a different event occurs in 25% of all trials, and the two can never occur together, then one or the other occurs on 65% of all trials because $40\% + 25\% = 65\%$.

We can use mathematical notation to state Facts 1 to 4 more concisely. Capital letters near the beginning of the alphabet denote events. If A is any event, we write its probability as $P(A)$. Here are our probability facts in formal language. As you apply these rules, remember that they are just another form of intuitively true facts about long-run proportions.

PROBABILITY RULES

Rule 1. The probability $P(A)$ of any event A satisfies $0 \leq P(A) \leq 1$.

Rule 2. If S is the sample space in a probability model, then $P(S) = 1$.

Rule 3. The probability that an event A does not occur is

$$P(A \text{ does not occur}) = 1 - P(A)$$

Rule 4. Two events A and B are **disjoint** if they have no outcomes in common and so can never occur simultaneously. If A and B are disjoint,

$$P(A \text{ or } B) = P(A) + P(B)$$

This is the **addition rule for disjoint events**.

EXAMPLE 4.4

Benford's Law

The probabilities assigned to first digits by Benford's Law are all between 0 and 1. They add to 1 because the first digits listed make up the sample space S . (Note that a first digit can't be 0.)

The probability that a first digit is anything other than 1 is, by Rule 3,

$$\begin{aligned} P(\text{not a 1}) &= 1 - P(\text{first digit is 1}) \\ &= 1 - 0.301 = 0.699 \end{aligned}$$

That is, if 30.1% of first digits are 1s, the other 69.9% are not 1s.

"First digit is a 1" and "first digit is a 2" are disjoint events. So the addition rule (Rule 4) says

$$\begin{aligned} P(1 \text{ or } 2) &= P(1) + P(2) \\ &= 0.301 + 0.176 = 0.477 \end{aligned}$$

About 48% of first digits in data governed by Benford's Law are 1s or 2s. Fraudulent records generally have many fewer 1s and 2s.



EXAMPLE 4.5**Probabilities for rolling dice**

Figure 4.2 displays the 36 possible outcomes of rolling two dice. What probabilities shall we assign to these outcomes?

Casino dice are carefully made. Their spots are not hollowed out, which would give the faces different weights, but are filled with white plastic of the same density as the colored plastic of the body. For casino dice it is reasonable to assign the same probability to each of the 36 outcomes in Figure 4.2. Because all 36 outcomes together must have probability 1 (Rule 2), each outcome must have probability 1/36.

What is the probability of rolling a 5? Because the event “roll a 5” contains the four outcomes displayed in Example 4.3, the addition rule (Rule 4) says that its probability is

$$\begin{aligned} P(\text{roll a 5}) &= P(\begin{array}{|c|c|}\hline \bullet & \bullet\bullet \\ \hline \end{array}) + P(\begin{array}{|c|c|}\hline \bullet\bullet & \bullet \\ \hline \end{array}) + P(\begin{array}{|c|c|}\hline \bullet\bullet & \bullet\bullet \\ \hline \end{array}) + P(\begin{array}{|c|c|}\hline \bullet\bullet & \bullet \\ \hline \end{array}) \\ &= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} \\ &= \frac{4}{36} = 0.111 \end{aligned}$$

What about the probability of rolling a 7? In Figure 4.2 you will find six outcomes for which the sum of the spots is 7. The probability is 6/36, or about 0.167.

APPLY YOUR KNOWLEDGE

- 4.14 **Moving up.** An economist studying economic class mobility finds that the probability that the son of a lower-class father remains in the lower class is 0.46. What is the probability that the son moves to one of the higher classes?
- 4.15 **Causes of death.** Government data on job-related deaths assign a single occupation for each such death that occurs in the United States. The data on occupational deaths in 1999 show that the probability is 0.134 that a randomly chosen death was agriculture-related, and 0.119 that it was manufacturing-related. What is the probability that a death was either agriculture-related or manufacturing-related? What is the probability that the death was related to some other occupation?
- 4.16 **Rating the economy.** A Gallup Poll (June 11–17, 2001) interviewed a random sample of 1004 adults (18 years or older). The people in the sample were asked how they would rate economic conditions in the United States today. Here are the results:

Outcome	Probability
Excellent	0.03
Good	0.39
Fair	?
Poor	0.12
No opinion	0.01

These proportions are probabilities for the random phenomenon of choosing an adult at random and asking the person's opinion on current economic conditions.

- What must be the probability that the person chosen says "fair"? Why?
- The official press release focused on the percent of adults giving the economy a "positive" rating where positive is defined as "good" or "excellent." What is this probability?

Assigning probabilities: finite number of outcomes

Examples 4.4 and 4.5 illustrate one way to assign probabilities to events: assign a probability to every individual outcome, then add these probabilities to find the probability of any event. If such an assignment is to satisfy the rules of probability, the probabilities of all the individual outcomes must sum to exactly 1.

PROBABILITIES IN A FINITE SAMPLE SPACE

Assign a probability to each individual outcome. These probabilities must be numbers between 0 and 1 and must have sum 1.

The probability of any event is the sum of the probabilities of the outcomes making up the event.

EXAMPLE 4.6

Random digits versus Benford's Law



You might think that first digits in financial records are distributed "at random" among the digits 1 to 9. The 9 possible outcomes would then be equally likely. The sample space for a single digit is

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

Call a randomly chosen first digit X for short. The probability model for X is completely described by this table:

First digit X	1	2	3	4	5	6	7	8	9
Probability	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9

The probability that a first digit is equal to or greater than 6 is

$$\begin{aligned} P(X \geq 6) &= P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9) \\ &= \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{4}{9} = 0.444 \end{aligned}$$

Note that this is not the same as the probability that a random digit is strictly greater than 6, $P(X > 6)$. The outcome $X = 6$ is included in the event $\{X \geq 6\}$ and is omitted from $\{X > 6\}$.

Compare this with the probability of the same event among first digits from financial records that obey Benford's Law. A first digit V chosen from such records has the distribution

First digit V	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

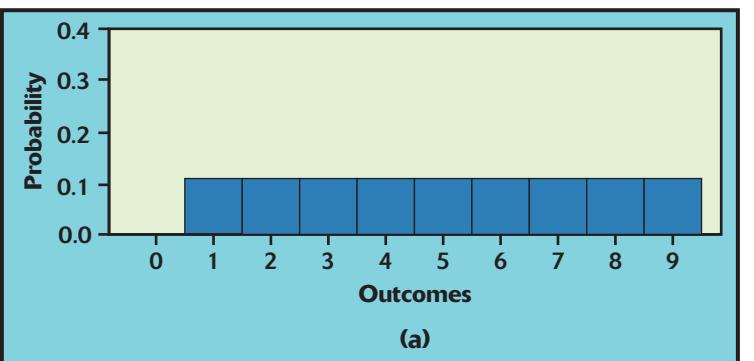
The probability that this digit is equal to or greater than 6 is

$$P(V \geq 6) = 0.067 + 0.058 + 0.051 + 0.046 = 0.222$$

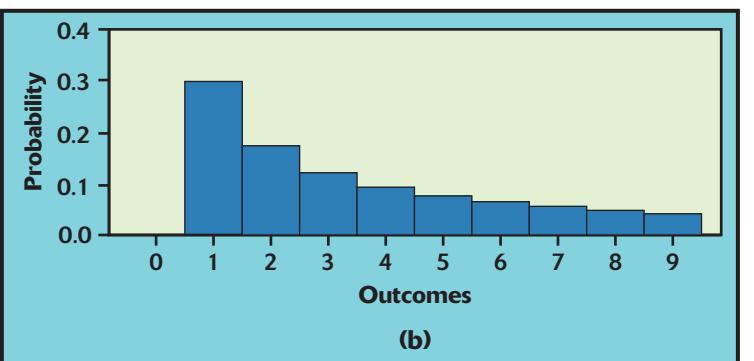
Benford's Law allows easy detection of phony financial records based on randomly generated numbers. These records tend to have too few first digits being 1s and 2s and too many first digits of 6 or greater.

probability histogram

We can use histograms to display probability distributions as well as distributions of data. Figure 4.3 displays **probability histograms** that compare the probability model for random digits with the model given by Benford's Law. The height of each bar shows the probability of the outcome at its base. Because the heights are probabilities, they add to 1. As usual, all the bars in a histogram have the same width. So the areas also display the assignment



(a)



(b)

FIGURE 4.3 Probability histograms for (a) random digits 1 to 9 and (b) Benford's Law. The height of each bar shows the probability assigned to a single outcome.

Outcome	Probability			
	Model 1	Model 2	Model 3	Model 4
	1/7	1/3	1/3	1
	1/7	1/6	1/6	1
	1/7	1/6	1/6	2
	1/7	0	1/6	1
	1/7	1/6	1/6	1
	1/7	1/6	1/6	2

FIGURE 4.4 Four assignments of probabilities to the six faces of a die, for Exercise 4.17.

of probability to outcomes. Think of these histograms as idealized pictures of the results of very many trials.

APPLY YOUR KNOWLEDGE

- 4.17 **Rolling a die.** Figure 4.4 displays several assignments of probabilities to the six faces of a die. We can learn which assignment is actually *accurate* for a particular die only by rolling the die many times. However, some of the assignments are not *legitimate* assignments of probability. That is, they do not obey the rules. Which are legitimate and which are not? In the case of the illegitimate models, explain what is wrong.
- 4.18 **Self-employed workers.** Draw a self-employed worker at random and record the industry in which the person works. “At random” means that we give every such person the same chance to be the one we choose. That is, we choose an SRS of size 1. The probability of any industry is just the proportion of all self-employed workers who work in that industry—if we drew many such workers, this is the proportion we would get. Here is the probability model:⁴

Industry	Probability
Agriculture	0.130
Construction	0.147
Finance, insurance, real estate	0.059
Manufacturing	0.042
Mining	0.002
Services	0.419
Trade	0.159
Transportation, public utilities	0.042

- (a) Show that this is a legitimate probability model.
 (b) What is the probability that a randomly chosen worker does not work in agriculture?
 (c) What is the probability that a randomly chosen worker works in construction, manufacturing, or mining?

- 4.19 **Job satisfaction.** We can use the results of a poll on job satisfaction to give a probability model for the job satisfaction rating of a randomly chosen employed (full-time or part-time) American.⁵ Here is the model:

Rating	Completely satisfied	Somewhat satisfied	Somewhat dissatisfied	Completely dissatisfied
Probability	?	0.47	0.12	0.02

- (a) What is the probability of a randomly selected employed American being completely satisfied with his or her job? Why?
- (b) What is the probability that a randomly selected employed American will be dissatisfied with his or her job?

Assigning probabilities: intervals of outcomes

A software random number generator is designed to produce a number between 0 and 1 chosen at random. It is only a slight idealization to consider *any* number in this range as a possible outcome. The sample space is then

$$S = \{\text{all numbers between 0 and 1}\}$$

Call the outcome of the random number generator Y for short. How can we assign probabilities to such events as $\{0.3 \leq Y \leq 0.7\}$? As in the case of selecting a random digit, we would like all possible outcomes to be equally likely. But we cannot assign probabilities to each individual value of Y and then add them, because there are infinitely many possible values.

We use a new way of assigning probabilities directly to events—as *areas under a density curve*. Any density curve has area exactly 1 underneath it, corresponding to total probability 1. We first met density curves as models for data in Chapter 1 (page 51).

EXAMPLE 4.7

Random numbers

Uniform density curve

The random number generator will spread its output uniformly across the entire interval from 0 to 1 as we allow it to generate a long sequence of numbers. The results of many trials are represented by the **Uniform density curve**, as shown in Figure 4.5. This density curve has height 1 over the interval from 0 to 1. The area under the curve is 1, and the probability of any event is the area under the curve and above the event in question.

As Figure 4.5(a) illustrates, the probability that the random number generator produces a number between 0.3 and 0.7 is

$$P(0.3 \leq Y \leq 0.7) = 0.4$$

because the area under the density curve and above the interval from 0.3 to 0.7 is 0.4. The height of the curve is 1 and the area of a rectangle is the product of height and length, so the probability of any interval of outcomes is just the length of the interval.

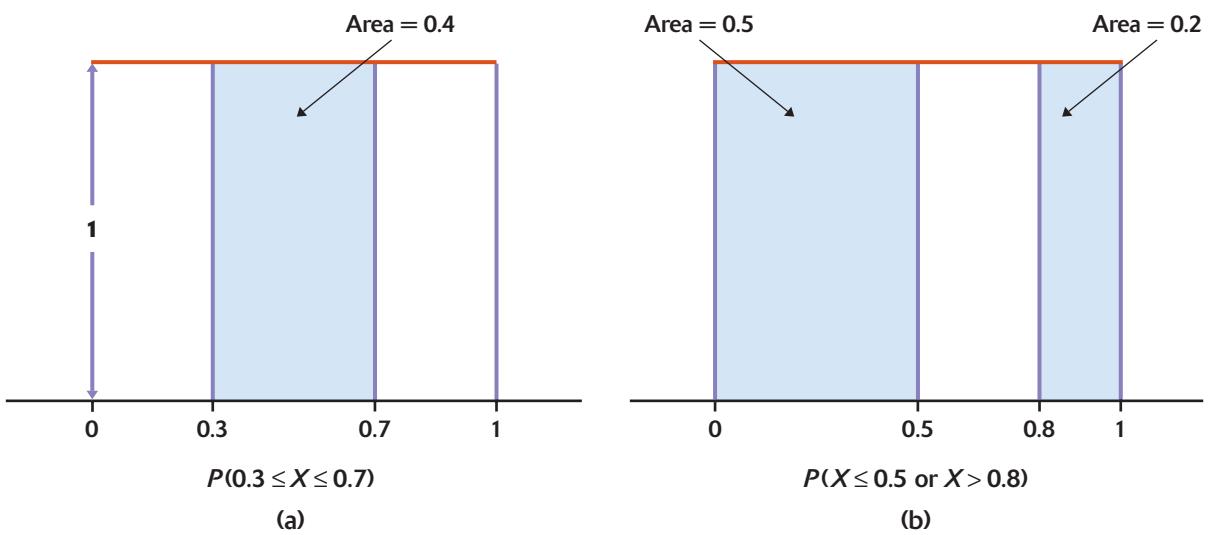


FIGURE 4.5 Probability as area under a density curve. These Uniform density curves spread probability evenly between 0 and 1.

Similarly,

$$P(Y \leq 0.5) = 0.5$$

$$P(Y > 0.8) = 0.2$$

$$P(Y \leq 0.5 \text{ or } Y > 0.8) = 0.7$$

Notice that the last event consists of two nonoverlapping intervals, so the total area above the event is found by adding two areas, as illustrated by Figure 4.5(b). This assignment of probabilities obeys all of our rules for probability.

Compare the density curves in Figure 4.5 with the probability histogram in Figure 4.3(a). Both describe Uniform distributions. In 4.3(a), there are only 9 possible outcomes, the whole numbers 1, 2, ..., 9. In Figure 4.5, the outcome can be any number between 0 and 1. In both figures, probability is given by area. Although the mathematics required to deal with density curves is more advanced, the most important ideas of probability apply to both kinds of probability models.

APPLY YOUR KNOWLEDGE

- 4.20 **Random numbers.** Let Y be a random number between 0 and 1 produced by the idealized Uniform random number generator described in Example 4.7 and Figure 4.5. Find the following probabilities:

- (a) $P(0 \leq Y \leq 0.4)$
- (b) $P(0.4 \leq Y \leq 1)$
- (c) $P(0.3 \leq Y \leq 0.5)$
- (d) $P(0.3 < Y < 0.5)$

- 4.21 **Adding random numbers.** Generate two random numbers between 0 and 1 and take T to be their sum. The sum T can take any value between 0 and 2. The density curve of T is the triangle shown in Figure 4.6.

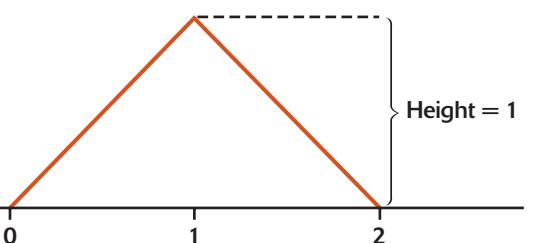


FIGURE 4.6 The density curve for the sum of two random numbers, for Exercise 4.21. This density curve spreads probability between 0 and 2.

- Verify by geometry that the area under this curve is 1.
- What is the probability that T is less than 1? (Sketch the density curve, shade the area that represents the probability, then find that area. Do this for (c) also.)
- What is $P(T < 0.5)$?

Normal probability models

Any density curve can be used to assign probabilities. The density curves that are most familiar to us are the Normal curves introduced in Section 1.3. So **Normal distributions are probability models**. There is a close connection between a Normal distribution as an idealized description for data and a Normal probability model. If we look at the weights of all 9-ounce bags of a particular brand of potato chip, we find that they closely follow the Normal distribution with mean $\mu = 9.12$ ounces and standard deviation $\sigma = 0.15$ ounces, $N(9.12, 0.15)$. That is a distribution for a large set of data. Now choose one 9-ounce bag at random. Call its weight W . If we repeat the random choice very many times, the distribution of values of W is the same Normal distribution.

EXAMPLE 4.8

The weight of bags of potato chips

What is the probability that a randomly chosen 9-ounce bag has weight between 9.33 and 9.45 ounces?

The weight W of the bag we choose has the $N(9.12, 0.15)$ distribution. Find the probability by standardizing and using Table A, the table of standard Normal probabilities. We will reserve capital Z for a standard Normal variable, $N(0, 1)$.

$$\begin{aligned} P(9.33 \leq W \leq 9.45) &= P\left(\frac{9.33 - 9.12}{0.15} \leq \frac{W - 9.12}{0.15} \leq \frac{9.45 - 9.12}{0.15}\right) \\ &= P(1.4 \leq Z \leq 2.2) \\ &= 0.9861 - 0.9192 = 0.0669 \end{aligned}$$

Figure 4.7 shows the areas under the standard Normal curve. The calculation is the same as those we did in Chapter 1. Only the language of probability is new.

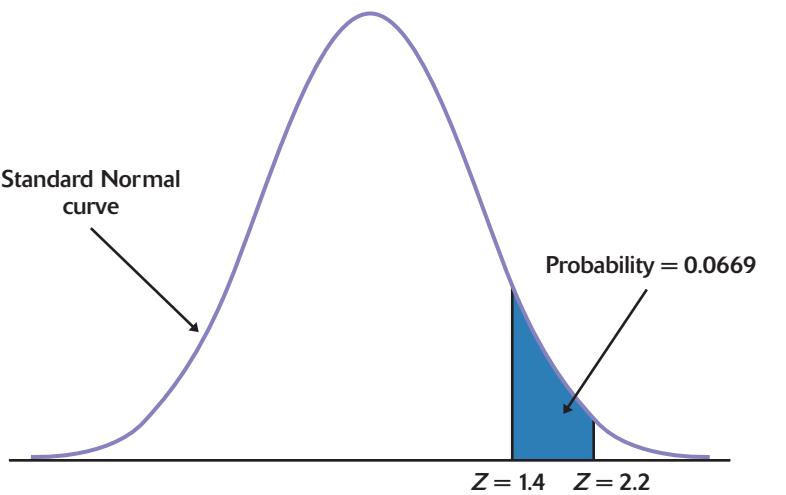


FIGURE 4.7 The probability in Example 4.8 as an area under the standard Normal curve.

APPLY YOUR KNOWLEDGE

- 4.22 MPG for 2001 vehicles. The Normal distribution with mean $\mu = 21.2$ miles per gallon and standard deviation $\sigma = 5.4$ MPG is an approximate model for the city gas mileage of 2001 model year vehicles. Figure 1.14 (page 51) pictures this distribution. Let X be the city MPG of one 2001 vehicle chosen at random.
- Write the event “the vehicle chosen has a city MPG of 32 or higher” in terms of X .
 - Find the probability of this event.
- 4.23 NAEP math scores. Scores on the National Assessment of Educational Progress 12th-grade mathematics test for the year 2000 were approximately Normal with mean 300 points (out of 500 possible) and standard deviation 35 points. Let Y stand for the score of a randomly chosen student. Express each of the following events in terms of Y and use the 68–95–99.7 rule to give the approximate probability.
- The student has a score above 300.
 - The student’s score is above 370.

SECTION 4.2 SUMMARY

- A probability model for a random phenomenon consists of a sample space S and an assignment of probabilities P .
- The sample space S is the set of all possible outcomes of the random phenomenon. Sets of outcomes are called **events**. P assigns a number $P(A)$ to an event A as its probability.
- Any assignment of probability must obey the rules that state the basic properties of probability:

1. $0 \leq P(A) \leq 1$ for any event A .
2. $P(S) = 1$.

3. For any event A , $P(A \text{ does not occur}) = 1 - P(A)$.
4. **Addition rule:** Events A and B are **disjoint** if they have no outcomes in common. If A and B are disjoint, then $P(A \text{ or } B) = P(A) + P(B)$.

- When a sample space S contains finitely many possible outcomes, a probability model assigns each of these outcomes a probability between 0 and 1 such that the sum of all the probabilities is exactly 1. The probability of any event is the sum of the probabilities of all the outcomes that make up the event.
- A sample space can contain all values in some interval of numbers. A probability model assigns probabilities as **areas under a density curve**. The probability of any event is the area under the curve above the outcomes that make up the event.

SECTION 4.2 EXERCISES

- 4.24 **Stock price movements.** You watch the price of Cisco Systems stock for four days. Give a sample space for each of these random phenomena:
- You record the sequence of up days and down days.
 - You record the number of up days.
- 4.25 **Land in Iowa.** Choose an acre of land in Iowa at random. The probability is 0.92 that it is farmland and 0.01 that it is forest.
- What is the probability that the acre chosen is not farmland?
 - What is the probability that it is either farmland or forest?
 - What is the probability that a randomly chosen acre in Iowa is something other than farmland or forest?
- 4.26 **Car colors.** Choose a new car or light truck at random and note its color. Here are the probabilities of the most popular colors for cars made in North America in 2000:⁶

Color	Silver	White	Black	Dark green	Dark blue	Medium red
Probability	0.176	0.172	0.113	0.089	0.088	0.067

What is the probability that the car you choose has any color other than the six listed? What is the probability that a randomly chosen car is either silver or white?

- 4.27 **Colors of M&Ms.** The colors of candies such as M&Ms are carefully chosen to match consumer preferences. The color of an M&M drawn at random from a bag has a probability distribution determined by the proportions of colors among all M&Ms of that type.
- Here is the distribution for plain M&Ms:

Color	Brown	Red	Yellow	Green	Orange	Blue
Probability	0.3	0.2	0.2	0.1	0.1	?

What must be the probability of drawing a blue candy?

- (b) The probabilities for peanut M&Ms are a bit different. Here they are:

Color	Brown	Red	Yellow	Green	Orange	Blue
Probability	0.2	0.2	0.2	0.1	0.1	?

What is the probability that a peanut M&M chosen at random is blue?

- (c) What is the probability that a plain M&M is any of red, yellow, or orange?
What is the probability that a peanut M&M has one of these colors?

- 4.28 **Crispy M&Ms.** Exercise 4.27 gives the probabilities that an M&M candy is each of brown, red, yellow, green, orange, and blue. “Crispy Chocolate” M&Ms are equally likely to be any of these colors. What is the probability of any one color?
- 4.29 **Legitimate probabilities?** In each of the following situations, state whether or not the given assignment of probabilities to individual outcomes is legitimate, that is, satisfies the rules of probability. If not, give specific reasons for your answer.
- (a) When a coin is spun, $P(H) = 0.55$ and $P(T) = 0.45$.
 - (b) When two coins are tossed, $P(HH) = 0.4$, $P(HT) = 0.4$, $P(TH) = 0.4$, and $P(TT) = 0.4$.
 - (c) Plain M&Ms have not always had the mixture of colors given in Exercise 4.27. In the past there were no red candies and no blue candies. Tan had probability 0.10, and the other four colors had the same probabilities that are given in Exercise 4.27.
- 4.30 **Who goes to Paris?** Abby, Deborah, Sam, Tonya, and Roberto work in a firm’s public relations office. Their employer must choose two of them to attend a conference in Paris. To avoid unfairness, the choice will be made by drawing two names from a hat. (This is an SRS of size 2.)
- (a) Write down all possible choices of two of the five names. This is the sample space.
 - (b) The random drawing makes all choices equally likely. What is the probability of each choice?
 - (c) What is the probability that Tonya is chosen?
 - (d) What is the probability that neither of the two men (Sam and Roberto) is chosen?
- 4.31 **How big are farms?** Choose an American farm at random and measure its size in acres. Here are the probabilities that the farm chosen falls in several acreage categories:

Acres	<10	10–49	50–99	100–179	180–499	500–999	1000–1999	≥ 2000
Probability	0.09	0.20	0.15	0.16	0.22	0.09	0.05	0.04

Let A be the event that the farm is less than 50 acres in size, and let B be the event that it is 500 acres or more.

- (a) Find $P(A)$ and $P(B)$.
- (b) Describe the event “ A does not occur” in words and find its probability by Rule 3.

- (c) Describe the event “ A or B ” in words and find its probability by the addition rule.
- 4.32 Roulette.** A roulette wheel has 38 slots, numbered 0, 00, and 1 to 36. The slots 0 and 00 are colored green, 18 of the others are red, and 18 are black. The dealer spins the wheel and at the same time rolls a small ball along the wheel in the opposite direction. The wheel is carefully balanced so that the ball is equally likely to land in any slot when the wheel slows. Gamblers can bet on various combinations of numbers and colors.
- (a) What is the probability of any one of the 38 possible outcomes? Explain your answer.
 - (b) If you bet on “red,” you win if the ball lands in a red slot. What is the probability of winning?
 - (c) The slot numbers are laid out on a board on which gamblers place their bets. One column of numbers on the board contains all multiples of 3, that is, 3, 6, 9, ..., 36. You place a “column bet” that wins if any of these numbers comes up. What is your probability of winning?
- 4.33 Consumer preference.** Suppose that half of all computer users prefer the new version of your company’s best-selling software, and half prefer the old version. If you randomly choose three consumers and record their preferences, there are 8 possible outcomes. For example, NNO means the first two consumers prefer the new (N) version and the third consumer prefers the old version (O). All 8 arrangements are equally likely.
- (a) Write down all 8 arrangements of preferences of three consumers. What is the probability of any one of these arrangements?
 - (b) Let X be the number of consumers who prefer the new version of the product. What is the probability that $X = 2$?
 - (c) Starting from your work in (a), list the values X can take and the probability for each value.
- 4.34 Moving up.** A study of class mobility in England looked at the economic class reached by the sons of lower-class fathers. Economic classes are numbered from 1 (low) to 5 (high). Take X to be the class of a randomly chosen son of a father in Class 1. The study found that the probability model for X is
- | Son’s class X | 1 | 2 | 3 | 4 | 5 |
|-----------------|------|------|------|------|------|
| Probability | 0.48 | 0.38 | 0.08 | 0.05 | 0.01 |
- (a) Check that this distribution satisfies the two requirements for a legitimate assignment of probabilities to individual outcomes.
 - (b) What is $P(X \leq 3)$? (Be careful: the event “ $X \leq 3$ ” includes the value 3.)
 - (c) What is $P(X < 3)$?
 - (d) Write the event “a son of a lower-class father reaches one of the two highest classes” in terms of values of X . What is the probability of this event?
- 4.35 How large are households?** Choose an American household at random and let X be the number of persons living in the household. If we ignore the few

households with more than seven inhabitants, the probability model for X is as follows:

Household size X	1	2	3	4	5	6	7
Probability	0.25	0.32	0.17	0.15	0.07	0.03	0.01

- (a) Verify that this is a legitimate probability distribution.
 - (b) What is $P(X \geq 5)$?
 - (c) What is $P(X > 5)$?
 - (d) What is $P(2 < X \leq 4)$?
 - (e) What is $P(X \neq 1)$?
 - (f) Write the event that a randomly chosen household contains more than two persons in terms of X . What is the probability of this event?
- 4.36 **Random numbers.** Many random number generators allow users to specify the range of the random numbers to be produced. Suppose that you specify that the random number Y can take any value between 0 and 2. Then the density curve of the outcomes has constant height between 0 and 2, and height 0 elsewhere.
- (a) What is the height of the density curve between 0 and 2? Draw a graph of the density curve.
 - (b) Use your graph from (a) and the fact that probability is area under the curve to find $P(Y \leq 1)$.
 - (c) Find $P(0.5 < Y < 1.3)$.
 - (d) Find $P(Y \geq 0.8)$.

4.3 Random Variables

Not all sample spaces are made up of numbers. When we toss a coin four times, we can record the outcome as a string of heads and tails, such as HTTH. In statistics, however, we are most often interested in numerical outcomes such as the count of heads in the four tosses. It is convenient to use a shorthand notation: Let X be the number of heads. If our outcome is HTTH, then $X = 2$. If the next outcome is TTTH, the value of X changes to $X = 1$. The possible values of X are 0, 1, 2, 3, and 4. Tossing a coin four times will give X one of these possible values. Tossing four more times will give X another and probably different value. We call X a *random variable* because its values vary when the coin tossing is repeated. Examples 4.6 and 4.7 used this shorthand notation. In Example 4.8, we let W stand for the weight of a randomly chosen 9-ounce bag of potato chips. We know that W would take a different value if we took another random sample. Because its value changes from one sample to another, W is also a random variable.

RANDOM VARIABLE

A **random variable** is a variable whose value is a numerical outcome of a random phenomenon.

We usually denote random variables by capital letters near the end of the alphabet, such as X or Y . Of course, the random variables of greatest interest to us are outcomes such as the mean \bar{x} of a random sample, for which we will keep the familiar notation.⁷ As we progress from general rules of probability toward statistical inference, we will concentrate on random variables. When a random variable X describes a random phenomenon, the sample space S just lists the possible values of the random variable. We usually do not mention S separately.

We will meet two types of random variables, corresponding to two ways of assigning probability: *discrete* and *continuous*. Compare the random variable X in Example 4.6 (page 243) and the random variable Y in Example 4.7 (page 246). Both have Uniform distributions that distribute probability evenly across the possible outcomes, but one is discrete and the other is continuous.

The possible values of X in Example 4.6 are the whole numbers $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. These nine values are separated on the number line by other numbers that are not possible values of X :



discrete random variable

X is a **discrete random variable**. “Discrete” describes the nature of the sample space for X . In contrast, the sample space for Y in Example 4.7 is $\{\text{all numbers between } 0 \text{ and } 1\}$. These possible values are not separated on the number line. As you move continuously from the value 0 to the value 1, you do not leave the sample space for Y :



continuous random variable

Y is a **continuous random variable**. Once again, “continuous” describes the nature of the sample space.

APPLY YOUR KNOWLEDGE

- 4.37 For each exercise listed below, decide whether the random variable described is discrete or continuous and give a brief explanation for your response.
 - (a) Exercise 4.12(b), page 239
 - (b) Exercise 4.12(e), page 240
 - (c) Exercise 4.22, page 249
 - (d) Exercise 4.35, page 252

- 4.38 For each exercise listed below, decide whether the random variable described is discrete or continuous and sketch the sample space on a number line.
 - (a) Exercise 4.13(a), page 240
 - (b) Exercise 4.13(c), page 240
 - (c) Exercise 4.13(e), page 240

Probability Distributions

The starting point for studying any random variable is its probability distribution, which is just the probability model for the outcomes.

PROBABILITY DISTRIBUTION

The **probability distribution** of a random variable X tells us what values X can take and how to assign probabilities to those values.

Because of the difference in the nature of sample spaces for discrete and continuous random variables, we describe probability distributions for the two types of random variables separately.

Discrete Random Variables

A discrete random variable like the random digit X in Example 4.6 (page 243) has a finite number of possible values.⁸ The probability distribution of X therefore simply lists the values and their probabilities.

DISCRETE PROBABILITY DISTRIBUTIONS

The **probability distribution** of a discrete random variable X lists the possible values of X and their probabilities:

Value of X	x_1	x_2	x_3	\cdots	x_k
Probability	p_1	p_2	p_3	\cdots	p_k

The probabilities p_i must satisfy two requirements:

1. Every probability p_i is a number between 0 and 1, $0 \leq p_i \leq 1$.
2. The sum of the probabilities is exactly 1, $p_1 + p_2 + \cdots + p_k = 1$.

To find the probability of any event, add the probabilities p_i of the individual values x_i that make up the event.

We can use a probability histogram to display a discrete distribution. Figure 4.3 (page 244) compares the distributions of random digits and first digits that obey Benford's Law.

EXAMPLE 4.9

Hard-drive sizes

Buyers of a laptop computer model may choose to purchase either a 10 GB (gigabyte), 20 GB, 30 GB, or 40 GB internal hard drive. Choose customers from the last 60 days at random to ask what influenced their choice of computer. To "choose at random" means to give every customer of the last 60 days the same

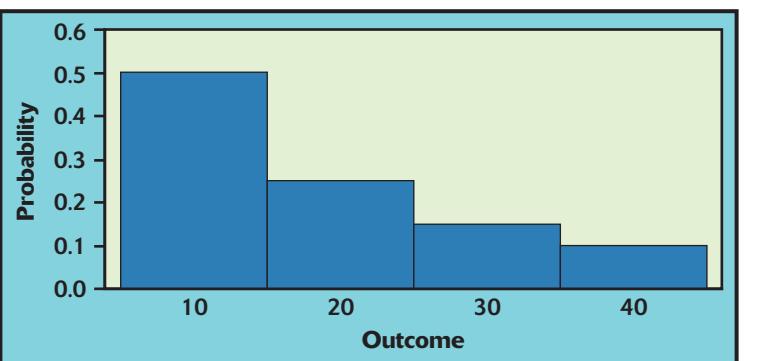


FIGURE 4.8 Probability histogram for the distribution of hard-drive sizes in Example 4.9.

chance to be chosen. The size (in gigabytes) of the internal hard drive chosen by a randomly selected customer is a random variable X .

The value of X changes when we repeatedly choose customers at random, but it is always one of 10, 20, 30, or 40 GB. Here is the distribution of X :

Hard-drive size X	10	20	30	40
Probability	0.50	0.25	0.15	0.10

The probability histogram in Figure 4.8 pictures this distribution.

The probability that a randomly selected customer chose at least a 30 GB hard drive is

$$\begin{aligned} P(\text{size is 30 or 40}) &= P(X = 30) + P(X = 40) \\ &= 0.15 + 0.10 = 0.25 \end{aligned}$$



EXAMPLE 4.10

Four coin tosses

Toss a balanced coin four times; the discrete random variable X counts the number of heads. How shall we find the probability distribution of X ?

The outcome of four tosses is a sequence of heads and tails such as HTTH. There are 16 possible outcomes in all. Figure 4.9 lists these outcomes along with the value of X for each outcome. A reasonable probability model says that the 16 outcomes are all equally likely; that is, each has probability 1/16. (This model is justified both by empirical observation and by a theoretical derivation that appears in the optional Chapter 5.)

The number of heads X has possible values 0, 1, 2, 3, and 4. These values are *not* equally likely. As Figure 4.9 shows, $X = 0$ can occur in only one way, when the outcome is TTTT. So $P(X = 0) = 1/16$. The outcome $X = 2$, on the other hand, can occur in six different ways, so that

$$\begin{aligned} P(X = 2) &= \frac{\text{count of ways } X = 2 \text{ can occur}}{16} \\ &= \frac{6}{16} \end{aligned}$$

			HTTH HTHT HTTT THTT TTHT TTTT	THTH HHTT THHT TTHH	HHHT HHTH HTHH THHH	HHHH
$X = 0$	$X = 1$	$X = 2$				

FIGURE 4.9 Possible outcomes in four tosses of a coin, for Example 4.10. The random variable X is the number of heads.

We can find the probability of each value of X from Figure 4.9 in the same way. Here is the result:

$$P(X = 0) = \frac{1}{16} = 0.0625$$

$$P(X = 1) = \frac{4}{16} = 0.25$$

$$P(X = 2) = \frac{6}{16} = 0.375$$

$$P(X = 3) = \frac{4}{16} = 0.25$$

$$P(X = 4) = \frac{1}{16} = 0.0625$$

These probabilities have sum 1, so this is a legitimate probability distribution. In table form, the distribution is

Number of heads X	0	1	2	3	4
Probability	0.0625	0.25	0.375	0.25	0.0625

Figure 4.10 is a probability histogram for this distribution. The probability distribution is exactly symmetric. It is an idealization of the distribution of the

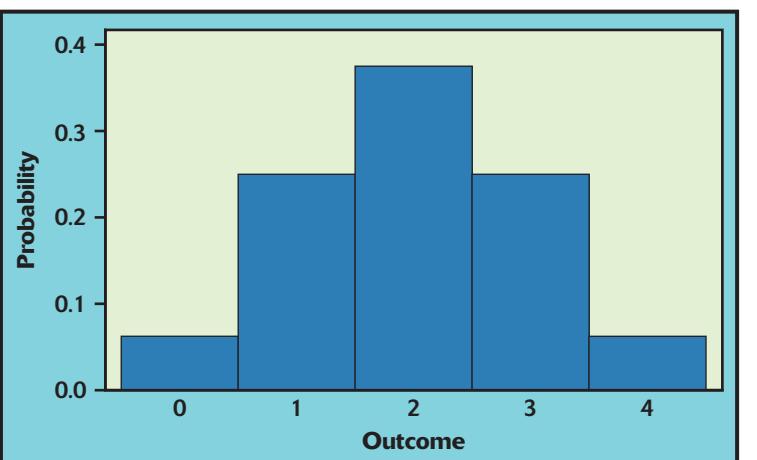


FIGURE 4.10 Probability histogram for the number of heads in four tosses of a coin, for Example 4.10.

proportions for numbers of heads after many tosses of four coins, which would be nearly symmetric but is unlikely to be exactly symmetric.

Any event involving the number of heads observed can be expressed in terms of X , and its probability can be found from the distribution of X . For example, the probability of tossing at least two heads is

$$P(X \geq 2) = 0.375 + 0.25 + 0.0625 = 0.6875$$

The probability of at least one head is most simply found by

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - 0.0625 = 0.9375 \end{aligned}$$

Why talk about tossing coins? The probability distribution for repeated coin tossing also fits many real-world problems: respondents answer Yes or No to a poll question; consumers buy or don't buy a product after viewing an advertisement; an inspector finds good or defective parts. Of course, we must allow the probability of a "head" to be something other than one-half in these settings, but that is a minor change. We will extend the results of Example 4.10 when we return to sampling distributions in the next section.

Continuous Random Variables

A continuous random variable like the Uniform random number Y in Example 4.7 (page 246) or the Normal package weight W in Example 4.8 (page 248) has an infinite number of possible values. Continuous probability distributions therefore assign probabilities directly to events as areas under a density curve.

CONTINUOUS PROBABILITY DISTRIBUTIONS

The **probability distribution** of a continuous random variable X is described by a density curve. The probability of any event is the area under the density curve and above the values of X that make up the event.

EXAMPLE 4.11

Tread life

The actual tread life X of a 40,000-mile automobile tire has a Normal probability distribution with $\mu = 50,000$ miles and $\sigma = 5500$ miles. We say X has a $N(50000, 5500)$ distribution. The probability that a randomly selected tire has a tread life less than 40,000 miles is

$$\begin{aligned} P(X < 40,000) &= P\left(\frac{X - 50,000}{5500} < \frac{40,000 - 50,000}{5500}\right) \\ &= P(Z < -1.82) \\ &= 0.0344 \end{aligned}$$

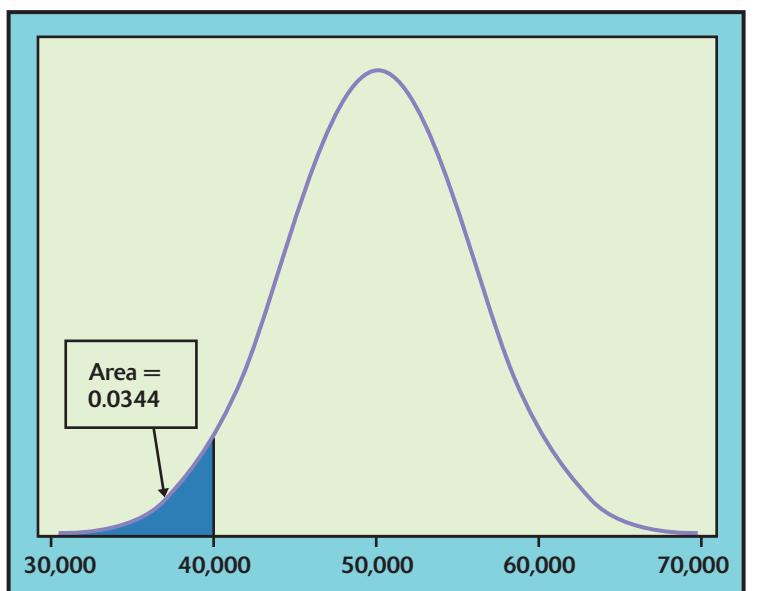


FIGURE 4.11 The Normal distribution with $\mu = 50,000$ and $\sigma = 5500$. The shaded area is $P(X < 40,000)$, calculated in Example 4.11.

The $N(50000, 5500)$ distribution is shown in Figure 4.11 with the area $P(X < 40,000)$ shaded.



The probability distribution for a continuous random variable assigns probabilities to intervals of outcomes rather than to individual outcomes. In fact, all continuous probability distributions assign probability 0 to every individual outcome. Only intervals of values have positive probability. To see that this makes sense, return to the Uniform random number generator of Example 4.7.

EXAMPLE 4.12

Can a random number be exactly 0.8?

What is $P(Y = 0.8)$, the probability that a random number generator produces exactly 0.8? Figure 4.12 shows the Uniform density curve for outcomes between 0 and 1. Probabilities are areas under this curve. The probability of any interval of outcomes is equal to its length. This implies that because the point 0.8 has no length, its probability is 0.

Although this fact may seem odd, it makes intuitive as well as mathematical sense. The random number generator produces a number between 0.79 and 0.81 with probability 0.02. An outcome between 0.799 and 0.801 has probability 0.002. A result between 0.799999 and 0.800001 has probability 0.000002. You see that as we home in on 0.8 the probability gets closer to 0. To be consistent, the probability of an outcome *exactly* equal to 0.8 must be 0.

Because there is no probability exactly at $Y = 0.8$, the two events $\{Y > 0.8\}$ and $\{Y \geq 0.8\}$ have the same probability. We can ignore the distinction between

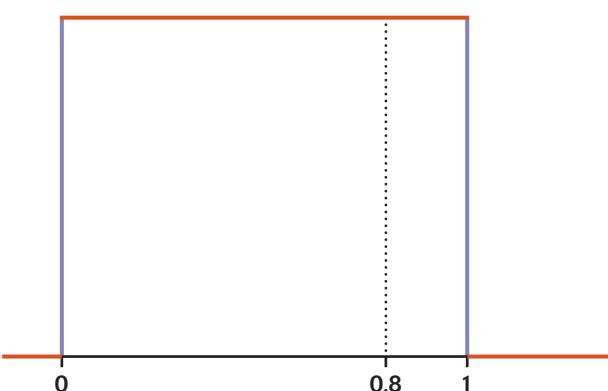


FIGURE 4.12 The density curve of the Uniform distribution of numbers chosen at random between 0 and 1. This model assigns probability 0 to the event that the number generated has any one specific value, such as 0.8.

> and \geq when finding probabilities for continuous (but not discrete) random variables.



Example 4.12 reminds us that continuous probability distributions are idealizations. We don't measure the weight of potato chip bags or the tread life of tires to an unlimited number of decimal places. If we measure tread life to the nearest mile, we could use a discrete distribution with a very small probability on each specific tread life. It is much more convenient to use a continuous Normal distribution. There is an element of “not exactly right but good enough to use in practice” in most choices of a probability model.

We began this chapter with a general discussion of the idea of probability and the properties of probability models. Two specific types of probability models are distributions of discrete and continuous random variables. Our study of statistics will employ only these two types of probability models.

- 4.39 **How many cars?** Choose an American household at random and let the random variable X be the number of cars (including SUVs and light trucks) they own. Here is the probability model if we ignore the few households that own more than 5 cars:

Number of cars X	0	1	2	3	4	5
Probability	0.09	0.36	0.35	0.13	0.05	0.02

- (a) Verify that this is a legitimate discrete distribution. Display the distribution in a probability histogram.
- (b) Say in words what the event $\{X \geq 1\}$ is. Find $P(X \geq 1)$.
- (c) Your company builds houses with two-car garages. What percent of households have more cars than the garage can hold?

APPLY YOUR KNOWLEDGE

- 4.40 Let Y be a random number between 0 and 1 produced by the idealized Uniform random number generator with density curve pictured in Figure 4.12. Find the following probabilities:
- $P(0 \leq Y \leq 0.4)$
 - $P(0.4 \leq Y \leq 1)$
 - $P(0.3 \leq Y \leq 0.5)$
 - $P(0.3 < Y < 0.5)$
 - $P(0.226 \leq Y \leq 0.713)$
- 4.41 Example 4.11 gives the Normal distribution $N(50000, 5500)$ for the tread life X of a type of tire (in miles). Calculate the following probabilities:
- The probability that a tire lasts more than 50,000 miles.
 - $P(X > 60,000)$
 - $P(X \geq 60,000)$

The mean of a random variable

Probability is the mathematical language that describes the long-run regular behavior of random phenomena. The probability distribution of a random variable is an idealized distribution of the proportions of outcomes in very many observations. The probability histograms and density curves that picture probability distributions resemble our earlier pictures of distributions of data. In describing data, we moved from graphs to numerical measures such as means and standard deviations. Now we will make the same move to expand our descriptions of the distributions of random variables. We can speak of the mean winnings in a game of chance or the standard deviation of the randomly varying number of calls a travel agency receives in an hour. We will learn more about how to compute these descriptive measures and about the laws they obey.

The mean \bar{x} of a set of observations is their ordinary average. The mean of a random variable X is also an average of the possible values of X , but with an essential change to take into account the fact that not all outcomes need be equally likely. An example will show what we must do.

EXAMPLE 4.13

Pick 3

Most states and Canadian provinces have government-sponsored lotteries. Here is a simple lottery wager, from the Tri-State Pick 3 game that New Hampshire shares with Maine and Vermont. You choose a three-digit number; the state chooses a three-digit winning number at random and pays you \$500 if your number is chosen. Because there are 1000 three-digit numbers, you have probability 1/1000 of winning. Taking X to be the amount your ticket pays you, the probability distribution of X is

Payoff X	\$0	\$500
Probability	0.999	0.001

What is your average payoff from many tickets? The ordinary average of the two possible outcomes \$0 and \$500 is \$250, but that makes no sense as the average because \$500 is much less likely than \$0. In the long run you receive \$500 once in every 1000 tickets and \$0 on the remaining 999 of 1000 tickets. The long-run average payoff is

$$\$500 \frac{1}{1000} + \$0 \frac{999}{1000} = \$0.50$$

or fifty cents. That number is the mean of the random variable X . (Tickets cost \$1, so in the long run the state keeps half the money you wager.)

mean μ

expected value

If you play Tri-State Pick 3 several times, we would as usual call the mean of the actual amounts your tickets pay \bar{x} . The mean in Example 4.13 is a different quantity—it is the long-run average payoff you expect if you play a very large number of times. Just as probabilities are an idealized description of long-run proportions, the mean of a probability distribution describes the long-run average outcome. The common symbol for the **mean of a probability distribution** is μ , the Greek letter mu. We used μ in Chapter 1 for the mean of a Normal distribution, so this is not a new notation. We will often be interested in several random variables, each having a different probability distribution with a different mean. To remind ourselves that we are talking about the mean of X , we can write μ_X rather than simply μ . In Example 4.13, $\mu_X = \$0.50$. Notice that, as often happens, the mean is not a possible value of X . You will often find the mean of a random variable X called the **expected value** of X . This term can be misleading, for we don't necessarily expect one observation on X to be close to its mean.

The mean of any discrete random variable is found just as in Example 4.13. It is an average of the possible outcomes, but a weighted average in which each outcome is weighted by its probability. Because the probabilities add to 1, we have total weight 1 to distribute among the outcomes. An outcome that occurs half the time has probability one-half and gets one-half the weight in calculating the mean. Here is the general definition.

MEAN OF A DISCRETE RANDOM VARIABLE

Suppose that X is a discrete random variable whose distribution is

Value of X	x_1	x_2	x_3	\cdots	x_k
Probability	p_1	p_2	p_3	\cdots	p_k

To find the **mean of X** , multiply each possible value by its probability, then add all the products:

$$\begin{aligned}\mu_X &= x_1 p_1 + x_2 p_2 + \cdots + x_k p_k \\ &= \sum x_i p_i\end{aligned}$$

EXAMPLE 4.14**First digits**

If first digits in a set of records appear “at random,” the nine possible digits 1 to 9 all have the same probability. The probability distribution of the first digit X is then

First digit X	1	2	3	4	5	6	7	8	9
Probability	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9

The mean of this distribution is

$$\begin{aligned}\mu_X &= 1\frac{1}{9} + 2\frac{1}{9} + 3\frac{1}{9} + 4\frac{1}{9} + 5\frac{1}{9} + 6\frac{1}{9} + 7\frac{1}{9} + 8\frac{1}{9} + 9\frac{1}{9} \\ &= 45 \times \frac{1}{9} = 5\end{aligned}$$

If, on the other hand, the records obey Benford’s Law, the distribution of the first digit V is

First digit V	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

The mean of V is

$$\begin{aligned}\mu_V &= (1)(0.301) + (2)(0.176) + (3)(0.125) + (4)(0.097) + (5)(0.079) \\ &\quad + (6)(0.067) + (7)(0.058) + (8)(0.051) + (9)(0.046) \\ &= 3.441\end{aligned}$$

The means reflect the greater probability of smaller first digits under Benford’s Law.

Figure 4.13 locates the means of X and V on the two probability histograms. Because the discrete Uniform distribution of Figure 4.13(a) is symmetric, the mean lies at the center of symmetry. We can’t locate the mean of the right-skewed distribution of Figure 4.13(b) by eye—calculation is needed.



What about continuous random variables? The probability distribution of a continuous random variable X is described by a density curve. Chapter 1 showed how to find the mean of the distribution: it is the point at which the area under the density curve would balance if it were made out of solid material. The mean lies at the center of symmetric density curves such as the Normal curves. Exact calculation of the mean of a distribution with a skewed density curve requires advanced mathematics.⁹ The idea that the mean is the balance point of the distribution applies to discrete random variables as well, but in the discrete case we have a formula that gives us the numerical value of μ .

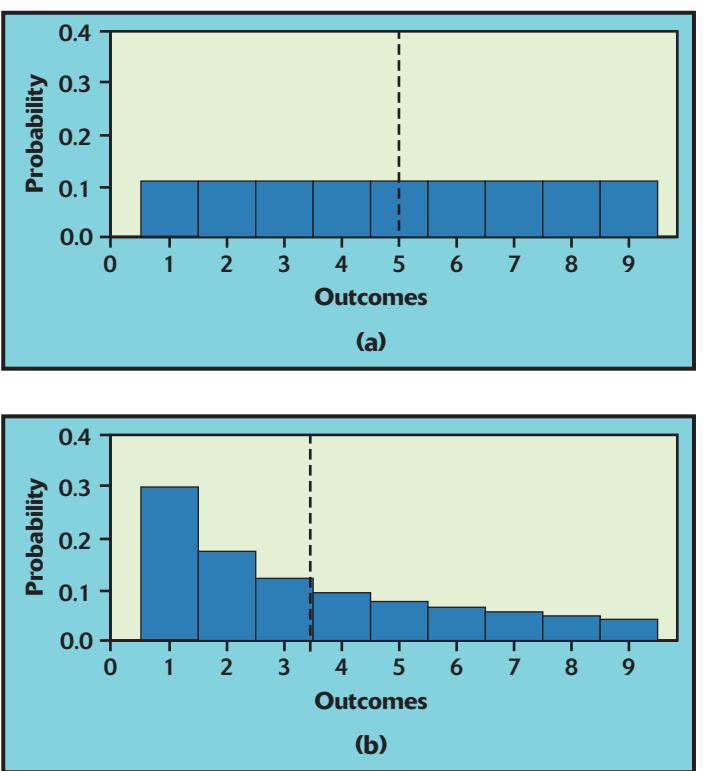


FIGURE 4.13 Locating the mean of a discrete random variable on the probability histogram for (a) digits between 1 and 9 chosen at random; (b) digits between 1 and 9 chosen from records that obey Benford’s Law.

APPLY YOUR KNOWLEDGE

- 4.42 **Household size.** Choose an American household at random and let X be the number of persons living in the household. If we ignore the few households with more than seven inhabitants, the probability model for X is as follows:

Household size X	1	2	3	4	5	6	7
Probability	0.25	0.32	0.17	0.15	0.07	0.03	0.01

Find the mean μ_X . Sketch a probability histogram of the distribution of X and mark the mean on your sketch. The Census Bureau reports that the mean size of American households is 2.65 people. Why is your result slightly less than this?

- 4.43 **Hard-drive sizes.** Example 4.9 (page 255) gives the distribution of customer choices of hard-drive size for a laptop computer model. Find the mean μ of this probability distribution. Explain in simple language what μ tells us about customer choices. Also explain why knowing μ is not very helpful to the computer maker.
- 4.44 **Customer orders.** Each week your business receives orders from customers who wish to have air-conditioning units installed in their homes. From past

data, you estimate the distribution of the number of units ordered in a week X to be

Units ordered X	0	1	2	3	4	5
Probability	0.05	0.15	0.27	0.33	0.13	0.07

- (a) Sketch a probability histogram for the distribution of X .
- (b) Calculate the mean of X and mark it on the horizontal axis of your probability histogram.
- (c) If you hire workers sufficient to handle the mean demand, what is the probability that you will be unable to handle all of a week's installation orders?

Rules for means



PORTFOLIO ANALYSIS

The *rate of return* of an investment over a time period is the percent change in the price during the time period, plus any income received. For example, Apple Computer's stock price was \$14.875 at the beginning of January 2001 and \$21.625 at the end of that month. Apple pays no dividends, so Apple's rate of return for that time period was

$$\frac{\text{change in price}}{\text{starting price}} = \frac{21.625 - 14.875}{14.875} = 0.45, \text{ or } 45\%$$

Investors want high returns, but they also want safety. Apple's stock price has often halved or doubled within a few months. The variability of returns, called *volatility* in finance, is a measure of the risk of an investment. A highly volatile stock, which may go either up or down a lot, is more risky than a Treasury bill, whose return is very predictable.

A *portfolio* is a collection of investments held by an individual or an institution. *Portfolio analysis* begins by studying how the risk and return of a portfolio are determined by the risk and return of the individual investments it contains. That's where statistics comes in: the return on an investment over some period of time is a random variable. We are interested in the *mean* return and we measure volatility by the *standard deviation* of returns. Let's say that Sadie's portfolio places 20% of her funds in Treasury bills and 80% in an "index fund" that represents all U.S. common stocks. If X is the annual return on T-bills and Y the annual return on stocks, the portfolio rate of return is

$$R = 0.2X + 0.8Y$$

How can we find the mean and standard deviation of the portfolio return R starting from information about X and Y ? We must now develop the machinery to do this.



Think first not about investments but about making refrigerators. Dimples and paint sags are two kinds of flaws in the painted finish of refrigerators. Not all refrigerators have the same number of dimples: many have none, some have one, some two, and so on. The inspectors report finding an average of 0.7 dimples and 1.4 sags per refrigerator. How many total imperfections of both kinds (on the average) are on a refrigerator? That's easy: if the average number of dimples is 0.7 and the average number of sags is 1.4, then counting both gives an average of $0.7 + 1.4 = 2.1$ flaws.

In more formal language, the number of dimples on a refrigerator is a random variable X that varies as we inspect one refrigerator after another. We know only that the mean number of dimples is $\mu_X = 0.7$. The number of paint sags is a second random variable Y having mean $\mu_Y = 1.4$. (As usual, the subscripts keep straight which variable we are talking about.) The total number of both dimples and sags is another random variable, the sum $X + Y$. Its mean μ_{X+Y} is the average number of dimples and sags together. It is just the sum of the individual means μ_X and μ_Y . That's an important rule for how means of random variables behave.

Here's another rule. A large lot of plastic coffee-can lids has mean diameter 4.2 inches. What is the mean in centimeters? There are 2.54 centimeters in an inch, so the diameter in centimeters of any lid is 2.54 times its diameter in inches. If we multiply every observation by 2.54, we also multiply their average by 2.54. The mean in centimeters must be 2.54×4.2 , or about 10.7 centimeters. More formally, the diameter in inches of a lid chosen at random from the lot is a random variable X with mean μ_X . The diameter in centimeters is $2.54X$, and this new random variable has mean $2.54\mu_X$.

The point of these examples is that means behave like averages. Here are the rules we need.

RULES FOR MEANS

Rule 1. If X is a random variable and a and b are fixed numbers, then

$$\mu_{a+bX} = a + b\mu_X$$

Rule 2. If X and Y are random variables, then

$$\mu_{X+Y} = \mu_X + \mu_Y$$

This is the addition rule for means.

EXAMPLE 4.15

Portfolio analysis

The past behavior of the two securities in Sadie's portfolio is pictured in Figure 4.14, which plots the annual returns on Treasury bills and common stocks for the years 1950 to 2000. We can calculate mean returns from the data behind the plot:¹⁰



$$\begin{array}{ll} X = \text{annual return on T-bills} & \mu_X = 5.2\% \\ Y = \text{annual return on stocks} & \mu_Y = 13.3\% \end{array}$$

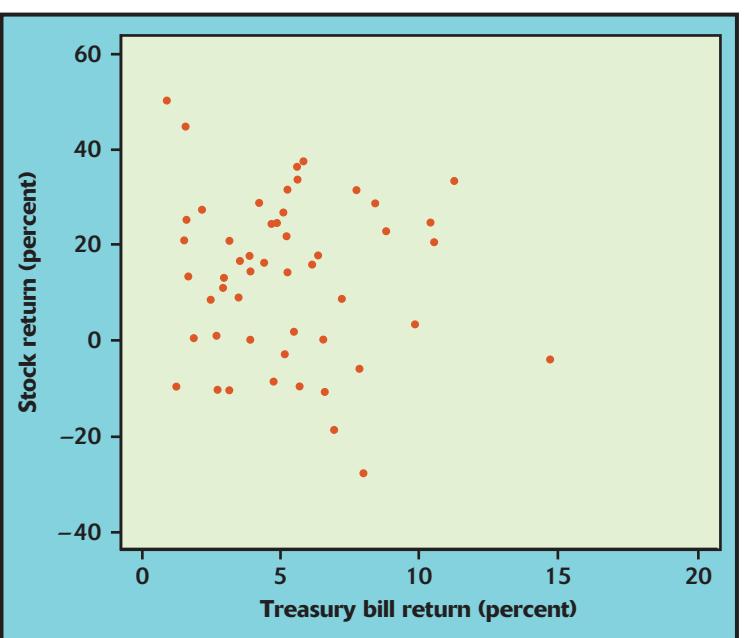


FIGURE 4.14 Annual returns on U.S. common stocks versus returns on Treasury bills for the years 1950 to 2000.

Sadie invests 20% in T-bills and 80% in stocks. We can find the mean return on her portfolio by combining Rules 1 and 2:

$$\begin{aligned} R &= 0.2X + 0.8Y \\ \mu_R &= 0.2\mu_X + 0.8\mu_Y \\ &= (0.2)(5.2) + (0.8)(13.3) = 11.68\% \end{aligned}$$

This calculation uses historical data on returns. Next year may of course be very different. It is usual in finance to use the term *expected return* in place of mean return. Remember our warning that we can't generally expect a future value to be close to the mean.

EXAMPLE 4.16

Gain Communications

Gain Communications sells aircraft communications units to both the military and the civilian markets. Next year's sales depend on market conditions that cannot be predicted exactly. Gain follows the modern practice of using probability estimates of sales. The military division estimates its sales as follows:

Units sold	1000	3000	5000	10,000
Probability	0.1	0.3	0.4	0.2

These are personal probabilities that express the informed opinion of Gain's executives. The corresponding sales estimates for the civilian division are

Units sold	300	500	750
Probability	0.4	0.5	0.1

Take X to be the number of military units sold and Y the number of civilian units. From the probability distributions we compute that

$$\begin{aligned}\mu_X &= (1000)(0.1) + (3000)(0.3) + (5000)(0.4) + (10,000)(0.2) \\ &= 100 + 900 + 2000 + 2000 = 5000 \text{ units} \\ \mu_Y &= (300)(0.4) + (500)(0.5) + (750)(0.1) \\ &= 120 + 250 + 75 = 445 \text{ units}\end{aligned}$$

Gain makes a profit of \$2000 on each military unit sold and \$3500 on each civilian unit. Next year's profit from military sales will be $2000X$, \$2000 times the number X of units sold. By Rule 1 for means, the mean military profit is

$$\mu_{2000X} = 2000\mu_X = (2000)(5000) = \$10,000,000$$

Similarly, the civilian profit is $3500Y$, and the mean profit from civilian sales is

$$\mu_{3500Y} = 3500\mu_Y = (3500)(445) = \$1,557,500$$

The total profit is the sum of the military and civilian profit:

$$Z = 2000X + 3500Y$$

Rule 2 for means says that the mean of this sum of two variables is the sum of the two individual means:

$$\begin{aligned}\mu_Z &= \mu_{2000X} + \mu_{3500Y} \\ &= \$10,000,000 + \$1,557,500 \\ &= \$11,557,500\end{aligned}$$

This mean is the company's best estimate of next year's profit, combining the probability estimates of the two divisions. We can do this calculation more quickly by combining Rules 1 and 2:

$$\begin{aligned}\mu_Z &= \mu_{2000X+3500Y} \\ &= 2000\mu_X + 3500\mu_Y \\ &= (2000)(5000) + (3500)(445) = \$11,557,500\end{aligned}$$

APPLY YOUR KNOWLEDGE

- 4.45 Selling mobile phones. You own a mobile-phone store with two locations, one in the local mall and one downtown on Main Street. Let X be the number of phones sold during the next month at the mall location, and let Y be the

number of phones sold during the next month at the downtown location. For the mall location, you estimate the following probability distribution:

Phones sold	200	300	400
Probability	0.4	0.4	0.2

For the downtown location, your estimated distribution is

Phones sold	100	200	300	400
Probability	0.3	0.5	0.15	0.05

- (a) Calculate μ_X and μ_Y .
 - (b) Your rental cost is higher at the mall location. You make \$25 profit on each phone sold at the mall location and \$35 profit on each phone sold at the downtown location. Calculate the mean profit for each location.
 - (c) Calculate the mean profit for both locations combined.
- 4.46 **Mutual funds.** The addition rule for means extends to sums of any number of random variables. Let's look at a portfolio containing three mutual funds. The monthly returns on Fidelity Magellan Fund, Fidelity Real Estate Fund, and Fidelity Japan Fund for the 36 months ending in December 2000 had approximately these means:¹¹



$$\begin{array}{ll} W = \text{Magellan monthly return} & \mu_W = 1.14\% \\ X = \text{Real Estate monthly return} & \mu_X = 0.16\% \\ Y = \text{Japan monthly return} & \mu_Y = 1.59\% \end{array}$$

What is the mean monthly return for a portfolio consisting of 50% Magellan, 30% Real Estate, and 20% Japan?

The variance of a random variable

The mean portfolio return in Example 4.15 is less than the mean return for stocks because Sadie added 20% Treasury bills. What does she gain by giving up some average return? T-bills are much less volatile than stocks, so she hopes that her portfolio returns will vary less from year to year than would be the case if she invested all her funds in stocks. The mean is a measure of the center of a distribution. Even the most basic numerical description requires in addition a measure of the spread or variability of the distribution.

The variance and the standard deviation are the measures of spread that accompany the choice of the mean to measure center. Just as for the mean, we need a distinct symbol to distinguish the variance of a random variable from the variance s^2 of a data set. We write the variance of a random variable X as σ_X^2 . The definition of the variance σ_X^2 of a random variable is similar to the definition of the sample variance s^2 given in Chapter 1. That is, the variance is an average of the squared deviation $(X - \mu_X)^2$ of the variable X from its mean μ_X . As for the mean, the average we use is a weighted average in which

each outcome is weighted by its probability to take account of outcomes that are not equally likely. Calculating this weighted average is straightforward for discrete random variables but requires advanced mathematics in the continuous case. Here is the definition.

VARIANCE OF A DISCRETE RANDOM VARIABLE

Suppose that X is a discrete random variable whose distribution is

Value of X	x_1	x_2	x_3	\cdots	x_k
Probability	p_1	p_2	p_3	\cdots	p_k

and that μ is the mean of X . The **variance** of X

$$\begin{aligned}\sigma_X^2 &= (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + \cdots + (x_k - \mu_X)^2 p_k \\ &= \sum (x_i - \mu_X)^2 p_i\end{aligned}$$

The **standard deviation** σ_X of X is the square root of the variance.

EXAMPLE 4.17

Gain Communications

In Example 4.16 we saw that the number X of communications units that the Gain Communications military division hopes to sell has distribution

Units sold	1000	3000	5000	10,000
Probability	0.1	0.3	0.4	0.2

We can find the mean and variance of X by arranging the calculation in the form of a table. Both μ_X and σ_X^2 are sums of columns in this table.

x_i	p_i	$x_i p_i$	$(x_i - \mu_X)^2 p_i$
1,000	0.1	100	$(1,000 - 5,000)^2(0.1) = 1,600,000$
3,000	0.3	900	$(3,000 - 5,000)^2(0.3) = 1,200,000$
5,000	0.4	2,000	$(5,000 - 5,000)^2(0.4) = 0$
10,000	0.2	2,000	$(10,000 - 5,000)^2(0.2) = 5,000,000$
$\mu_X = 5,000$			$\sigma_X^2 = 7,800,000$

We see that $\sigma_X^2 = 7,800,000$. The standard deviation of X is $\sigma_X = \sqrt{7,800,000} = 2792.8$. The standard deviation is a measure of how variable the number of units sold is. As in the case of distributions for data, the standard deviation of a probability distribution is easiest to understand for Normal distributions.



APPLY YOUR KNOWLEDGE

- 4.47 **Civilian sales.** Example 4.16 also gives the distribution of Gain's civilian sales Y . Find the variance σ_Y^2 and the standard deviation σ_Y .
- 4.48 **Hard-drive sizes.** Example 4.9 (page 255) gives the distribution of hard-drive sizes for sales of a particular model computer. You found the mean hard-drive size in Exercise 4.43 (page 264). Find the standard deviation of the distribution of hard-drive sizes.
- 4.49 **Selling mobile phones.** Exercise 4.45 gives the distribution of weekly mobile phone sales in two locations.
- Calculate the variance and the standard deviation of the number of phones sold at the mall location.
 - Calculate σ_Y^2 and σ_Y for the downtown location.

Rules for variances

What are the facts for variances that parallel Rules 1 and 2 for means? The mean of a sum of random variables is always the sum of their means, but this addition rule is not always true for variances. To understand why, take X to be the percent of a family's after-tax income that is spent and Y the percent that is saved. When X increases, Y decreases by the same amount. Though X and Y may vary widely from year to year, their sum $X + Y$ is always 100% and does not vary at all. It is the association between the variables X and Y that prevents their variances from adding. If random variables are *independent*, this kind of association between their values is ruled out and their variances do add. Two random variables X and Y are **independent** if knowing that any event involving X alone did or did not occur tells us nothing about the occurrence of any event involving Y alone. Probability models often assume independence when the random variables describe outcomes that appear unrelated to each other. You should ask in each instance whether the assumption of independence seems reasonable.

When random variables are not independent, the variance of their sum depends on the **correlation** between them as well as on their individual variances. In Chapter 2, we met the correlation r between two observed variables measured on the same individuals. We defined (page 103) the correlation r as an average of the products of the standardized x and y observations. The correlation between two random variables is defined in the same way, once again using a weighted average with probabilities as weights. We won't give the details—it is enough to know that the correlation between two random variables has the same basic properties as the correlation r calculated from data. We use ρ , the Greek letter rho, for the correlation between two random variables. The correlation ρ is a number between -1 and 1 that measures the direction and strength of the linear relationship between two variables. **The correlation between two independent random variables is zero.**

Returning to family finances, if X is the percent of a family's after-tax income that is spent and Y the percent that is saved, then $Y = 100 - X$. This is a perfect linear relationship with a negative slope, so the correlation between X and Y is $\rho = -1$. With the correlation at hand, we can state the rules for manipulating variances.

independence

correlation

RULES FOR VARIANCES

Rule 1. If X is a random variable and a and b are fixed numbers, then

$$\sigma_{a+bX}^2 = b^2 \sigma_X^2$$

Rule 2. If X and Y are independent random variables, then

$$\begin{aligned}\sigma_{X+Y}^2 &= \sigma_X^2 + \sigma_Y^2 \\ \sigma_{X-Y}^2 &= \sigma_X^2 + \sigma_Y^2\end{aligned}$$

This is the **addition rule for variances of independent random variables**.

Rule 3. If X and Y have correlation ρ , then

$$\begin{aligned}\sigma_{X+Y}^2 &= \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y \\ \sigma_{X-Y}^2 &= \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y\end{aligned}$$

This is the **general addition rule for variances of random variables**.

Notice that because a variance is the average of *squared* deviations from the mean, multiplying X by a constant b multiplies σ_X^2 by the *square* of the constant. Adding a constant a to a random variable changes its mean but does not change its variability. The variance of $X + a$ is therefore the same as the variance of X . Because the square of -1 is 1 , the addition rule says that the variance of a difference of independent random variables is the *sum* of the variances. For independent random variables, the difference $X - Y$ is more variable than either X or Y alone because variations in both X and Y contribute to variation in their difference.

As with data, we prefer the standard deviation to the variance as a measure of the variability of a random variable. Rule 2 for variances implies that standard deviations of independent random variables do *not* add. To combine standard deviations, use the rules for variances rather than trying to remember separate rules for standard deviations. For example, the standard deviations of $2X$ and $-2X$ are both equal to $2\sigma_X$ because this is the square root of the variance $4\sigma_X^2$.

EXAMPLE 4.18**Pick 3**

The payoff X of a \$1 ticket in the Tri-State Pick 3 game is \$500 with probability $1/1000$ and 0 the rest of the time. Here is the combined calculation of mean and variance:

x_i	p_i	$x_i p_i$	$(x_i - \mu_X)^2 p_i$
0	0.999	0	$(0 - 0.5)^2(0.999) = 0.2497$
500	0.001	0.5	$(500 - 0.5)^2(0.001) = 249.50025$
		$\mu_X = 0.5$	$\sigma_X^2 = 249.75$

The standard deviation is $\sigma_X = \sqrt{249.75} = \15.80 . Games of chance usually have large standard deviations because large variability makes gambling exciting.

If you buy a Pick 3 ticket, your net winnings are $W = X - 1$ because the dollar you pay for the ticket must be subtracted from the payoff. By the rules for means, the mean amount you win is

$$\mu_W = \mu_X - 1 = -\$0.50$$

That is, you lose an average of 50 cents on a ticket. The rules for variances remind us that the variance and standard deviation of the winnings $W = X - 1$ are the same as those of the payoff X . Subtracting a fixed number changes the mean but not the variance.

Suppose now that you buy a \$1 ticket on each of two different days. The payoffs X and Y on the two tickets are independent because separate drawings are held each day. Your total payoff $X + Y$ has mean

$$\mu_{X+Y} = \mu_X + \mu_Y = \$0.50 + \$0.50 = \$1.00$$

Because X and Y are independent, the variance of $X + Y$ is

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 = 249.75 + 249.75 = 499.5$$

The standard deviation of the total payoff is

$$\sigma_{X+Y} = \sqrt{499.5} = \$22.35$$

This is not the same as the sum of the individual standard deviations, which is $\$15.80 + \$15.80 = \$31.60$. Variances of independent random variables add; standard deviations do not.

EXAMPLE 4.19

SAT scores

A college uses SAT scores as one criterion for admission. Experience has shown that the distribution of SAT scores among the college's entire population of applicants is such that

$$\begin{aligned} X &= \text{SAT math score} & \mu_X &= 625 & \sigma_X &= 90 \\ Y &= \text{SAT verbal score} & \mu_Y &= 590 & \sigma_Y &= 100 \end{aligned}$$

What are the mean and standard deviation of the total score $X + Y$ among students applying to this college?

The mean overall SAT score is

$$\mu_{X+Y} = \mu_X + \mu_Y = 625 + 590 = 1215$$

The variance and standard deviation of the total *cannot be computed* from the information given. SAT verbal and math scores are not independent, because students who score high on one exam tend to score high on the other also. Therefore, Rule 2 does not apply and we need to know ρ , the correlation between X and Y , to apply Rule 3.

Nationally, the correlation between SAT math and verbal scores is about $\rho = 0.7$. If this is true for these students,

$$\begin{aligned}\sigma_{X+Y}^2 &= \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y \\ &= (90)^2 + (100)^2 + (2)(0.7)(90)(100) \\ &= 30,700\end{aligned}$$

The variance of the sum $X + Y$ is greater than the sum of the variances $\sigma_X^2 + \sigma_Y^2$ because of the positive correlation between SAT math scores and SAT verbal scores. We find the standard deviation from the variance,

$$\sigma_{X+Y} = \sqrt{30,700} = 175$$



EXAMPLE 4.20

Portfolio analysis



Now we can complete our initial analysis of Sadie's portfolio. Based on annual returns between 1950 and 2000, we have

$$\begin{aligned}X &= \text{annual return on T-bills} & \mu_X &= 5.2\% & \sigma_X &= 2.9\% \\ Y &= \text{annual return on stocks} & \mu_Y &= 13.3\% & \sigma_Y &= 17.0\% \\ \text{Correlation between } X \text{ and } Y: & & \rho &= -0.1\end{aligned}$$

We see that stocks had higher returns than Treasury bills on the average but were also much more volatile. This is a basic relationship in finance: a more volatile investment must offer a higher return to compensate investors for its greater risk. For the return R on Sadie's portfolio of 20% T-bills and 80% stocks,

$$\begin{aligned}R &= 0.2X + 0.8Y \\ \mu_R &= 0.2\mu_X + 0.8\mu_Y \\ &= (0.2 \times 5.2) + (0.8 \times 13.3) = 11.68\%\end{aligned}$$

To find the variance of the portfolio return, combine Rules 1 and 3:

$$\begin{aligned}\sigma_R^2 &= \sigma_{0.2X}^2 + \sigma_{0.8Y}^2 + 2\rho\sigma_{0.2X}\sigma_{0.8Y} \\ &= (0.2)^2\sigma_X^2 + (0.8)^2\sigma_Y^2 + 2\rho(0.2 \times \sigma_X)(0.8 \times \sigma_Y) \\ &= (0.2)^2(2.9)^2 + (0.8)^2(17.0)^2 + (2)(-0.1)(0.2 \times 2.9)(0.8 \times 17.0) \\ &= 183.719 \\ \sigma_R &= \sqrt{183.719} = 13.55\%\end{aligned}$$

The portfolio has a smaller mean return than an all-stock portfolio, but it is also less volatile. As a proportion of the all-stock values, the reduction in standard deviation is greater than the reduction in mean return. That's why Sadie put some funds into Treasury bills.



Example 4.20 illustrates the first step in modern finance, using mean and standard deviation to describe the behavior of a portfolio. The next step is to seek best-possible portfolios based on a given set of securities by finding the combination with the highest return for a specified level of risk. The investor says how much risk (measured as standard deviation of returns) she is willing to bear; the best portfolio has the highest mean return for this standard deviation. Our examples also suggest the characteristic weakness of portfolio analysis: risk and return in the future may be very different from the σ and μ we get from past returns.

APPLY YOUR KNOWLEDGE

- 4.50 **Comparing sales.** Tamara and Derek are sales associates in a large electronics and appliance store. Their store tracks each associate's daily sales in dollars. Tamara's sales total X varies from day to day with mean and standard deviation

$$\mu_X = \$1100 \text{ and } \sigma_X = \$100$$

Derek's sales total Y also varies, with

$$\mu_Y = \$1000 \text{ and } \sigma_Y = \$80$$

Because the store is large and Tamara and Derek work in different departments, we might assume that their daily sales totals vary independently of each other. What are the mean and standard deviation of the difference $X - Y$ between Tamara's daily sales and Derek's daily sales? Tamara sells more on the average. Do you think she sells more every day? Why?

- 4.51 **Comparing sales.** It is unlikely that the daily sales of Tamara and Derek in the previous problem are independent. They will both sell more during the Christmas season, for example. Suppose that the correlation between their sales is $\rho = 0.4$. What are now the mean and standard deviation of the difference $X - Y$? Can you explain conceptually why positive correlation between two variables reduces the variability of the difference between them?
- 4.52 Exercise 4.45 (page 268) gives the distributions of X , the number of wireless phones sold at the mall location of a store during the next month, and Y , the number of wireless phones sold at the downtown location during the next month. You did some useful variance calculations in Exercise 4.49 (page 271). Each phone sold at the mall location results in \$25 profit, and each phone sold at the downtown location results in \$35 profit.
- Calculate the standard deviation of the profit for each location using Rule 1 for variances.
 - Assuming phone sales at the two locations are independent, calculate the standard deviation for total profit of both locations combined.
 - Assuming $\rho = 0.8$, calculate the standard deviation for total profit of both locations combined.
 - Assuming $\rho = 0$, calculate the standard deviation for total profit of both locations combined. How does this compare with your result in part (b)? In part (c)?
 - Assuming $\rho = -0.8$, calculate the standard deviation for total profit of both locations combined. How does this compare with your result in part (b)? In part (c)? In part (d)?

SECTION 4.3 SUMMARY

- A random variable is a variable taking numerical values determined by the outcome of a random phenomenon. The **probability distribution** of a random variable X tells us what the possible values of X are and how probabilities are assigned to those values.
- A random variable X and its distribution can be **discrete** or **continuous**.
- A **discrete random variable** has finitely many possible values. The probability distribution assigns each of these values a probability between 0 and 1 such that the sum of all the probabilities is exactly 1. The probability of any event is the sum of the probabilities of all the values that make up the event.
- A **continuous random variable** takes all values in some interval of numbers. A **density curve** describes the probability distribution of a continuous random variable. The probability of any event is the area under the curve above the values that make up the event. **Normal distributions** are one type of continuous probability distribution.
- You can picture a probability distribution by drawing a **probability histogram** in the discrete case or by graphing the density curve in the continuous case.
- The probability distribution of a random variable X , like a distribution of data, has a **mean** μ_X and a **standard deviation** σ_X .
- The **mean** μ is the balance point of the probability histogram or density curve. If X is discrete with possible values x_i having probabilities p_i , the mean is the average of the values of X , each weighted by its probability:

$$\mu_X = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

- The **variance** σ_X^2 is the average squared deviation of the values of the variable from their mean. For a discrete random variable,

$$\sigma_X^2 = (x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2 + \cdots + (x_k - \mu)^2 p_k$$

- The **standard deviation** σ_X is the square root of the variance. The standard deviation measures the variability of the distribution about the mean. It is easiest to interpret for Normal distributions.
- The mean and variance of a continuous random variable can be computed from the density curve, but to do so requires more advanced mathematics.
- The means and variances of random variables obey the following rules. If a and b are fixed numbers, then

$$\begin{aligned}\mu_{a+bX} &= a + b\mu_X \\ \sigma_{a+bX}^2 &= b^2 \sigma_X^2\end{aligned}$$

If X and Y are any two random variables having correlation ρ , then

$$\begin{aligned}\mu_{X+Y} &= \mu_X + \mu_Y \\ \sigma_{X+Y}^2 &= \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y \\ \sigma_{X-Y}^2 &= \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y\end{aligned}$$

If X and Y are independent, then $\rho = 0$. In this case,

$$\begin{aligned}\sigma_{X+Y}^2 &= \sigma_X^2 + \sigma_Y^2 \\ \sigma_{X-Y}^2 &= \sigma_X^2 + \sigma_Y^2\end{aligned}$$

SECTION 4.3 EXERCISES

- 4.53 How many rooms?** Furniture makers and others are interested in how many rooms housing units have, because more rooms can generate more sales. Here are the distributions of the number of rooms for owner-occupied units and renter-occupied units in San Jose, California.¹²

Rooms	1	2	3	4	5	6	7	8	9	10
Owned	0.003	0.002	0.023	0.104	0.210	0.224	0.197	0.149	0.053	0.035
Rented	0.008	0.027	0.287	0.363	0.164	0.093	0.039	0.013	0.003	0.003

- (a) Make probability histograms of these two distributions, using the same scales. What are the most important differences between the distributions for owner-occupied and rented housing units?
(b) Find the mean number of rooms for both types of housing unit. How do the means reflect the differences you found in (a)?
- 4.54 Households and families.** In government data, a household consists of all occupants of a dwelling unit, while a family consists of two or more persons who live together and are related by blood or marriage. Here are the distributions of household size and of family size in the United States:

Number of persons	1	2	3	4	5	6	7
Household probability	0.25	0.32	0.17	0.15	0.07	0.03	0.01
Family probability	0	0.42	0.23	0.21	0.09	0.03	0.02

You have considered the distribution of household size in Exercises 4.35 and 4.42. Compare the two distributions using probability histograms, means, and standard deviations. Write a brief comparison, using your calculations to back up your statements.

- 4.55 How many rooms?** Which of the two distributions for room counts in Exercise 4.53 appears more spread out in the probability histograms? Why? Find the standard deviation for both distributions. The standard deviation provides a numerical measure of spread.

- 4.56 **Tossing four coins.** The distribution of the count X of heads in four tosses of a balanced coin was found in Example 4.10 to be

Number of heads x_i	0	1	2	3	4
Probability p_i	0.0625	0.25	0.375	0.25	0.0625

Find the mean μ_x from this distribution. Then find the mean number of heads for a single coin toss and show that your two results are related by the addition rule for means.

- 4.57 **Pick 3 once more.** The Tri-State Pick 3 lottery game offers a choice of several bets. You choose a three-digit number. The lottery commission announces the winning three-digit number, chosen at random, at the end of each day. The “box” pays \$83.33 if the number you choose has the same digits as the winning number, in any order. Find the expected payoff for a \$1 bet on the box. (Assume that you chose a number having three different digits.)
- 4.58 **Tossing four coins.** The distribution of outcomes for tossing four coins given in Exercise 4.56 assumes that the tosses are independent of each other. Find the variance for four tosses. Then find the variance of the number of heads (0 or 1) on a single toss and show that your two results are related by the addition rule for variances of independent random variables.
- 4.59 **Independent random variables?** For each of the following situations, would you expect the random variables X and Y to be independent? Explain your answers.
- (a) X is the rainfall (in inches) on November 6 of this year and Y is the rainfall at the same location on November 6 of next year.
 - (b) X is the amount of rainfall today and Y is the rainfall at the same location tomorrow.
 - (c) X is today’s rainfall at the Orlando, Florida, airport, and Y is today’s rainfall at Disney World just outside Orlando.
- 4.60 **Independent random variables?** In which of the following games of chance would you be willing to assume independence of X and Y in making a probability model? Explain your answer in each case.
- (a) In blackjack, you are dealt two cards and examine the total points X on the cards (face cards count 10 points). You can choose to be dealt another card and compete based on the total points Y on all three cards.
 - (b) In craps, the betting is based on successive rolls of two dice. X is the sum of the faces on the first roll, and Y the sum of the faces on the next roll.
- 4.61 **Time and motion studies.** A time and motion study measures the time required for an assembly line worker to perform a repetitive task. The data show that the time required to bring a part from a bin to its position on an automobile chassis varies from car to car with mean 11 seconds and standard deviation 2 seconds. The time required to attach the part to the chassis varies with mean 20 seconds and standard deviation 4 seconds.
- (a) What is the mean time required for the entire operation of positioning and attaching the part?

- (b) If the variation in the worker's performance is reduced by better training, the standard deviations will decrease. Will this decrease change the mean you found in (a) if the mean times for the two steps remain as before?
- (c) The study finds that the times required for the two steps are independent. A part that takes a long time to position, for example, does not take more or less time to attach than other parts. How would your answer in (a) change if the two variables were dependent, with correlation 0.8? With correlation 0.3?
- 4.62 **A chemical production process.** Laboratory data show that the time required to complete two chemical reactions in a production process varies. The first reaction has a mean time of 40 minutes and a standard deviation of 2 minutes; the second has a mean time of 25 minutes and a standard deviation of 1 minute. The two reactions are run in sequence during production. There is a fixed period of 5 minutes between them as the product of the first reaction is pumped into the vessel where the second reaction will take place. What is the mean time required for the entire process?
- 4.63 **Time and motion studies.** Find the standard deviation of the time required for the two-step assembly operation studied in Exercise 4.61, assuming that the study shows the two times to be independent. Redo the calculation assuming that the two times are dependent, with correlation 0.3. Can you explain in nontechnical language why positive correlation increases the variability of the total time?
- 4.64 **A chemical production process.** The times for the two reactions in the chemical production process described in Exercise 4.62 are independent. Find the standard deviation of the time required to complete the process.
- 4.65 **Get the mean you want.** Here is a simple way to create a random variable X that has mean μ and standard deviation σ : X takes only the two values $\mu - \sigma$ and $\mu + \sigma$, each with probability 0.5. Use the definition of the mean and variance for discrete random variables to show that X does have mean μ and standard deviation σ .
- 4.66 **Combining measurements.** You have two scales for measuring weight. Both scales give answers that vary a bit in repeated weighings of the same item. If the true weight of an item is 2 grams (g), the first scale produces readings X that have mean 2.000 g and standard deviation 0.002 g. The second scale's readings Y have mean 2.001 g and standard deviation 0.001 g.
- What are the mean and standard deviation of the difference $Y - X$ between the readings? (The readings X and Y are independent.)
 - You measure once with each scale and average the readings. Your result is $Z = (X + Y)/2$. What are μ_Z and σ_Z ? Is the average Z more or less variable than the reading Y of the less variable scale?
- 4.67 **Gain Communications.** Examples 4.16 and 4.17 concern a probabilistic projection of sales and profits by an electronics firm, Gain Communications. The mean and variance of military sales X appear in Example 4.17 (page 270). You found the mean and variance of civilian sales Y in Exercise 4.47 (page 271).
- Because the military budget and the civilian economy are not closely linked, Gain is willing to assume that its military and civilian sales

- vary independently. What is the standard deviation of Gain's total sales $X + Y$?
- (b) Find the standard deviation of the estimated profit, $Z = 2000X + 3500Y$.
- 4.68 **Gain Communications, continued.** Redo Exercise 4.67 assuming correlation 0.01 between civilian and military sales. Redo the exercise assuming correlation 0.99. Comment on the effect of small and large correlations on the uncertainty of Gain's sales projections.
- 4.69 **Study habits.** The academic motivation and study habits of female students as a group are better than those of males. The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures these factors. The distribution of SSHA scores among the women at a college has mean 120 and standard deviation 28, and the distribution of scores among men students has mean 105 and standard deviation 35. You select a single male student and a single female student at random and give them the SSHA test.
- (a) Explain why it is reasonable to assume that the scores of the two students are independent.
 - (b) What are the mean and standard deviation of the difference (female minus male) between their scores?
 - (c) From the information given, can you find the probability that the woman chosen scores higher than the man? If so, find this probability. If not, explain why you cannot.

Portfolio analysis. Here are the means, standard deviations, and correlations for the monthly returns from three Fidelity mutual funds for the 36 months ending in December 2000.¹³ Because there are three random variables, there are three correlations. We use subscripts to show which pair of random variables a correlation refers to.

$$\begin{aligned} W &= \text{monthly return on Magellan Fund} & \mu_W &= 1.14\% & \sigma_W &= 4.64\% \\ X &= \text{monthly return on Real Estate Fund} & \mu_X &= 0.16\% & \sigma_X &= 3.61\% \\ Y &= \text{monthly return on Japan Fund} & \mu_Y &= 1.59\% & \sigma_Y &= 6.75\% \end{aligned}$$

Correlations

$$\rho_{WX} = 0.19 \quad \rho_{WY} = 0.54 \quad \rho_{XY} = -0.17$$

Exercises 4.70 to 4.72 make use of these historical data.

- 4.70 **Diversification.** Many advisors recommend using roughly 20% foreign stocks to diversify portfolios of U.S. stocks. Michael owns Fidelity Magellan Fund, which concentrates on stocks of large American companies. He decides to move to a portfolio of 80% Magellan and 20% Fidelity Japan Fund. Show that (based on historical data) this portfolio has both a *higher* mean return and *less* volatility than Magellan alone. This illustrates the beneficial effects of diversifying among investments.
- 4.71 **More on diversification.** Diversification works better when the investments in a portfolio have small correlations. To demonstrate this, suppose that returns on Magellan Fund and Japan Fund had the means and standard deviations we have given but were uncorrelated ($\rho_{WY} = 0$). Show that the standard deviation of a portfolio that combines



80% Magellan with 20% Japan is then smaller than your result from the previous exercise. What happens to the mean return if the correlation is 0?

- 4.72 **(Optional) Larger portfolios.** Portfolios often contain more than two investments. The rules for means and variances continue to apply, though the arithmetic gets messier. A portfolio containing proportions a of Magellan Fund, b of Real Estate Fund, and c of Japan Fund has return $R = aW + bX + cY$. Because a , b , and c are the proportions invested in the three funds, $a + b + c = 1$. The mean and variance of the portfolio return R are:

$$\begin{aligned}\mu_R &= a\mu_W + b\mu_X + c\mu_Y \\ \sigma_R^2 &= a^2\sigma_W^2 + b^2\sigma_X^2 + c^2\sigma_Y^2 + 2ab\rho_{WX}\sigma_W\sigma_X + 2ac\rho_{WY}\sigma_W\sigma_Y \\ &\quad + 2bc\rho_{XY}\sigma_X\sigma_Y\end{aligned}$$



Having seen the advantages of diversification, Michael decides to invest his funds 60% in Magellan, 20% in Real Estate, and 20% in Japan. What are the (historical) mean and standard deviation of the monthly returns for this portfolio?

- 4.73 **(Optional) Perfectly correlated variables.** We know that variances add if the random variables involved are uncorrelated ($\rho = 0$), but not otherwise. The opposite extreme is perfect positive correlation ($\rho = 1$). Show by using the general addition rule for variances that in this case the standard deviations add. That is, $\sigma_{X+Y} = \sigma_X + \sigma_Y$ if $\rho_{XY} = 1$.
- 4.74 **A hot stock.** You purchase a hot stock for \$1000. The stock either gains 30% or loses 25% each day, each with probability 0.5. Its returns on consecutive days are independent of each other. This implies that all four possible combinations of gains and losses in two days are equally likely, each having probability 0.25. You plan to sell the stock after two days.
- (a) What are the possible values of the stock after two days, and what is the probability for each value? What is the probability that the stock is worth more after two days than the \$1000 you paid for it?
 - (b) What is the mean value of the stock after two days? You see that these two criteria give different answers to the question “Should I invest?”
- 4.75 **Making glassware.** In a process for manufacturing glassware, glass stems are sealed by heating them in a flame. The temperature of the flame varies a bit. Here is the distribution of the temperature X measured in degrees Celsius:

Temperature	540°	545°	550°	555°	560°
Probability	0.1	0.25	0.3	0.25	0.1

- (a) Find the mean temperature μ_X and the standard deviation σ_X .
- (b) The target temperature is 550° C. Use the rules for means and variances to find the mean and standard deviation of the number of degrees off target, $X - 550$.

- (c) A manager asks for results in degrees Fahrenheit. The conversion of X into degrees Fahrenheit is given by

$$Y = \frac{9}{5}X + 32$$

What are the mean μ_Y and standard deviation σ_Y of the temperature of the flame in the Fahrenheit scale?

- 4.76 Irrational decision making?** The psychologist Amos Tversky did many studies of our perception of chance behavior. In its obituary of Tversky (June 6, 1996), the *New York Times* cited the following example.
- (a) Tversky asked subjects to choose between two public health programs that affect 600 people. One has probability $1/2$ of saving all 600 and probability $1/2$ that all 600 will die. The other is guaranteed to save exactly 400 of the 600 people. Find the mean number of people saved by the first program.
 - (b) Tversky then offered a different choice. One program has probability $1/2$ of saving all 600 and probability $1/2$ of losing all 600, while the other will definitely lose exactly 200 lives. What is the difference between this choice and that in (a)?
 - (c) Given option (a), most people choose the second program. Given option (b), most people choose the first program. Do people appear to use means in making their decisions? Why do you think their choices differed in the two cases?

4.4 The Sampling Distribution of a Sample Mean

Section 3.3 (page 206) motivated our study of probability by looking at *sampling variability* and *sampling distributions*. A statistic from a random sample will take different values if we take more samples from the same population. That is, sample statistics are random variables. The values of a statistic do not vary haphazardly from sample to sample but have a regular pattern, the sampling distribution, in many samples. This is the distribution of the random variable. We can now use the language of probability to examine one very important sampling distribution, that of a sample mean \bar{x} . Some new and important facts about probability emerge from this examination. Here is the example we will follow to introduce the role of probability in statistical inference.

EXAMPLE 4.21

Does this wine smell bad?

Sulfur compounds such as dimethyl sulfide (DMS) are sometimes present in wine. DMS causes “off-odors” in wine, so winemakers want to know the odor threshold, the lowest concentration of DMS that the human nose can detect. Different people have different thresholds, so we start by asking about the mean threshold μ in the population of all adults. The number μ is a *parameter* that describes this population.

To estimate μ , we present tasters with both natural wine and the same wine spiked with DMS at different concentrations to find the lowest concentration at

which they can identify the spiked wine. Here are the odor thresholds (measured in micrograms of DMS per liter of wine) for 10 randomly chosen subjects:

28 40 28 33 20 31 29 27 17 21

The mean threshold for these subjects is $\bar{x} = 27.4$. This sample mean is a *statistic* that we use to estimate the parameter μ , but it is probably not exactly equal to μ . Moreover, we know that a different 10 subjects would give us a different \bar{x} .

A parameter, such as the mean odor threshold μ of all adults, is in practice a fixed but unknown number. A statistic, such as the mean threshold \bar{x} of a random sample of 10 adults, is a random variable. It seems reasonable to use \bar{x} to estimate μ . An SRS should fairly represent the population, so the mean \bar{x} of the sample should be somewhere near the mean μ of the population. Of course, we don't expect \bar{x} to be exactly equal to μ , and we realize that if we choose another SRS, the luck of the draw will probably produce a different \bar{x} .

Statistical estimation and the law of large numbers

If \bar{x} is rarely exactly right and varies from sample to sample, why is it nonetheless a reasonable estimate of the population mean μ ? Here is one answer: if we keep on taking larger and larger samples, the statistic \bar{x} is *guaranteed* to get closer and closer to the parameter μ . We have the comfort of knowing that if we can afford to keep on measuring more subjects, eventually we will estimate the mean odor threshold of all adults very accurately. This remarkable fact is called the *law of large numbers*. It is remarkable because it holds for *any* population, not just for some special class such as Normal distributions.

LAW OF LARGE NUMBERS

Draw independent observations at random from any population with finite mean μ . As the number of observations drawn increases, the mean \bar{x} of the observed values gets closer and closer to the mean μ of the population.

The mean μ of a random variable is the average value of the variable in two senses. By its definition, μ is the average of the possible values, weighted by their probability of occurring. The law of large numbers says that μ is also the long-run average of many independent observations on the variable. The law of large numbers can be proved mathematically starting from the basic laws of probability. The behavior of \bar{x} is similar to the idea of probability. In the long run, the proportion of outcomes taking any value gets close to the probability of that value, and the average outcome gets close to the population mean. Figure 4.1 (page 233) shows how proportions approach probability in one example. Here is an example of how sample means approach the population mean.

EXAMPLE 4.22**The law of large numbers in action**

In fact, the distribution of odor thresholds among all adults has mean 25. The mean $\mu = 25$ is the true value of the parameter we seek to estimate. Typically, the value of a parameter like μ is unknown. We use an example in which μ is known to illustrate facts about the general behavior of \bar{x} that hold even when μ is unknown. Figure 4.15 shows how the sample mean \bar{x} of an SRS drawn from this population changes as we add more subjects to our sample.

The first subject in Example 4.21 had threshold 28, so the line in Figure 4.15 starts there. The mean for the first two subjects is

$$\bar{x} = \frac{28 + 40}{2} = 34$$

This is the second point on the graph. At first, the graph shows that the mean of the sample changes as we take more observations. Eventually, however, the mean of the observations gets close to the population mean $\mu = 25$ and settles down at that value.

If we started over, again choosing people at random from the population, we would get a different path from left to right in Figure 4.15. The law of large numbers says that whatever path we get will always settle down at 25 as we draw more and more people.

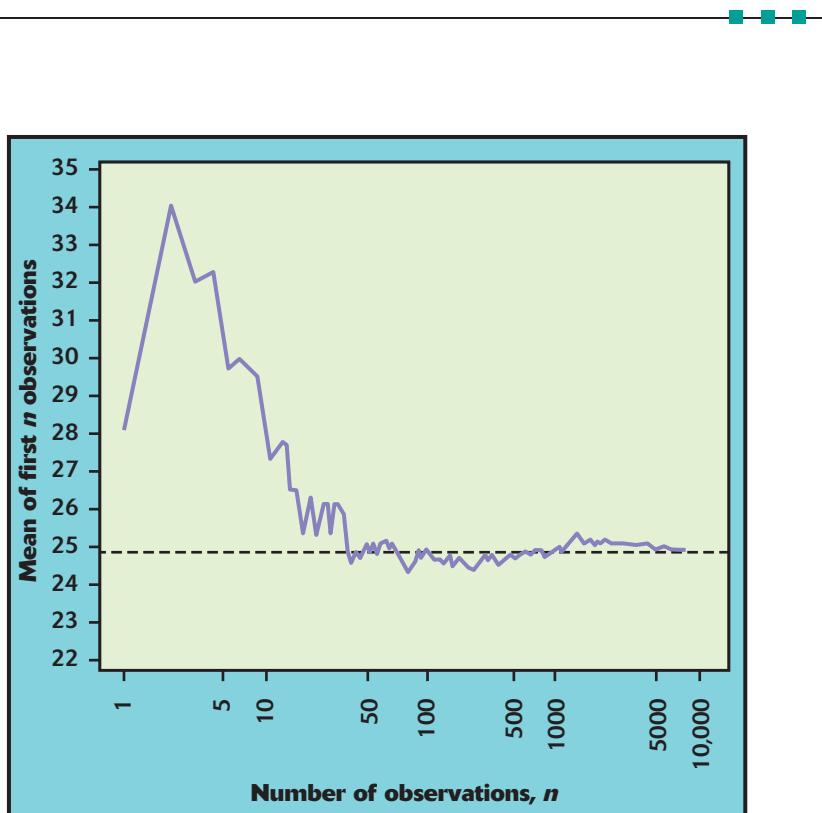


FIGURE 4.15 The law of large numbers in action: as we take more observations, the sample mean \bar{x} always approaches the mean μ of the population.



As in the case of probability, an animated version of Figure 4.15 makes the idea of the mean as long-run average outcome clearer. The *Mean of a Random Variable* applet at the text Web site, www.whfreeman.com/pbs, gives you control of an animation.

The law of large numbers is the foundation of such business enterprises as gambling casinos and insurance companies. The winnings (or losses) of a gambler on a few plays are uncertain—that's why gambling is exciting. In Figure 4.15, the mean of even 100 observations is not yet very close to μ . It is only *in the long run* that the mean outcome is predictable. The house plays tens of thousands of times. So the house, unlike individual gamblers, can count on the long-run regularity described by the law of large numbers. The average winnings of the house on tens of thousands of plays will be very close to the mean of the distribution of winnings. Needless to say, this mean guarantees the house a profit. That's why gambling can be a business.

APPLY YOUR KNOWLEDGE

- 4.77 **Comparing computers.** Pfeiffer Consulting, a technology consulting group, designed benchmark tests to compare the speed with which computer models complete a variety of tasks. Pfeiffer announced that the mean completion time was 15 minutes for 733-MHz Power Mac G4 models and 23.5 minutes for 400-MHz Power Mac G3 models.¹⁴ Do you think the bold numbers are parameters or statistics? Explain your reasoning carefully.
- 4.78 Figure 4.15 shows how the mean of n observations behaves as we keep adding more observations to those already in hand. The first 10 observations are given in Example 4.21. Demonstrate that you grasp the idea of Figure 4.15: find the mean of the first one, then two, three, four, and five of these observations and plot the successive means against n . Verify that your plot agrees with the first part of the plot in Figure 4.15.
- 4.79 **Playing the numbers.** The numbers racket is a well-entrenched illegal gambling operation in most large cities. One version works as follows. You choose one of the 1000 three-digit numbers 000 to 999 and pay your local numbers runner a dollar to enter your bet. Each day, one three-digit number is chosen at random and pays off \$600. The mean payoff for the population of thousands of bets is $\mu = 60$ cents. Joe makes one bet every day for many years. Explain what the law of large numbers says about Joe's results as he keeps on betting.

Thinking about the law of large numbers

The law of large numbers says broadly that the average results of many independent observations are stable and predictable. Casinos are not the only businesses that base forecasts on this fact. A grocery store deciding how many gallons of milk to stock and a fast-food restaurant deciding how many beef patties to prepare can predict demand even though their many customers make independent decisions. The law of large numbers says that these many individual decisions will produce a stable result. It is worth the effort to think a bit more closely about so important a fact.

The “Law of Small Numbers”

Both the rules of probability and the law of large numbers describe the regular behavior of chance phenomena *in the long run*. Psychologists have

discovered that the popular understanding of randomness is quite different from the true laws of chance.¹⁵ Most people believe in an incorrect “law of small numbers.” That is, we expect even short sequences of random events to show the kind of average behavior that in fact appears only in the long run.

Try this experiment: Write down a sequence of heads and tails that you think imitates 10 tosses of a balanced coin. How long was the longest string (called a *run*) of consecutive heads or consecutive tails in your tosses? Most people will write a sequence with no runs of more than two consecutive heads or tails. Longer runs don’t seem “random” to us. In fact, the probability of a run of three or more consecutive heads or tails in 10 tosses is greater than 0.8, and the probability of *both* a run of three or more heads and a run of three or more tails is almost 0.2.¹⁶ This and other probability calculations suggest that a short sequence of coin tosses will often not seem random to most people. The runs of consecutive heads or consecutive tails that appear in real coin tossing (and that are predicted by the mathematics of probability) surprise us. Because we don’t expect to see long runs, we may conclude that the coin tosses are not independent or that some influence is disturbing the random behavior of the coin.

Belief in the law of small numbers influences behavior. If a basketball player makes several consecutive shots, both the fans and his teammates believe that he has a “hot hand” and is more likely to make the next shot. This is doubtful. Careful study suggests that runs of baskets made or missed are no more frequent in basketball than would be expected if each shot were independent of the player’s previous shots. Players perform consistently, not in streaks. (Of course, some players make a higher percent of their shots in the long run than others.) Our perception of hot or cold streaks simply shows that we don’t perceive random behavior very well.¹⁷

Gamblers often follow the hot-hand theory, betting that a run will continue. At other times, however, they draw the opposite conclusion when confronted with a run of outcomes. If a coin gives 10 straight heads, some gamblers feel that it must now produce some extra tails to get back to the average of half heads and half tails. Not so. If the next 10,000 tosses give about 50% tails, those 10 straight heads will be swamped by the later thousands of heads and tails. No compensation is needed to get back to the average in the long run. Remember that it is *only* in the long run that the regularity described by probability and the law of large numbers takes over.

Our inability to accurately distinguish random behavior from systematic influences points out once more the need for statistical inference to supplement exploratory analysis of data. Probability calculations can help verify that what we see in the data is more than a random pattern.

How Large Is a Large Number?

The law of large numbers says that the actual mean outcome of many trials gets close to the distribution mean μ as more trials are made. It doesn’t say how many trials are needed to guarantee a mean outcome close to μ . That depends on the *variability* of the random outcomes. The more variable the

outcomes, the more trials are needed to ensure that the mean outcome \bar{x} is close to the distribution mean μ .

BEYOND THE BASICS: MORE LAWS OF LARGE NUMBERS

The law of large numbers is one of the central facts about probability. It helps us understand the mean μ of a random variable. It explains why gambling casinos and insurance companies make money. It assures us that statistical estimation will be accurate if we can afford enough observations. The basic law of large numbers applies to independent observations that all have the same distribution. Mathematicians have extended the law to many more general settings. Here are two of these.

Is There a Winning System for Gambling?

Serious gamblers often follow a system of betting in which the amount bet on each play depends on the outcome of previous plays. You might, for example, double your bet on each spin of the roulette wheel until you win—or, of course, until your fortune is exhausted. Such a system tries to take advantage of the fact that you have a memory even though the roulette wheel does not. Can you beat the odds with a system based on the outcomes of past plays? No. Mathematicians have established a stronger version of the law of large numbers that says that if you do not have an infinite fortune to gamble with, your long-run average winnings μ remain the same as long as successive trials of the game (such as spins of the roulette wheel) are independent.

What If Observations Are Not Independent?

You are in charge of a process that manufactures video screens for computer monitors. Your equipment measures the tension on the metal mesh that lies behind each screen and is critical to its image quality. You want to estimate the mean tension μ for the process by the average \bar{x} of the measurements. Alas, the tension measurements are not independent. If the tension on one screen is a bit too high, the tension on the next is more likely to also be high. Many real-world processes are like this—the process stays stable in the long run, but observations made close together are likely to be both above or both below the long-run mean. Again the mathematicians come to the rescue: as long as the dependence dies out fast enough as we take measurements farther and farther apart in time, the law of large numbers still holds.

APPLY YOUR KNOWLEDGE

4.80 Help this man.

- (a) A gambler knows that red and black are equally likely to occur on each spin of a roulette wheel. He observes five consecutive reds and bets heavily on red at the next spin. Asked why, he says that “red is hot” and that the run of reds is likely to continue. Explain to the gambler what is wrong with this reasoning.
- (b) After hearing you explain why red and black remain equally probable after five reds on the roulette wheel, the gambler moves to a poker game.

He is dealt five straight red cards. He remembers what you said and assumes that the next card dealt in the same hand is equally likely to be red or black. Is the gambler right or wrong? Why?

- 4.81 **The “law of averages.”** The baseball player Tony Gwynn got a hit about 34% of the time over his 20-year career. After he failed to hit safely in six straight at-bats, the TV commentator said, “Tony is due for a hit by the law of averages.” Is that right? Why?



- 4.82 **The law of large numbers.** Figure 4.2 (page 239) shows the 36 possible outcomes of rolling two dice and counting the spots on the up-faces. These 36 outcomes are equally likely. You can calculate that the mean for the sum of the two up-faces is $\mu = 7$. This is the population mean μ for the idealized population that contains the results of rolling two dice forever. The law of large numbers says that the average \bar{x} from a finite number of rolls gets closer and closer to 7 as we do more and more rolls.
- Go to the *Mean of a Random Variable* applet. Click “More dice” once to get two dice. Click “Show mean” to see the mean 7 on the graph. Leaving the number of rolls at 1, click “Roll dice” three times. Note the count of spots for each roll (what were they?) and the average for the three rolls. You see that the graph displays at each point the average number of spots for all rolls up to the last one. Now you understand the display.
 - Set the number of rolls to 100 and click “Roll dice.” The applet rolls the two dice 100 times. The graph shows how the average count of spots changes as we make more rolls. That is, the graph shows \bar{x} as we continue to roll the dice. Make a rough sketch of the final graph.
 - Repeat your work from (b). Click “Reset” to start over, then roll two dice 100 times. Make a sketch of the final graph of the mean \bar{x} against the number of rolls. Your two graphs will often look very different. What they have in common is that the average eventually gets close to the population mean $\mu = 7$. The law of large numbers says that this will *always* happen if you keep on rolling the dice.

Sampling distributions

The law of large numbers assures us that if we measure enough subjects, the statistic \bar{x} will eventually get very close to the unknown parameter μ . But our study in Example 4.21 had just 10 subjects. What can we say about \bar{x} from 10 subjects as an estimate of μ ? We ask: “What would happen if we took many samples of 10 subjects from this population?” We learned in Section 3.3 how to answer this question:

- Take a large number of samples of size 10 from the same population.
- Calculate the sample mean \bar{x} for each sample.
- Make a histogram of the values of \bar{x} . This histogram shows how \bar{x} varies in many samples.

In Section 3.3 (page 208) we used computer simulation to take many samples in a different setting. The histogram of values of the statistic

approximates the *sampling distribution* that we would see if we kept on sampling forever. The idea of a sampling distribution is the foundation of statistical inference. One reason for studying probability is that the laws of probability can tell us about sampling distributions without the need to actually choose or simulate a large number of samples.

The mean and the standard deviation of \bar{x}

The first important fact about the sampling distribution of \bar{x} follows by using the rules for means and variances, though we won't do the algebra. Here is the result.¹⁸

MEAN AND STANDARD DEVIATION OF A SAMPLE MEAN

Suppose that \bar{x} is the mean of an SRS of size n drawn from a large population with mean μ and standard deviation σ . Then the **mean** of the sampling distribution of \bar{x} is μ and its **standard deviation** is σ/\sqrt{n} .

Both the mean and the standard deviation of the sampling distribution of \bar{x} have important implications for statistical inference.

- The mean of the statistic \bar{x} is always the same as the mean μ of the population. The sampling distribution of \bar{x} is centered at μ . In repeated sampling, \bar{x} will sometimes fall above the true value of the parameter μ and sometimes below, but there is no systematic tendency to overestimate or underestimate the parameter. This makes the idea of lack of bias in the sense of "no favoritism" more precise. Because the mean of \bar{x} is equal to μ , we say that the statistic \bar{x} is an **unbiased estimator** of the parameter μ .
- An unbiased estimator is "correct on the average" in many samples. How close the estimator falls to the parameter in most samples is determined by the spread of the sampling distribution. If individual observations have standard deviation σ , then sample means \bar{x} from samples of size n have standard deviation σ/\sqrt{n} . **Averages are less variable than individual observations.**

Not only is the standard deviation of the distribution of \bar{x} smaller than the standard deviation of individual observations, but it gets smaller as we take larger samples. **The results of large samples are less variable than the results of small samples.** If n is large, the standard deviation of \bar{x} is small and almost all samples will give values of \bar{x} that lie very close to the true parameter μ . That is, the sample mean from a large sample can be trusted to estimate the population mean accurately. Notice, however, that the standard deviation of the sampling distribution gets smaller only at the rate \sqrt{n} . To cut the standard deviation of \bar{x} in half, we must take four times as many observations, not just twice as many.

APPLY YOUR KNOWLEDGE

- 4.83 **Generating a sampling distribution.** Let's illustrate the idea of a sampling distribution in the case of a very small sample from a very small population. The population is the sizes of 10 medium-sized businesses where size is measured in terms of the number of employees. For convenience, the 10 companies have been labeled with the integers 0 to 9.

Company	0	1	2	3	4	5	6	7	8	9
Size	82	62	80	58	72	73	65	66	74	62

The parameter of interest is the mean size μ in this population. The sample is an SRS of size $n = 4$ drawn from the population. Because the companies are labeled 0 to 9, a single random digit from Table B chooses one company for the sample.

- (a) Find the mean of the 10 sizes in the population. This is the population mean μ .
 - (b) Use Table B to draw an SRS of size 4 from this population. Write the four sizes in your sample and calculate the mean \bar{x} of the sample sizes. This statistic is an estimate of μ .
 - (c) Repeat this process 10 times using different parts of Table B. Make a histogram of the 10 values of \bar{x} . You are constructing the sampling distribution of \bar{x} . Is the center of your histogram close to μ ?
- 4.84 **Measurements on the production line.** Sodium content (in milligrams) is measured for bags of potato chips sampled from a production line. The standard deviation of the sodium content measurements is $\sigma = 10$ mg. The sodium content is measured 3 times and the mean \bar{x} of the 3 measurements is recorded.
- (a) What is the standard deviation of the mean result? (That is, if you kept on making 3 measurements and averaging them, what would be the standard deviation of all the \bar{x} 's?)
 - (b) How many times must we repeat the measurement to reduce the standard deviation of \bar{x} to 5? Explain to someone who knows no statistics the advantage of reporting the average of several measurements rather than the result of a single measurement.
- 4.85 **Measuring blood cholesterol.** A study of the health of teenagers plans to measure the blood cholesterol level of an SRS of youth of ages 13 to 16 years. The researchers will report the mean \bar{x} from their sample as an estimate of the mean cholesterol level μ in this population.
- (a) Explain to someone who knows no statistics what it means to say that \bar{x} is an “unbiased” estimator of μ .
 - (b) The sample result \bar{x} is an unbiased estimator of the population truth μ no matter what size SRS the study chooses. Explain to someone who knows no statistics why a large sample gives more trustworthy results than a small sample.

The central limit theorem

We have described the center and spread of the sampling distribution of a sample mean \bar{x} , but not its shape. The shape of the distribution of \bar{x} depends

on the shape of the population distribution. Here is one important case: if the population distribution is Normal, then so is the distribution of the sample mean.

SAMPLING DISTRIBUTION OF A SAMPLE MEAN

If a population has the $N(\mu, \sigma)$ distribution, then the sample mean \bar{x} of n independent observations has the $N(\mu, \sigma/\sqrt{n})$ distribution.

EXAMPLE 4.23

Estimating odor thresholds

Adults differ in the smallest amount of DMS they can detect in wine. Extensive studies have found that the DMS odor threshold of adults follows roughly a Normal distribution with mean $\mu = 25$ micrograms per liter ($\mu\text{g/l}$) and standard deviation $\sigma = 7 \mu\text{g/l}$. Because the population distribution is Normal, the sampling distribution of \bar{x} is also Normal. Figure 4.16 displays the Normal curve for odor thresholds in the adult population and also the Normal curve for the average threshold in random samples of size 10.

Both distributions have the same mean. But means \bar{x} from samples of 10 adults vary less than do measurements on individual adults. The standard deviation of \bar{x} is

$$\frac{\sigma}{\sqrt{n}} = \frac{7}{\sqrt{10}} = 2.21 \mu\text{g/l}$$

What happens when the population distribution is not Normal? As the sample size increases, the distribution of \bar{x} changes shape: it looks less like

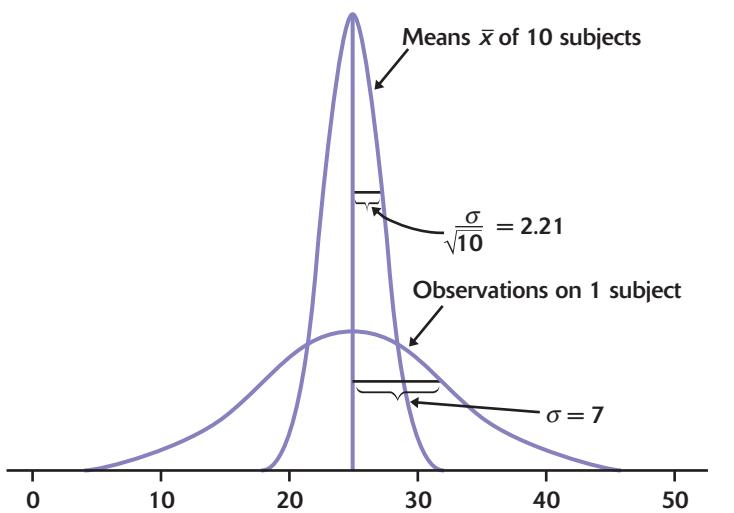


FIGURE 4.16 The distribution of single observations compared with the distribution of the means \bar{x} of 10 observations. Averages are less variable than individual observations.

that of the population and more like a Normal distribution. When the sample is large enough, the distribution of \bar{x} is very close to Normal. This is true no matter what shape the population distribution has, as long as the population has a finite standard deviation σ . This famous fact of probability theory is called the *central limit theorem*. It is much more useful than the fact that the distribution of \bar{x} is exactly Normal if the population is exactly Normal.

CENTRAL LIMIT THEOREM

Draw an SRS of size n from any population with mean μ and finite standard deviation σ . When n is large, the sampling distribution of the sample mean \bar{x} is approximately Normal:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

More general versions of the central limit theorem say that the distribution of a sum or average of many small random quantities is close to Normal. This is true even if the quantities are not independent (as long as they are not too highly correlated) and even if they have different distributions (as long as no one random quantity is so large that it dominates the others). The central limit theorem suggests why the Normal distributions are common models for observed data. Any variable that is a sum of many small influences will have approximately a Normal distribution.

How large a sample size n is needed for \bar{x} to be close to Normal depends on the population distribution. More observations are required if the shape of the population distribution is far from Normal.

EXAMPLE 4.24

The central limit theorem in action

Figure 4.17 shows how the central limit theorem works for a very non-Normal population. Figure 4.17(a) displays the density curve of a single observation, that is, of the population. The distribution is strongly right-skewed, and the most probable outcomes are near 0. The mean μ of this distribution is 1, and its standard deviation σ is also 1. This particular distribution is called an *Exponential distribution*. Exponential distributions are used as models for the lifetime in service of electronic components and for the time required to serve a customer or repair a machine.

Figures 4.17(b), (c), and (d) are the density curves of the sample means of 2, 10, and 25 observations from this population. As n increases, the shape becomes more Normal. The mean remains at $\mu = 1$, and the standard deviation decreases, taking the value $1/\sqrt{n}$. The density curve for the sample mean of 10 observations is still somewhat skewed to the right but already resembles a Normal curve having $\mu = 1$ and $\sigma = 1/\sqrt{10} = 0.32$. The density curve for $n = 25$ is yet more Normal. The contrast between the shapes of the population distribution and of the distribution of the mean of 10 or 25 observations is striking.

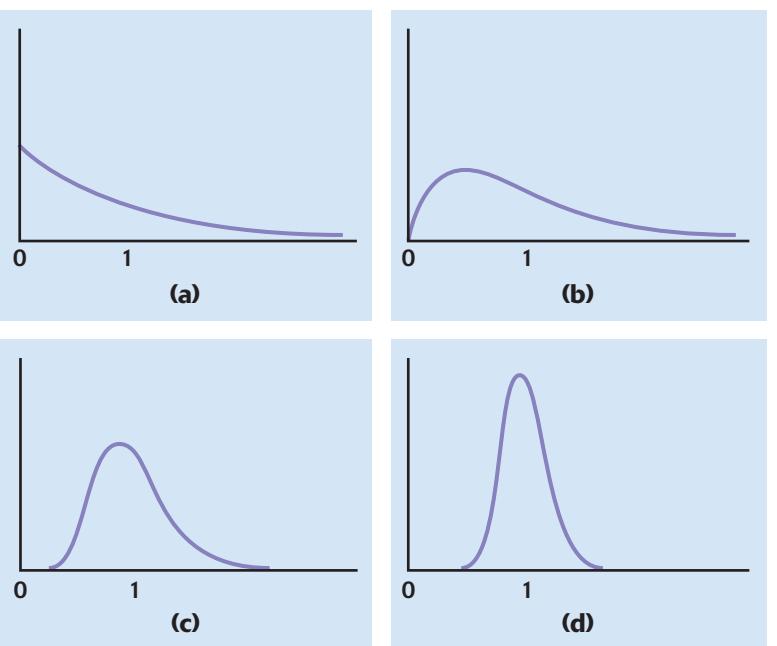


FIGURE 4.17 The central limit theorem in action: the distribution of sample means \bar{x} from a strongly non-Normal population becomes more Normal as the sample size increases. (a) The distribution of 1 observation. (b) The distribution of \bar{x} for 2 observations. (c) The distribution of \bar{x} for 10 observations. (d) The distribution of \bar{x} for 25 observations.

The central limit theorem allows us to use Normal probability calculations to answer questions about sample means from many observations even when the population distribution is not Normal.

EXAMPLE 4.25

Maintaining air conditioners

The time X that a technician requires to perform preventive maintenance on an air-conditioning unit is governed by the Exponential distribution whose density curve appears in Figure 4.17(a). The mean time is $\mu = 1$ hour and the standard deviation is $\sigma = 1$ hour. Your company operates 70 of these units. What is the probability that their average maintenance time exceeds 50 minutes?

The central limit theorem says that the sample mean time \bar{x} (in hours) spent working on 70 units has approximately the Normal distribution with mean equal to the population mean $\mu = 1$ hour and standard deviation

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{70}} = 0.12 \text{ hour}$$

The distribution of \bar{x} is therefore approximately $N(1, 0.12)$. This Normal curve is the solid curve in Figure 4.18.

Because 50 minutes is $50/60$ of an hour, or 0.83 hour, the probability we want is $P(\bar{x} > 0.83)$. A Normal distribution calculation gives this probability as 0.9222. This is the area to the right of 0.83 under the solid Normal curve in Figure 4.18.

Using more mathematics, we could start with the Exponential distribution and find the actual density curve of \bar{x} for 70 observations. This is the dashed curve in

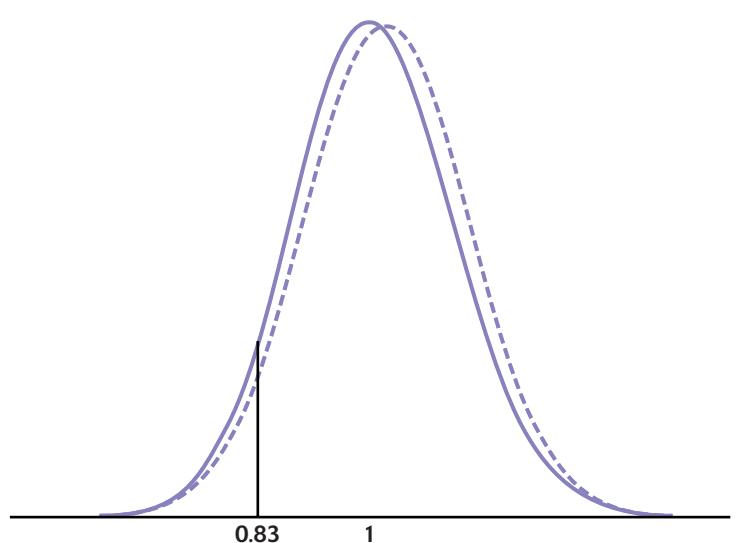


FIGURE 4.18 The exact distribution (dashed) and the Normal approximation from the central limit theorem (solid) for the average time needed to maintain an air conditioner, for Example 4.25.

Figure 4.18. You can see that the solid Normal curve is a good approximation. The exactly correct probability is the area under the dashed density curve. It is 0.9294. The central limit theorem Normal approximation is off by only about 0.007.

APPLY YOUR KNOWLEDGE

- 4.86 **ACT scores.** The scores of students on the ACT college entrance examination in a recent year had the Normal distribution with mean $\mu = 18.6$ and standard deviation $\sigma = 5.9$.
- What is the probability that a single student randomly chosen from all those taking the test scores 21 or higher?
 - Now take an SRS of 50 students who took the test. What are the mean and standard deviation of the sample mean score \bar{x} of these 50 students?
 - What is the probability that the mean score \bar{x} of these students is 21 or higher?
- 4.87 **Flaws in carpets.** The number of flaws per square yard in a type of carpet material varies with mean 1.6 flaws per square yard and standard deviation 1.2 flaws per square yard. The population distribution cannot be Normal, because a count takes only whole-number values. An inspector samples 200 square yards of the material, records the number of flaws found in each square yard, and calculates \bar{x} , the mean number of flaws per square yard inspected. Use the central limit theorem to find the approximate probability that the mean number of flaws exceeds 2 per square yard.
- 4.88 **Returns on stocks.** The distribution of annual returns on common stocks is roughly symmetric, but extreme observations are somewhat more frequent than in Normal distributions. Because the distribution is not strongly non-Normal, the mean return over a number of years is close to Normal. We have

seen that annual returns on common stocks (not adjusted for inflation) have varied with mean about 13% and standard deviation about 17%. Andrew plans to retire in 45 years. He assumes (this is dubious) that the past pattern of variation will continue. What is the probability that the mean annual return on common stocks over the next 45 years will exceed 15%? What is the probability that the mean return will be less than 7%?

SECTION 4.4 SUMMARY

- When we want information about the **population mean** μ for some variable, we often take an SRS and use the **sample mean** \bar{x} to estimate the unknown parameter μ .
- The **law of large numbers** states that the actually observed mean outcome \bar{x} must approach the mean μ of the population as the number of observations increases.
- The **sampling distribution** of \bar{x} describes how the statistic \bar{x} varies in all possible samples of the same size from the same population.
- The **mean** of the sampling distribution is μ , so that \bar{x} is an **unbiased estimator** of μ .
- The **standard deviation** of the sampling distribution of \bar{x} is σ/\sqrt{n} for an SRS of size n if the population has standard deviation σ . That is, averages are less variable than individual observations.
- If the population has a Normal distribution, so does \bar{x} .
- The **central limit theorem** states that for large n the sampling distribution of \bar{x} is approximately Normal for any population with finite standard deviation σ . That is, averages are more Normal than individual observations. We can use the $N(\mu, \sigma/\sqrt{n})$ distribution to calculate approximate probabilities for events involving \bar{x} .

SECTION 4.4 EXERCISES

- 4.89 **Business employees.** There are more than 5 million businesses in the United States. The mean number of employees in these businesses is about 19. A university selects a random sample of 100 businesses in North Dakota and finds that they average about 14 employees. Is each of the bold numbers a parameter or a statistic?
- 4.90 **Personal income.** The Annual Demographic Supplement to the Current Population Survey interviewed more than 128,000 people in March 2001. The Census Bureau reports that the mean income of the Hispanic males interviewed was \$22,771. The mean income for the non-Hispanic white males interviewed was \$41,798. Is each of the bold numbers a parameter or a statistic?
- 4.91 **Roulette.** A roulette wheel has 38 slots, of which 18 are black, 18 are red, and 2 are green. When the wheel is spun, the ball is equally likely to come

to rest in any of the slots. One of the simplest wagers chooses red or black. A bet of \$1 on red returns \$2 if the ball lands in a red slot. Otherwise, the player loses his dollar. When gamblers bet on red or black, the two green slots belong to the house. Because the probability of winning \$2 is $18/38$, the mean payoff from a \$1 bet is twice $18/38$, or 94.7 cents. Explain what the law of large numbers tells us about what will happen if a gambler makes a large number of bets on red.

- 4.92 **Generating a sampling distribution.** Table 1.13 (page 82) gives the population counts of the 58 counties in California. Consider these counties to be the population of interest.
- Make a histogram of the 58 counts. This is the population distribution. It is strongly skewed to the right.
 - Find the mean of the 58 counts. This is the population mean μ . Mark μ on the x axis of your histogram.
 - Label the members of the population 01 to 58 and use Table B to choose an SRS of size $n = 10$. What is \bar{x} for your sample? Mark the value of \bar{x} with a point on the axis of your histogram from (a).
 - Choose four more SRSs of size 10, using different parts of Table B. Find \bar{x} for each sample and mark the values on the axis of your histogram from (a). Would you be surprised if all five \bar{x} 's fell on the same side of μ ? Why?
 - If you chose a large number of SRSs of size 10 from this population and made a histogram of the \bar{x} -values, where would you expect the center of this sampling distribution to lie?
- 4.93 **Dust in coal mines.** A laboratory weighs filters from a coal mine to measure the amount of dust in the mine atmosphere. Repeated measurements of the weight of dust on the same filter vary Normally with standard deviation $\sigma = 0.08$ milligram (mg) because the weighing is not perfectly precise. The dust on a particular filter actually weighs 123 mg. Repeated weighings will then have the Normal distribution with mean 123 mg and standard deviation 0.08 mg.
- The laboratory reports the mean of 3 weighings. What is the distribution of this mean?
 - What is the probability that the laboratory reports a weight of 124 mg or higher for this filter?
- 4.94 **Making auto parts.** An automatic grinding machine in an auto parts plant prepares axles with a target diameter $\mu = 40.125$ millimeters (mm). The machine has some variability, so the standard deviation of the diameters is $\sigma = 0.002$ mm. A sample of 4 axles is inspected each hour for process control purposes, and records are kept of the sample mean diameter. What will be the mean and standard deviation of the numbers recorded?
- 4.95 **Bottling cola.** A bottling company uses a filling machine to fill plastic bottles with cola. The bottles are supposed to contain 300 milliliters (ml). In fact, the contents vary according to a Normal distribution with mean $\mu = 298$ ml and standard deviation $\sigma = 3$ ml.
- What is the probability that an individual bottle contains less than 295 ml?
 - What is the probability that the mean contents of the bottles in a six-pack is less than 295 ml?

- 4.96 **Glucose testing and medical costs.** Shelia's doctor is concerned that she may suffer from gestational diabetes (high blood glucose levels during pregnancy). There is variation both in the actual glucose level and in the blood test that measures the level. A patient is classified as having gestational diabetes if the glucose level is above 140 milligrams per deciliter one hour after a sugary drink is ingested. Shelia's measured glucose level one hour after ingesting the sugary drink varies according to the Normal distribution with $\mu = 125 \text{ mg/dl}$ and $\sigma = 10 \text{ mg/dl}$.
- If a single glucose measurement is made, what is the probability that Shelia is diagnosed as having gestational diabetes?
 - If measurements are made instead on 4 separate days and the mean result is compared with the criterion 140 mg/dl, what is the probability that Shelia is diagnosed as having gestational diabetes?
 - If Shelia is incorrectly diagnosed with gestational diabetes, then she (and her insurance company) will incur unnecessary additional expenses (insulin, needles, doctor visits) for treating the condition during the remainder of her pregnancy. These additional expenses are greater than the cost of repeating the test on three additional days as suggested in (b). Considering the probabilities you calculated in (a) and (b), comment on which testing method seems more appropriate given that Shelia has an acceptable mean glucose level of 125 mg/dl.
- 4.97 **Manufacturing defects.** Newly manufactured automobile radiators may have small leaks. Most have no leaks, but some have 1, 2, or more. The number of leaks in radiators made by one supplier has mean 0.15 and standard deviation 0.4. The distribution of number of leaks cannot be Normal because only whole-number counts are possible. The supplier ships 400 radiators per day to an auto assembly plant. Take \bar{x} to be the mean number of leaks in these 400 radiators. Over several years of daily shipments, what range of values will contain the middle 95% of the many \bar{x} 's?
- 4.98 **Auto accidents.** The number of accidents per week at a hazardous intersection varies with mean 2.2 and standard deviation 1.4. This distribution takes only whole-number values, so it is certainly not Normal.
- Let \bar{x} be the mean number of accidents per week at the intersection during a year (52 weeks). What is the approximate distribution of \bar{x} according to the central limit theorem?
 - What is the approximate probability that \bar{x} is less than 2?
 - What is the approximate probability that there are fewer than 100 accidents at the intersection in a year? (Hint: Restate this event in terms of \bar{x} .)
- 4.99 **Budgeting for expenses.** Weekly postage expenses for your company have a mean of \$312 and a standard deviation of \$58. Your company has allowed for \$400 postage per week in its budget.
- What is the approximate probability that the average weekly postage expense for the past 52 weeks will exceed the budgeted amount of \$400?
 - What information would you need to determine the probability that postage for one particular week will exceed \$400? Why wasn't this information required for (a)?

- 4.100 **Pollutants in auto exhausts.** The level of nitrogen oxides (NOX) in the exhaust of a particular car model varies with mean 0.9 grams per mile (g/mi) and standard deviation 0.15 g/mi. A company has 125 cars of this model in its fleet.
- What is the approximate distribution of the mean NOX emission level \bar{x} for these cars?
 - What is the level L such that the probability that \bar{x} is greater than L is only 0.01? (Hint: This requires a backward Normal calculation. See page 66 of Chapter 1 if you need to review.)
- 4.101 **Glucose testing and medical costs.** In Exercise 4.96, Shelia's measured glucose level one hour after ingesting the sugary drink varies according to the Normal distribution with $\mu = 125$ mg/dl and $\sigma = 10$ mg/dl. Find the level L such that there is probability only 0.05 that the mean glucose level of 4 test results falls above L for Shelia's glucose level distribution. What is the value of L ? (Hint: This requires a backward Normal calculation. See page 66 of Chapter 1 if you need to review.)

STATISTICS IN SUMMARY

This chapter lays the foundations for the study of statistical inference. Statistical inference uses data to draw conclusions about the population or process from which the data come. What is special about inference is that the conclusions include a statement, in the language of probability, about how reliable they are. The statement gives a probability that answers the question “What would happen if I used this method very many times?” The probabilities we need come from the sampling distributions of sample statistics.

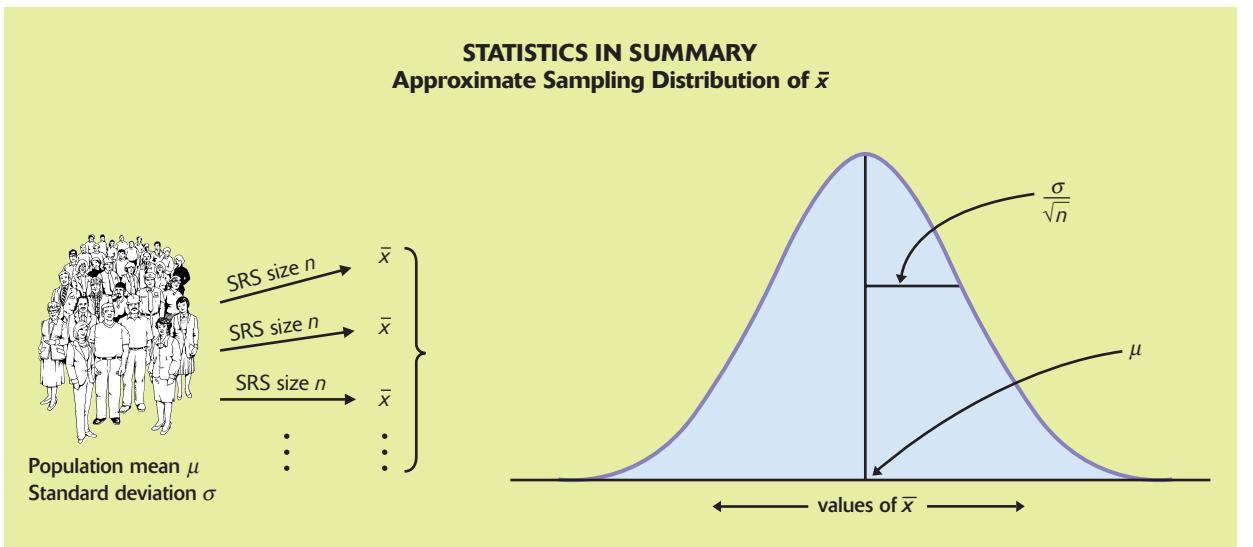
Sampling distributions are the key to understanding statistical inference. The Statistics in Summary figure (on the next page) summarizes the facts about the sampling distribution of \bar{x} in a way that reminds us of the big idea of a sampling distribution. Keep taking random samples of size n from a population with mean μ . Find the sample mean \bar{x} for each sample. Collect all the \bar{x} 's and display their distribution. That's the sampling distribution of \bar{x} . Keep this figure in mind as you go forward.

To think more effectively about sampling distributions, we use the language of probability. Probability, the mathematics that describes randomness, is important in many areas of study. Here, we concentrate on informal probability as the conceptual foundation for statistical inference. Because random samples and randomized comparative experiments use chance, their results vary according to the laws of probability. Here is a review list of the most important things you should be able to do after studying this chapter.

A. PROBABILITY

- Recognize that some phenomena are random. Probability describes the long-run regularity of random phenomena.
- Understand that the probability of an event is the proportion of times the event occurs in very many repetitions of a random phenomenon.

Use the idea of probability as long-run proportion to think about probability.



3. Know the rules that are obeyed by any legitimate assignment of probabilities to events. Use these rules to find the probabilities of events that are formed from other events.
4. Know what properties an assignment of probabilities to a finite number of outcomes must satisfy. Find probabilities of events in a finite sample space by adding the probabilities of their outcomes.
5. Know what a density curve is. Find probabilities of events as areas under a density curve.

B. RANDOM VARIABLES

1. Use the language of random variables to give compact descriptions of events and their probabilities, such as $P(X \geq 2.5)$.
2. Verify that the probability distribution of a discrete random variable is legitimate. Use the distribution to find probabilities of events involving the random variable.
3. Use the density curve of a continuous random variable to find probabilities of events involving the random variable.
4. Calculate the mean and standard deviation of a discrete random variable from its probability distribution.
5. Know and use the rules for means and variances of sums and differences of random variables, both for independent random variables and correlated random variables.

C. THE SAMPLING DISTRIBUTION OF A SAMPLE MEAN

1. Interpret the sampling distribution of a statistic as describing the probabilities of its possible values.

2. Recognize when a problem involves the mean \bar{x} of a sample. Understand that \bar{x} estimates the mean μ of the population from which the sample is drawn.
3. Use the law of large numbers to describe the behavior of \bar{x} as the size of the sample increases.
4. Find the mean and standard deviation of a sample mean \bar{x} from an SRS of size n when the mean μ and standard deviation σ of the population are known.
5. Understand that \bar{x} is an unbiased estimator of μ and that the variability of \bar{x} about its mean μ gets smaller as the sample size increases.
6. Understand that \bar{x} has approximately a Normal distribution when the sample is large (central limit theorem). Use this Normal distribution to calculate probabilities that concern \bar{x} .

CHAPTER 4 REVIEW EXERCISES

- 4.102 **Customer backgrounds.** A company that offers courses to prepare would-be M.B.A. students for the GMAT examination has the following information about its customers: 20% are currently undergraduate students in business; 15% are undergraduate students in other fields of study; 60% are college graduates who are currently employed; and 5% are college graduates who are not employed.
- (a) This is a legitimate assignment of probabilities to customer backgrounds. Why?
 - (b) What percent of customers are currently undergraduates?
- 4.103 **Who gets promoted?** Exactly one of Brown, Chavez, and Williams will be promoted to partner in the law firm that employs them all. Brown thinks that she has probability 0.25 of winning the promotion and that Williams has probability 0.2. What probability does Brown assign to the outcome that Chavez is the one promoted?
- 4.104 **Predicting the ACC champion.** Las Vegas Zeke, when asked to predict the Atlantic Coast Conference basketball champion, follows the modern practice of giving probabilistic predictions. He says, “North Carolina’s probability of winning is twice Duke’s. North Carolina State and Virginia each have probability 0.1 of winning, but Duke’s probability is three times that. Nobody else has a chance.” Has Zeke given a legitimate assignment of probabilities to the eight teams in the conference? Explain your answer.
- 4.105 **Classifying occupations.** Choose an American worker at random and assign his or her occupation to one of the following classes. These classes are used in government employment data.
- A Managerial and professional
B Technical, sales, administrative support
C Service occupations
D Precision production, craft, and repair
E Operators, fabricators, and laborers
F Farming, forestry, and fishing

The table below gives the probabilities that a randomly chosen worker falls into each of 12 sex-by-occupation classes:

Class	A	B	C	D	E	F
Male	0.14	0.11	0.06	0.11	0.12	0.03
Female	0.09	0.20	0.08	0.01	0.04	0.01

- (a) Verify that this is a legitimate assignment of probabilities to these outcomes.
- (b) What is the probability that the worker is female?
- (c) What is the probability that the worker is not engaged in farming, forestry, or fishing?
- (d) Classes D and E include most mechanical and factory jobs. What is the probability that the worker holds a job in one of these classes?
- (e) What is the probability that the worker does not hold a job in Classes D or E?
- 4.106 Rolling dice.** Figure 4.2 (page 239) shows the possible outcomes for rolling two dice. If the dice are carefully made, each of these 36 outcomes has probability $1/36$. The outcome of interest to a gambler is the sum of the spots on the two up-faces. Call this random variable X . Example 4.5 shows that $P(X = 5) = 4/36$.
- (a) Give the complete probability distribution of X in the form of a table of the possible values and their probabilities. Draw a probability histogram to display the distribution.
- (b) One bet available in craps wins if a 7 or an 11 comes up on the next roll of two dice. What is the probability of rolling a 7 or an 11 on the next roll?
- (c) Several bets in craps lose if a 7 is rolled. If any outcome other than 7 occurs, these bets either win or continue to the next roll. What is the probability that anything other than a 7 is rolled?
- 4.107 Nonstandard dice.** You have two balanced, six-sided dice. One is a standard die, with faces having 1, 2, 3, 4, 5, and 6 spots. The other die has three faces with 0 spots and three faces with 6 spots. Find the probability distribution for the total number of spots Y on the up-faces when you roll these two dice.
- 4.108 An IQ test.** The Wechsler Adult Intelligence Scale (WAIS) is a common “IQ test” for adults. The distribution of WAIS scores for persons over 16 years of age is approximately Normal with mean 100 and standard deviation 15.
- (a) What is the probability that a randomly chosen individual has a WAIS score of 105 or higher?
- (b) What are the mean and standard deviation of the average WAIS score \bar{x} for an SRS of 60 people?
- (c) What is the probability that the average WAIS score of an SRS of 60 people is 105 or higher?
- (d) Would your answers to any of (a), (b), or (c) be affected if the distribution of WAIS scores in the adult population were distinctly non-Normal?

- 4.109 **Weights of eggs.** The weight of the eggs produced by a certain breed of hen is Normally distributed with mean 65 grams (g) and standard deviation 5 g. Think of cartons of such eggs as SRSs of size 12 from the population of all eggs. What is the probability that the weight of a carton falls between 750 g and 825 g?
- 4.110 **How many people in a car?** A study of rush-hour traffic in San Francisco counts the number of people in each car entering a freeway at a suburban interchange. Suppose that this count has mean 1.5 and standard deviation 0.75 in the population of all cars that enter at this interchange during rush hours.
- Could the exact distribution of the count be Normal? Why or why not?
 - Traffic engineers estimate that the capacity of the interchange is 700 cars per hour. According to the central limit theorem, what is the approximate distribution of the mean number of persons \bar{x} in 700 randomly selected cars at this interchange?
 - What is the probability that 700 cars will carry more than 1075 people?
(Hint: Restate this event in terms of the mean number of people \bar{x} per car.)

- 4.111 **A grade distribution.** North Carolina State University posts the grade distributions for its courses online.¹⁹ You can find that the distribution of grades in Accounting 210 in the spring 2001 semester was

Grade	A	B	C	D	F
Probability	0.18	0.32	0.34	0.09	0.07

- Verify that this is a legitimate assignment of probabilities to grades.
 - Using the common scale A = 4, B = 3, C = 2, D = 1, F = 0, what is the mean grade in Accounting 210?
- 4.112 **Simulating a mean.** One consequence of the law of large numbers is that once we have a probability distribution for a random variable, we can find its mean by simulating many outcomes and averaging them. The law of large numbers says that if we take enough outcomes, their average value is sure to approach the mean of the distribution.

I have a little bet to offer you. Toss a coin 10 times. If there is no run of three or more straight heads or tails in the 10 outcomes, I'll pay you \$2. If there is a run of three or more, you pay me just \$1. Surely you will want to take advantage of me and play this game?

Simulate enough plays of this game (the outcomes are +\$2 if you win and -\$1 if you lose) to estimate the mean outcome. Is it to your advantage to play?

Insurance. *The business of selling insurance is based on probability and the law of large numbers. Consumers (including businesses) buy insurance because we all face risks that are unlikely but carry high cost—think of a fire destroying your home. So we form a group to share the risk: we all pay a small amount, and the insurance policy pays a large amount to those few of us whose homes burn down. The insurance company sells many policies, so it can rely on the law of large numbers. Exercises 4.113 to 4.118 explore aspects of insurance.*

- 4.113 Fire insurance.** An insurance company looks at the records for millions of homeowners and sees that the mean loss from fire in a year is $\mu = \$250$ per person. (Most of us have no loss, but a few lose their homes. The \$250 is the average loss.) The company plans to sell fire insurance for \$250 plus enough to cover its costs and profit. Explain clearly why it would be stupid to sell only 12 policies. Then explain why selling thousands of such policies is a safe business.
- 4.114 More about fire insurance.** In fact, the insurance company sees that in the entire population of homeowners, the mean loss from fire is $\mu = \$250$ and the standard deviation of the loss is $\sigma = \$300$. The distribution of losses is strongly right-skewed: many policies have \$0 loss, but a few have large losses. If the company sells 10,000 policies, what is the approximate probability that the average loss will be greater than \$260?
- 4.115 Life insurance.** A life insurance company sells a term insurance policy to a 21-year-old male that pays \$100,000 if the insured dies within the next 5 years. The probability that a randomly chosen male will die each year can be found in mortality tables. The company collects a premium of \$250 each year as payment for the insurance. The amount X that the company earns on this policy is \$250 per year, less the \$100,000 that it must pay if the insured dies. Here is the distribution of X . Fill in the missing probability in the table and calculate the mean earnings μ_X .

Age at death	21	22	23	24	25	≥ 26
Earnings X	-\$99,750	-\$99,500	-\$99,250	-\$99,000	-\$98,750	\$1250
Probability	0.00183	0.00186	0.00189	0.00191	0.00193	?

- 4.116 More about life insurance.** It would be quite risky for you to insure the life of a 21-year-old friend under the terms of Exercise 4.115. There is a high probability that your friend would live and you would gain \$1250 in premiums. But if he were to die, you would lose almost \$100,000. Explain carefully why selling insurance is not risky for an insurance company that insures many thousands of 21-year-old men.
- 4.117 The risk of selling insurance.** We have seen that the risk of an investment is often measured by the standard deviation of the return on the investment. The more variable the return is (the larger σ is), the riskier the investment. We can measure the great risk of insuring one person's life in Exercise 4.115 by computing the standard deviation of the income X that the insurer will receive. Find σ_X , using the distribution and mean you found in Exercise 4.115.
- 4.118 The risk of selling insurance, continued.** The risk of insuring one person's life is reduced if we insure many people. Use the result of the previous exercise and the rules for means and variances to answer the following questions.
- (a) Suppose that we insure two 21-year-old males, and that their ages at death are independent. If X and Y are the insurer's income from the two insurance policies, the insurer's average income on the two policies is

$$Z = \frac{X + Y}{2} = 0.5X + 0.5Y$$

Find the mean and standard deviation of Z . You see that the mean income is the same as for a single policy but the standard deviation is less.

- (b) If four 21-year-old men are insured, the insurer's average income is

$$Z = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$$

where X_i is the income from insuring one man. The X_i are independent and each has the same distribution as before. Find the mean and standard deviation of Z . Compare your results with the results of (a). We see that averaging over many insured individuals reduces risk.

- 4.119 Finite sample spaces.** Choose one employee of your company at random and let X be the number of years that person has been employed by the company (rounded to the nearest year).

- (a) Give a reasonable sample space S for the possible values of X .
- (b) Is X a discrete random variable or a continuous random variable? Why?
- (c) How many possible values of X did you include in your definition of S ?

- 4.120 (Optional) Infinite sample spaces, Part I.** Buy one share of Apple Computer (AAPL) stock and let Y be the number of days until the closing market value of your share is double what you initially paid for the share.

- (a) Give the sample space S for the possible values of Y .
- (b) Is Y a discrete random variable or a continuous random variable? Why?
- (c) How many possible values of Y did you include in your definition of S ? (This exercise illustrates that “discrete” and “finite” are not the same thing, though we have made use only of finite discrete sample spaces.)²⁰

- 4.121 (Optional) Infinite sample spaces, Part II.** Randomly choose one 30-milliliter (ml) ink cartridge from the lot of cartridges produced at your plant in the last hour. Let W be the actual amount of ink in the cartridge. The cartridges are supposed to contain 30 ml of ink but are designed to hold up to a maximum of 35 ml of ink.

- (a) Give a reasonable sample space S for the possible values of W .
- (b) Is W a discrete random variable or a continuous random variable? Why?
- (c) How many possible values of W did you include in your definition of S ?



CHAPTER 4 CASE STUDY EXERCISES

CASE STUDY 4.1: Benford's Law. Case 4.1 (page 243) concerns Benford's Law, which gives the distribution of first digits in many sets of data from business and science. Locate at least two long tables whose entries could plausibly begin with any digit 1 through 9. You may choose data tables, such as populations of many cities or the number of shares traded on the New York Stock Exchange on many days, or mathematical tables such as logarithms or square roots. We hope it's clear that you can't use the table of random digits. Let's require that your examples each contain at least 300 numbers. Tally the first digits of all entries in each table. Report the distributions (in percents) and compare them with each other, with Benford's Law, and with the “equally likely” distribution.

CASE STUDY 4.2: State lotteries. Most American states and Canadian provinces operate lotteries. State lotteries combine probability with economics and politics. States sometimes add opportunities for gambling in hard times to increase their take from operating the games or from taxing games run by others. Investigate the presence of legal gambling in your state. Include probability aspects, such as the probability of winning and the expected payoff on different games. Also look at economic aspects of gambling: how much does the state earn? What percent of the state budget is this? Has the state's take from gambling changed in recent years? Probability and economics are related, because a state lottery keeps only what it doesn't pay out in winnings to gamblers. What percent of the money bet in the lottery does the state pay out in prizes? How does this compare with casino games?

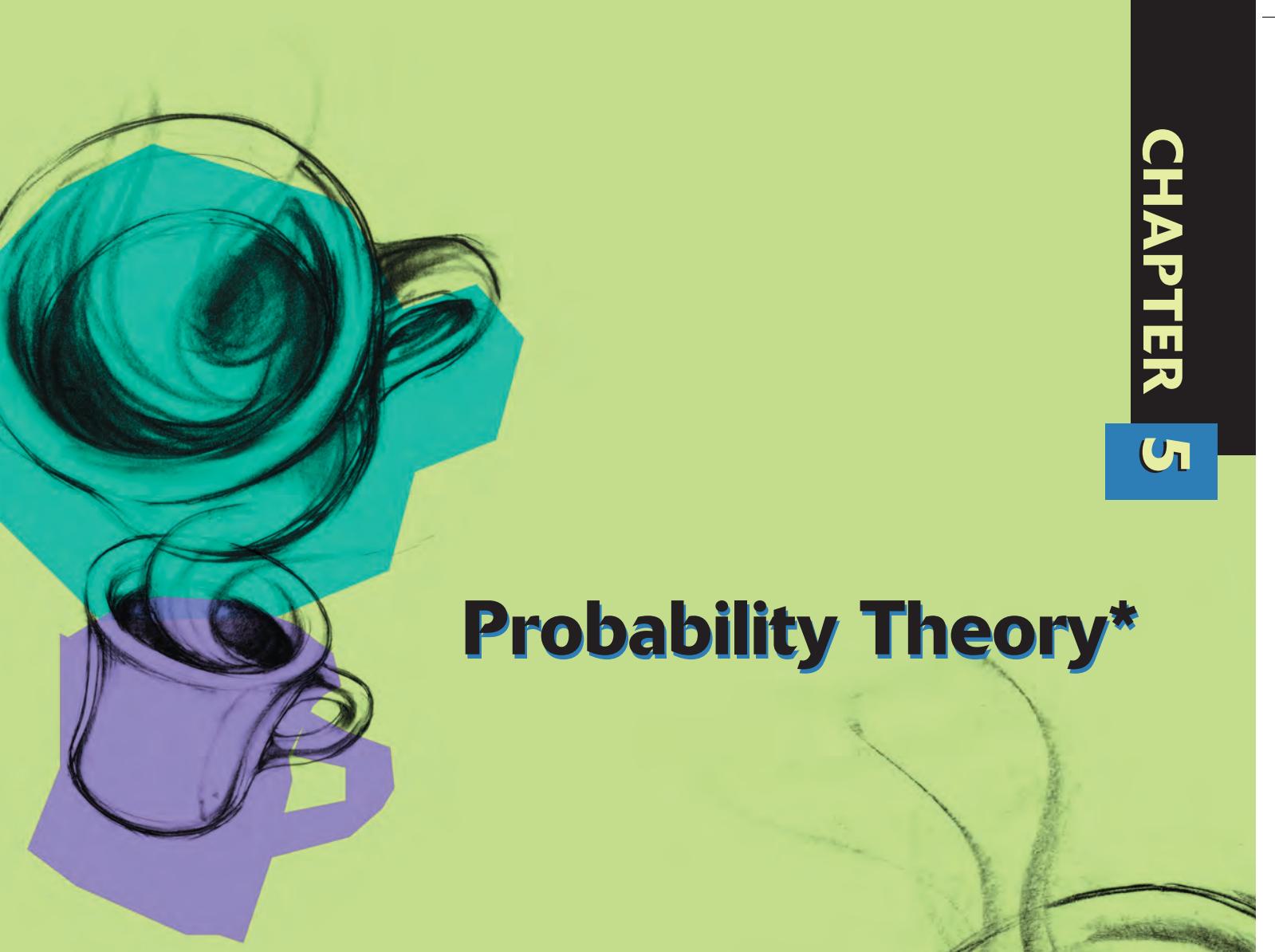
Prelude

The case of the Pentium bug

Intel's Pentium processors power a majority of the world's computers. In 1994, a bug was discovered in the Pentium chip. The bug, which came to be known as the Pentium FDIV bug, caused some floating-point division operations to give incorrect values. Intel claimed that the probability of an error was only 1 in 9 billion. At that rate, said the chip firm, a spreadsheet user who performs 1000 floating-point divisions per day will encounter an incorrect division only about once in 25,000 years. The probability of one or more errors due to the FDIV bug in 365 working days is only about 0.00004.

An IBM research group disagreed. Based on the results of probability simulations, IBM concluded that a more reasonable estimate of the probability of a floating-point division error was 1 in 100 million, 90 times the estimate given by Intel. What is more, simulations of typical financial calculations done by spreadsheet users found that an average user would perform nearly 4.2 million divisions per day. The probability of one or more errors due to the FDIV bug in 365 days, said IBM, is not 0.00004 but 0.999998.

At first, Intel claimed that the bug would cause errors so rarely that it would not replace the faulty chips that had already been sold—as many as 2 million chips. However, faced with results like IBM's and growing customer concern, Intel agreed to replace the chips for anyone who wanted a replacement. The Pentium FDIV bug incident is reported to have cost Intel approximately 475 million dollars.¹



CHAPTER
5

Probability Theory*

- 5.1 General Probability Rules**
- 5.2 The Binomial Distributions**
- 5.3 The Poisson Distributions**
- 5.4 Conditional Probability**

*This more advanced chapter gives more detail about probability. It is not needed to read the rest of the book.



Introduction

The mathematics of probability can provide models to describe the flow of traffic through a highway system, a telephone interchange, or a computer processor; the product preferences of consumers; the spread of epidemics or computer viruses; and the rate of return on risky investments. Although we are interested in probability because of its usefulness in statistics, the mathematics of chance is important in many fields of study. This chapter presents a bit more of the theory of probability.

5.1 General Probability Rules

Our study of probability in Chapter 4 concentrated on sampling distributions. Now we return to the general laws of probability. With more probability at our command, we can model more complex random phenomena. We have already met and used four rules.

RULES OF PROBABILITY

Rule 1. $0 \leq P(A) \leq 1$ for any event A

Rule 2. $P(S) = 1$

Rule 3. Complement rule: For any event A ,

$$P(A^c) = 1 - P(A)$$

Rule 4. Addition rule: If A and B are disjoint events, then

$$P(A \text{ or } B) = P(A) + P(B)$$

The complement rule takes its name from the fact that the set of all outcomes that are *not* in an event A is often called the **complement** of A . As a convenient short notation, we will write the complement of A as A^c . For example, if A is the event that a randomly chosen corporate CEO is female, then A^c is the event that the CEO is male.

Independence and the multiplication rule

Rule 4, the addition rule for disjoint events, describes the probability that *one or the other* of two events A and B occurs when A and B cannot occur together. Now we will describe the probability that *both* events A and B occur, again only in a special situation.

You may find it helpful to draw a picture to display relations among several events. A picture like Figure 5.1 that shows the sample space S as a rectangular area and events as areas within S is called a **Venn diagram**. The events A and B in Figure 5.1 are disjoint because they do not overlap.

Venn diagram

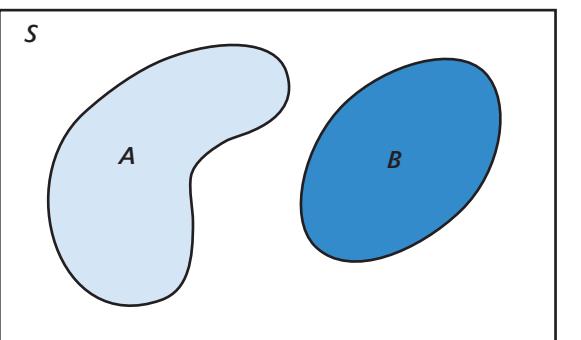


FIGURE 5.1 Venn diagram showing disjoint events A and B .

The Venn diagram in Figure 5.2 illustrates two events that are not disjoint. The event $\{A \text{ and } B\}$ appears as the overlapping area that is common to both A and B .

Suppose that you toss a balanced coin twice. You are counting heads, so two events of interest are

$$A = \text{first toss is a head}$$

$$B = \text{second toss is a head}$$

The events A and B are not disjoint. They occur together whenever both tosses give heads. We want to find the probability of the event $\{A \text{ and } B\}$ that *both* tosses are heads.

The coin tossing of Buffon, Pearson, and Kerrich described at the beginning of Chapter 4 makes us willing to assign probability $1/2$ to a head when we toss a coin. So

$$P(A) = 0.5$$

$$P(B) = 0.5$$

What is $P(A \text{ and } B)$? Our common sense says that it is $1/4$. The first coin will give a head half the time and then the second will give a head on

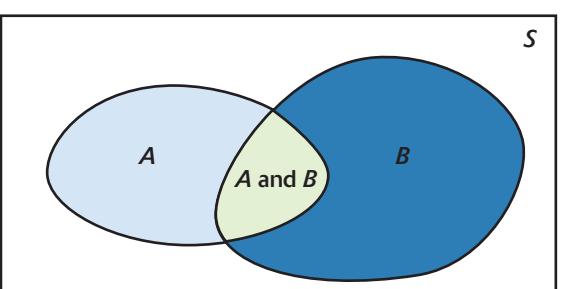


FIGURE 5.2 Venn diagram showing events A and B that are not disjoint. The event $\{A \text{ and } B\}$ consists of outcomes common to A and B .

independence

half of those trials, so both coins will give heads on $1/2 \times 1/2 = 1/4$ of all trials in the long run. This reasoning assumes that the second coin still has probability 1/2 of a head after the first has given a head. This is true—we can verify it by tossing two coins many times and observing the proportion of heads on the second toss after the first toss has produced a head. We say that the events “head on the first toss” and “head on the second toss” are **independent**. Independence means that the outcome of the first toss cannot influence the outcome of the second toss.

EXAMPLE 5.1**Independent or not?**

Because a coin has no memory and most coin tossers cannot influence the fall of the coin, it is safe to assume that successive coin tosses are independent. For a balanced coin this means that after we see the outcome of the first toss, we still assign probability 1/2 to heads on the second toss.

On the other hand, the colors of successive cards dealt from the same deck are not independent. A standard 52-card deck contains 26 red and 26 black cards. For the first card dealt from a shuffled deck, the probability of a red card is $26/52 = 0.50$ (equally likely outcomes). Once we see that the first card is red, we know that there are only 25 reds among the remaining 51 cards. The probability that the second card is red is therefore only $25/51 = 0.49$. Knowing the outcome of the first deal changes the probabilities for the second.

If an employer does two tests for illegal drugs on the same blood sample from a job applicant, it is reasonable to assume that the two results are independent because the first result does not influence the instrument that makes the second reading. But if the applicant takes an IQ test or other mental test twice in succession, the two test scores are not independent. The learning that occurs on the first attempt influences the second attempt.

**MULTIPLICATION RULE FOR INDEPENDENT EVENTS**

Two events A and B are **independent** if knowing that one occurs does not change the probability that the other occurs. If A and B are independent,

$$P(A \text{ and } B) = P(A)P(B)$$

EXAMPLE 5.2**Mendel's peas**

Gregor Mendel used garden peas in some of the experiments that revealed that inheritance operates randomly. The seed color of Mendel's peas can be either green or yellow. Two parent plants are “crossed” (one pollinates the other) to produce seeds. Each parent plant carries two genes for seed color, and each of these genes has probability 1/2 of being passed to a seed. The two genes that the seed receives, one from each parent, determine its color. The parents contribute their genes independently of each other.

Suppose that both parents carry the G (green) and the Y (yellow) genes. The seed will be green if both parents contribute a G gene; otherwise it will be yellow. If M is the event that the male contributes a G gene and F is the event that the female contributes a G gene, then the probability of a green seed is

$$\begin{aligned}P(M \text{ and } F) &= P(M)P(F) \\&= (0.5)(0.5) = 0.25\end{aligned}$$

In the long run, 1/4 of all seeds produced by crossing these plants will be green.



The multiplication rule $P(A \text{ and } B) = P(A)P(B)$ holds if A and B are *independent* but not otherwise. The addition rule $P(A \text{ or } B) = P(A) + P(B)$ holds if A and B are *disjoint* but not otherwise. Resist the temptation to use these simple rules when the circumstances that justify them are not present. You must also be certain not to confuse disjointness and independence. If A and B are disjoint, then the fact that A occurs tells us that B cannot occur—look again at Figure 5.1. So disjoint events are not independent. Unlike disjointness, we cannot picture independence in a Venn diagram, because it involves the probabilities of the events rather than just the outcomes that make up the events.

APPLY YOUR KNOWLEDGE

- 5.1 **High school rank.** Select a first-year college student at random and ask what his or her academic rank was in high school. Here are the probabilities, based on proportions from a large sample survey of first-year students:

Rank	Top 20%	Second 20%	Third 20%	Fourth 20%	Lowest 20%
Probability	0.41	0.23	0.29	0.06	0.01

- (a) Choose two first-year college students at random. Why is it reasonable to assume that their high school ranks are independent?
(b) What is the probability that both were in the top 20% of their high school classes?
(c) What is the probability that the first was in the top 20% and the second was in the lowest 20%?
- 5.2 **College-educated laborers?** Government data show that 27% of employed people have at least 4 years of college and that 14% of employed people work as laborers or operators of machines or vehicles. Can you conclude that because $(0.27)(0.14) = 0.038$ about 3.8% of employed people are college-educated laborers or operators? Explain your answer.

Applying the multiplication rule

If two events A and B are independent, the event that A does not occur is also independent of B , and so on. Suppose, for example, that 75% of all registered voters in a suburban district are Republicans. If an opinion poll interviews two voters chosen independently, the probability that the

first is a Republican and the second is not a Republican is $(0.75)(0.25) = 0.1875$. The multiplication rule also extends to collections of more than two events, provided that all are independent. Independence of events A , B , and C means that no information about any one or any two can change the probability of the remaining events. Independence is often assumed in setting up a probability model when the events we are describing seem to have no connection. We can then use the multiplication rule freely, as in this example.

EXAMPLE 5.3

Undersea cables

The first successful transatlantic telegraph cable was laid in 1866. The first telephone cable across the Atlantic did not appear until 1956—the barrier was designing “repeaters,” amplifiers needed to boost the signal, that could operate for years on the sea bottom. This first cable had 52 repeaters. The last copper cable, laid in 1983 and retired in 1994, had 662 repeaters. The first fiber-optic cable was laid in 1988 and has 109 repeaters. There are now more than 400,000 miles of undersea cable, with more being laid every year to handle the flood of Internet traffic.

Repeaters in undersea cables must be very reliable. To see why, suppose that each repeater has probability 0.999 of functioning without failure for 25 years. Repeaters fail independently of each other. (This assumption means that there are no “common causes” such as earthquakes that would affect several repeaters at once.) Denote by A_i the event that the i th repeater operates successfully for 25 years.

The probability that 2 repeaters both last 25 years is

$$\begin{aligned} P(A_1 \text{ and } A_2) &= P(A_1)P(A_2) \\ &= 0.999 \times 0.999 = 0.998 \end{aligned}$$

For a cable with 10 repeaters the probability of no failures in 25 years is

$$\begin{aligned} P(A_1 \text{ and } A_2 \text{ and } \dots \text{ and } A_{10}) &= P(A_1)P(A_2) \cdots P(A_{10}) \\ &= 0.999 \times 0.999 \times \cdots \times 0.999 \\ &= 0.999^{10} = 0.990 \end{aligned}$$

Cables with 2 or 10 repeaters would be quite reliable. Unfortunately, the last copper transatlantic cable had 662 repeaters. The probability that all 662 work for 25 years is

$$P(A_1 \text{ and } A_2 \text{ and } \dots \text{ and } A_{662}) = 0.999^{662} = 0.516$$

This cable will fail to reach its 25-year design life about half the time if each repeater is 99.9% reliable over that period. The multiplication rule for probabilities shows that repeaters must be much more than 99.9% reliable.

By combining the rules we have learned, we can compute probabilities for rather complex events. Here is an example.

EXAMPLE 5.4**False positives in HIV testing**

Screening large numbers of blood samples for HIV, the virus that causes AIDS, uses an enzyme immunoassay (EIA) test that detects antibodies to the virus. Samples that test positive are retested using a more accurate “Western blot” test. Applied to people who have no HIV antibodies, EIA has probability about 0.006 of producing a false positive. If the 140 employees of a medical clinic are tested and all 140 are free of HIV antibodies, what is the probability that at least 1 false positive will occur?

It is reasonable to assume as part of the probability model that the test results for different individuals are independent. The probability that the test is positive for a single person is 0.006, so the probability of a negative result is $1 - 0.006 = 0.994$ by the complement rule. The probability of at least 1 false positive among the 140 people tested is therefore

$$\begin{aligned}P(\text{at least one positive}) &= 1 - P(\text{no positives}) \\&= 1 - P(140 \text{ negatives}) \\&= 1 - 0.994^{140} \\&= 1 - 0.431 = 0.569\end{aligned}$$

The probability is greater than 1/2 that at least 1 of the 140 people will test positive for HIV, even though no one has the virus.

APPLY YOUR KNOWLEDGE

- 5.3 **Telemarketing.** Telephone marketers and opinion polls use random-digit-dialing equipment to call residential telephone numbers at random. The telephone polling firm Zogby International reports that the probability that a call reaches a live person is 0.2.² Calls are independent.
 - (a) A telemarketer places 5 calls. What is the probability that none of them reaches a person?
 - (b) When calls are made to New York City, the probability of reaching a person is only 0.08. What is the probability that none of 5 calls made to New York City reaches a person?
- 5.4 **Detecting drug use.** An employee suspected of having used an illegal drug is given two tests that operate independently of each other. Test A has probability 0.9 of being positive if the illegal drug has been used. Test B has probability 0.8 of being positive if the illegal drug has been used. What is the probability that *neither* test is positive if the illegal drug has been used?
- 5.5 **Bright lights?** A string of holiday lights contains 20 lights. The lights are wired in series, so that if any light fails the whole string will go dark. Each light has probability 0.02 of failing during a 3-year period. The lights fail independently of each other. What is the probability that the string of lights will remain bright for 3 years?

The general addition rule

We know that if A and B are disjoint events, then $P(A \text{ or } B) = P(A) + P(B)$. This addition rule extends to more than two events that are disjoint in the

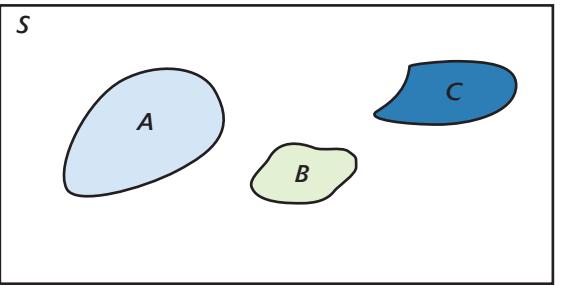


FIGURE 5.3 The addition rule for disjoint events:
 $P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$ when events A ,
 B , and C are disjoint.

sense that no two have any outcomes in common. The Venn diagram in Figure 5.3 shows three disjoint events A , B , and C . The probability that one of these events occurs is $P(A) + P(B) + P(C)$.

If events A and B are *not* disjoint, they can occur simultaneously. The probability that one or the other occurs is then *less* than the sum of their probabilities. As Figure 5.4 suggests, the outcomes common to both are counted twice when we add probabilities, so we must subtract this probability once. Here is the addition rule for any two events, disjoint or not.

GENERAL ADDITION RULE FOR ANY TWO EVENTS

For any two events A and B ,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

If A and B are disjoint, the event $\{A \text{ and } B\}$ that both occur contains no outcomes and therefore has probability 0. So the general addition rule includes Rule 4, the addition rule for disjoint events.

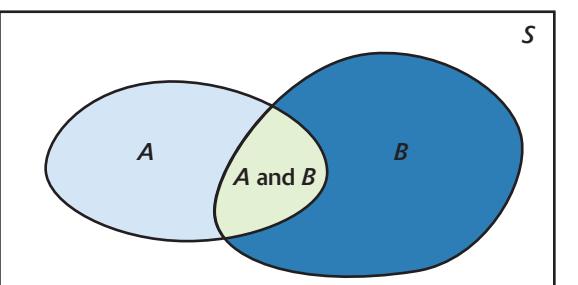


FIGURE 5.4 The general addition rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ for any events A and B .

EXAMPLE 5.5**Making partner**

Deborah and Matthew are anxiously awaiting word on whether they have been made partners of their law firm. Deborah guesses that her probability of making partner is 0.7 and that Matthew's is 0.5. (These are personal probabilities reflecting Deborah's assessment of chance.) This assignment of probabilities does not give us enough information to compute the probability that at least one of the two is promoted. In particular, adding the individual probabilities of promotion gives the impossible result 1.2. If Deborah also guesses that the probability that *both* she and Matthew are made partners is 0.3, then by the general addition rule

$$P(\text{at least one is promoted}) = 0.7 + 0.5 - 0.3 = 0.9$$

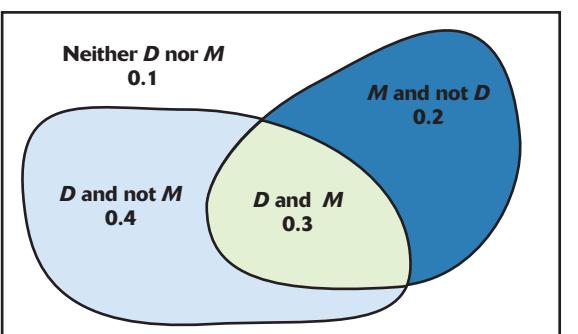
The probability that *neither* is promoted is then 0.1 by the complement rule.



Venn diagrams are a great help in finding probabilities because you can just think of adding and subtracting areas. Figure 5.5 shows some events and their probabilities for Example 5.5. What is the probability that Deborah is promoted and Matthew is not? The Venn diagram shows that this is the probability that Deborah is promoted minus the probability that both are promoted, $0.7 - 0.3 = 0.4$. Similarly, the probability that Matthew is promoted and Deborah is not is $0.5 - 0.3 = 0.2$. The four probabilities that appear in the figure add to 1 because they refer to four disjoint events that make up the entire sample space.

APPLY YOUR KNOWLEDGE

- 5.6 Prosperity and education. Call a household prosperous if its income exceeds \$100,000. Call the household educated if the householder completed college. Select an American household at random, and let A be the event that the selected household is prosperous and B the event that it is educated. According to the Current Population Survey, $P(A) = 0.134$, $P(B) = 0.254$,



D = Deborah is made partner
 M = Matthew is made partner

FIGURE 5.5 Venn diagram and probabilities for Example 5.5.

and the probability that a household is both prosperous and educated is $P(A \text{ and } B) = 0.080$.

- Draw a Venn diagram that shows the relation between the events A and B . What is the probability $P(A \text{ or } B)$ that the household selected is either prosperous or educated?
- In your diagram, shade the event that the household is educated but not prosperous. What is the probability of this event?

- 5.7 **Caffeine in the diet.** Common sources of caffeine are coffee, tea, and cola drinks. Suppose that

55% of adults drink coffee

25% of adults drink tea

45% of adults drink cola

and also that

15% drink both coffee and tea

5% drink all three beverages

25% drink both coffee and cola

5% drink only tea

Draw a Venn diagram marked with this information. Use it along with the addition rules to answer the following questions.

- What percent of adults drink only cola?
- What percent drink none of these beverages?

SECTION 5.1 SUMMARY

- Events A and B are **disjoint** if they have no outcomes in common. Events A and B are **independent** if knowing that one event occurs does not change the probability we would assign to the other event.
- Any assignment of probability obeys these more general rules in addition to those stated in Chapter 4:

Addition rule: If events A, B, C, \dots are all disjoint in pairs, then

$$P(\text{at least one of these events occurs}) = P(A) + P(B) + P(C) + \dots$$

Multiplication rule: If events A and B are independent, then

$$P(A \text{ and } B) = P(A)P(B)$$

General addition rule: For any two events A and B ,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

SECTION 5.1 EXERCISES

- 5.8 **Hiring strategy.** A chief executive officer (CEO) has resources to hire one vice-president or three managers. He believes that he has probability 0.6 of successfully recruiting the vice-president candidate and probability 0.8

of successfully recruiting each of the manager candidates. The three candidates for manager will make their decisions independently of each other. The CEO must successfully recruit either the vice-president or all three managers to consider his hiring strategy a success. Which strategy should he choose?

- 5.9 **Playing the lottery.** An instant lottery game gives you probability 0.02 of winning on any one play. Plays are independent of each other. If you play 5 times, what is the probability that you win at least once?
- 5.10 **Nonconforming chips.** Automobiles use semiconductor chips for engine and emission control, repair diagnosis, and other purposes. An auto manufacturer buys chips from a supplier. The supplier sends a shipment of which 5% fail to conform to performance specifications. Each chip chosen from this shipment has probability 0.05 of being nonconforming, and each automobile uses 12 chips selected independently. What is the probability that all 12 chips in a car will work properly?
- 5.11 **A random walk on Wall Street?** The “random walk” theory of securities prices holds that price movements in disjoint time periods are independent of each other. Suppose that we record only whether the price is up or down each year, and that the probability that our portfolio rises in price in any one year is 0.65. (This probability is approximately correct for a portfolio containing equal dollar amounts of all common stocks listed on the New York Stock Exchange.)
 - (a) What is the probability that our portfolio goes up for three consecutive years?
 - (b) If you know that the portfolio has risen in price 2 years in a row, what probability do you assign to the event that it will go down next year?
 - (c) What is the probability that the portfolio’s value moves in the same direction in both of the next 2 years?
- 5.12 **Getting into an MBA program.** Ramon has applied to MBA programs at both Harvard and Stanford. He thinks the probability that Harvard will admit him is 0.4, the probability that Stanford will admit him is 0.5, and the probability that both will admit him is 0.2.
 - (a) Make a Venn diagram with the probabilities given marked.
 - (b) What is the probability that neither university admits Ramon?
 - (c) What is the probability that he gets into Stanford but not Harvard?
- 5.13 **Will we get the jobs?** Consolidated Builders has bid on two large construction projects. The company president believes that the probability of winning the first contract (event A) is 0.6, that the probability of winning the second (event B) is 0.5, and that the probability of winning both jobs (event $\{A \text{ and } B\}$) is 0.3. What is the probability of the event $\{A \text{ or } B\}$ that Consolidated will win at least one of the jobs?
- 5.14 **Tastes in music.** Musical styles other than rock and pop are becoming more popular. A survey of college students finds that 40% like country music, 30% like gospel music, and 10% like both.
 - (a) Make a Venn diagram with these results.
 - (b) What percent of college students like country but not gospel?
 - (c) What percent like neither country nor gospel?

- 5.15 **Independent?** In the setting of Exercise 5.13, are events A and B independent? Do a calculation that proves your answer.

Blood types. All human blood can be “ABO-typed” as one of O, A, B, or AB, but the distribution of the types varies a bit among groups of people. Here is the distribution of blood types for a randomly chosen person in the United States:

Blood type	O	A	B	AB
U.S. probability	0.45	0.40	0.11	0.04

Choose a married couple at random. It is reasonable to assume that the blood types of husband and wife are independent and follow this distribution. Exercises 5.16 to 5.19 concern this setting.

- 5.16 **Is transfusion safe?** Someone with type B blood can safely receive transfusions only from persons with type B or type O blood. What is the probability that the husband of a woman with type B blood is an acceptable blood donor for her?
- 5.17 **Same type?** What is the probability that a wife and husband share the same blood type?
- 5.18 **Blood types, continued.** What is the probability that the wife has type A blood and the husband has type B? What is the probability that one of the couple has type A blood and the other has type B?
- 5.19 **Don't forget Rh.** Human blood is typed as O, A, B, or AB and also as Rh-positive or Rh-negative. ABO type and Rh-factor type are independent because they are governed by different genes. In the American population, 84% of people are Rh-positive. Give the probability distribution of blood type (ABO and Rh together) for a randomly chosen person.
- 5.20 **Age effects in medical care.** The type of medical care a patient receives may vary with the age of the patient. A large study of women who had a breast lump investigated whether or not each woman received a mammogram and a biopsy when the lump was discovered. Here are some probabilities estimated by the study. The entries in the table are the probabilities that *both* of two events occur; for example, 0.321 is the probability that a patient is under 65 years of age *and* the tests were done. The four probabilities in the table have sum 1 because the table lists all possible outcomes.

Age	Tests done?	
	Yes	No
Under 65	0.321	0.124
65 or over	0.365	0.190

- (a) What is the probability that a patient in this study is under 65? That a patient is 65 or over?
- (b) What is the probability that the tests were done for a patient? That they were not done?

- (c) Are the events $A = \{\text{the patient was 65 or older}\}$ and $B = \{\text{the tests were done}\}$ independent? Were the tests omitted on older patients more or less frequently than would be the case if testing were independent of age?
- 5.21 Playing the odds? A writer on casino games says that the odds against throwing an 11 in the dice game craps are 17 to 1. He then says that the odds against three 11s in a row are $17 \times 17 \times 17$ to 1, or 4913 to 1.³
- What is the probability that the sum of the up-faces is 11 when you throw two balanced dice? (See Figure 4.2 on page 239.) What is the probability of three 11s in three independent throws?
 - If an event A has probability P , the odds against A are

$$\text{odds against } A = \frac{1 - P}{P}$$

Gamblers often speak of odds rather than probabilities. The odds against an event that has probability $1/3$ are 2 to 1, for example. Find the odds against throwing an 11 and the odds against throwing three straight 11s. Which of the writer's statements are correct?

5.2 The Binomial Distributions

A company's human resources manager asks 100 employees if job stress is affecting their personal lives. How many will say "Yes"? A new treatment for pancreatic cancer is tried on 25 patients. How many will survive for 5 years? A store sells 10 computers with 1-year warranties. How many will not need repair within 1 year? In all these situations, we want a probability model for a *count* of successful outcomes.

The Binomial setting

The distribution of a count depends on how the data are produced. Here is a common situation.

THE BINOMIAL SETTING

- There are a fixed number n of observations.
- The n observations are all **independent**. That is, knowing the result of one observation tells you nothing about the other observations.
- Each observation falls into one of just two categories, which for convenience we call "success" and "failure."
- The probability of a success, call it p , is the same for each observation.

Think of tossing a coin n times as an example of the Binomial setting. Each toss gives either heads or tails. Knowing the outcome of one toss doesn't tell us anything about other tosses, so the n tosses are independent.

If we call heads a success, then p is the probability of a head and remains the same as long as we toss the same coin. The number of heads we count is a random variable X . The distribution of X is called a *Binomial distribution*.

BINOMIAL DISTRIBUTION

The distribution of the count X of successes in the Binomial setting is the **Binomial distribution** with parameters n and p . The parameter n is the number of observations, and p is the probability of a success on any one observation. The possible values of X are the whole numbers from 0 to n .

The Binomial distributions are an important class of probability distributions. Pay attention to the Binomial setting, because not all counts have Binomial distributions.

EXAMPLE 5.6

Determining consumer preferences

Market research to determine the product preferences of consumers is an increasingly important area in the intersection of business and statistics. With some companies competing in markets with little product discrimination, determining what features consumers most prefer is critical to the success of a product. The probability of a “typical” consumer purchasing a product with a particular combination of features is the probability of interest in market research.

Suppose that your product is actually preferred over competitors’ products by 25% of all consumers. If X is the count of the number of consumers who prefer your product in a group of 5 consumers, then X has a Binomial distribution with $n = 5$ and $p = 0.25$ provided the 5 consumers make choices independently. Some business schools and companies are doing research on innovative ways to collect independent consumer data for use in statistical analyses.⁴



EXAMPLE 5.7

Dealing cards

Deal 10 cards from a shuffled deck and count the number X of red cards. There are 10 observations, and each gives either a red or a black card. A “success” is a red card. But the observations are *not* independent. If the first card is black, the second is more likely to be red because there are more red cards than black cards left in the deck. The count X does *not* have a Binomial distribution.



INSPECTING A SUPPLIER'S PRODUCTS

A manufacturing firm purchases components for its products from suppliers. Good practice calls for suppliers to manage their production processes to ensure good quality. You can find some discussion of statistical methods

for managing and improving quality in Chapter 12. There have, however, been quality lapses in the switches supplied by a regular vendor. While working with the supplier to improve its processes, the manufacturing firm temporarily institutes an *acceptance sampling* plan to assess the quality of shipments of switches. If a random sample from a shipment contains too many switches that don't conform to specifications, the firm will not accept the shipment.

An engineer at the firm chooses an SRS of 10 switches from a shipment of 10,000 switches. Suppose that (unknown to the engineer) 10% of the switches in the shipment are nonconforming. The engineer counts the number X of nonconforming switches in the sample.

This is not quite a Binomial setting. Just as removing 1 card in Example 5.7 changed the makeup of the deck, removing 1 switch changes the proportion of nonconforming switches remaining in the shipment. If there are initially 1000 nonconforming switches, the proportion remaining is $1000/9999 = 0.10001$ if the first switch drawn is OK and $999/9999 = 0.09991$ if the first switch fails inspection. That is, the state of the second switch chosen is not independent of the first. But removing 1 switch from a shipment of 10,000 changes the makeup of the remaining 9999 switches very little. In practice, the distribution of X is very close to the Binomial distribution with $n = 10$ and $p = 0.1$.



Case 5.1 shows how we can use the Binomial distributions in the statistical setting of selecting an SRS. When the population is much larger than the sample, a count of successes in an SRS of size n has approximately the Binomial distribution with n equal to the sample size and p equal to the proportion of successes in the population.

APPLY YOUR KNOWLEDGE

In each of Exercises 5.22 to 5.24, X is a count. Does X have a Binomial distribution? Give your reasons in each case.

- 5.22 You observe the sex of the next 20 children born at a local hospital; X is the number of girls among them.
- 5.23 A couple decides to continue to have children until their first girl is born; X is the total number of children the couple has.
- 5.24 A company uses a computer-based system to teach clerical employees new office software. After a lesson, the computer presents 10 exercises. The student solves each exercise and enters the answer. The computer gives additional instruction between exercises if the answer is wrong. The count X is the number of exercises that the student gets right.

Binomial probabilities*

We can find a formula for the probability that a Binomial random variable takes any value by adding probabilities for the different ways of getting

*The derivation and use of the exact formula for Binomial probabilities are optional.

exactly that many successes in n observations. An example will guide us toward the formula we want.

EXAMPLE 5.8

Determining consumer preferences

Each consumer has probability 0.25 of preferring your product over competitors' products. If we question 5 consumers, what is the probability that exactly 2 of them prefer your product?

The count of consumers preferring your product is a Binomial random variable X with $n = 5$ tries and probability $p = 0.25$ of a success on each try. We want $P(X = 2)$.



Because the method doesn't depend on the specific example, let's use "S" for success and "F" for failure for short. Do the work in two steps.

Step 1. Find the probability that a specific 2 of the 5 tries, say the first and the third, give successes. This is the outcome SFSFF. Because tries are independent, the multiplication rule for independent events applies. The probability we want is

$$\begin{aligned} P(\text{SFSFF}) &= P(\text{S})P(\text{F})P(\text{S})P(\text{F})P(\text{F}) \\ &= (0.25)(0.75)(0.25)(0.75)(0.75) \\ &= (0.25)^2(0.75)^3 \end{aligned}$$

Step 2. Observe that the probability of *any one* arrangement of 2 S's and 3 F's has this same probability. This is true because we multiply together 0.25 twice and 0.75 three times whenever we have 2 S's and 3 F's. The probability that $X = 2$ is the probability of getting 2 S's and 3 F's in any arrangement whatsoever. Here are all the possible arrangements:

SSFFF	SFSFF	SFFSF	SFFFS	FSSFF
FSFSF	FSFFS	FFSSF	FFSFS	FFFSS

There are 10 of them, all with the same probability. The overall probability of 2 successes is therefore

$$P(X = 2) = 10(0.25)^2(0.75)^3 = 0.2637$$

Approximately 26% of the time, samples of 5 independent consumers will produce exactly 2 who prefer your product over competitors' products.

The pattern of this calculation works for any Binomial probability. To use it, we must count the number of arrangements of k successes in n observations. We use the following fact to do the counting without actually listing all the arrangements.

BINOMIAL COEFFICIENT

The number of ways of arranging k successes among n observations is given by the **Binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

for $k = 0, 1, 2, \dots, n$.

factorial

The formula for Binomial coefficients uses the **factorial** notation. For any positive whole number n , its factorial $n!$ is

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

Also, $0! = 1$ by definition.

The larger of the two factorials in the denominator of a Binomial coefficient will cancel much of the $n!$ in the numerator. For example, the Binomial coefficient we need for Example 5.8 is

$$\begin{aligned}\binom{5}{2} &= \frac{5!}{2!3!} \\ &= \frac{(5)(4)(3)(2)(1)}{(2)(1) \times (3)(2)(1)} \\ &= \frac{(5)(4)}{(2)(1)} = \frac{20}{2} = 10\end{aligned}$$

The notation $\binom{n}{k}$ is *not* related to the fraction $\frac{n}{k}$. A helpful way to remember its meaning is to read it as “Binomial coefficient n choose k .” Binomial coefficients have many uses in mathematics, but we are interested in them only as an aid to finding Binomial probabilities. The Binomial coefficient $\binom{n}{k}$ counts the number of different ways in which k successes can be arranged among n observations. The Binomial probability $P(X = k)$ is this count multiplied by the probability of any specific arrangement of the k successes. Here is the result we seek.

BINOMIAL PROBABILITY

If X has the Binomial distribution with n observations and probability p of success on each observation, the possible values of X are $0, 1, 2, \dots, n$. If k is any one of these values,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

EXAMPLE 5.9**Inspecting switches**

The number X of switches that fail inspection in Case 5.1 closely follows the Binomial distribution with $n = 10$ and $p = 0.1$.

The probability that no more than 1 switch fails is

$$\begin{aligned} P(X \leq 1) &= P(X = 1) + P(X = 0) \\ &= \binom{10}{1}(0.1)^1(0.9)^9 + \binom{10}{0}(0.1)^0(0.9)^{10} \\ &= \frac{10!}{1! 9!}(0.1)(0.3874) + \frac{10!}{0! 10!}(1)(0.3487) \\ &= (10)(0.1)(0.3874) + (1)(1)(0.3487) \\ &= 0.3874 + 0.3487 = 0.7361 \end{aligned}$$

This calculation uses the facts that $0! = 1$ and that $a^0 = 1$ for any number a other than 0. We see that about 74% of all samples will contain no more than 1 bad switch. In fact, 35% of the samples will contain no bad switches. A sample of size 10 cannot be trusted to alert the engineer to the presence of unacceptable items in the shipment. Calculations such as this are used to design acceptance sampling schemes.

APPLY YOUR KNOWLEDGE

- 5.25 Inheriting blood type.** Genetics says that children receive genes from their parents independently. Each child of a particular pair of parents has probability 0.25 of having type O blood. If these parents have 5 children, the number who have type O blood is the count X of successes in 5 independent trials with probability 0.25 of a success on each trial. So X has the Binomial distribution with $n = 5$ and $p = 0.25$.
- (a) What are the possible values of X ?
 - (b) Find the probability of each value of X . Draw a probability histogram to display this distribution. (Because probabilities are long-run proportions, a histogram with the probabilities as the heights of the bars shows what the distribution of X would be in very many repetitions.)
- 5.26 Hispanic representation.** A factory employs several thousand workers, of whom 30% are Hispanic. If the 15 members of the union executive committee were chosen from the workers at random, the number of Hispanics on the committee would have the Binomial distribution with $n = 15$ and $p = 0.3$.
- (a) What is the probability that exactly 3 members of the committee are Hispanic?
 - (b) What is the probability that 3 or fewer members of the committee are Hispanic?
- 5.27 Do our athletes graduate?** A university claims that 80% of its basketball players get degrees. An investigation examines the fate of all 20 players who entered the program over a period of several years that ended six years ago. Of these players, 11 graduated and the remaining 9 are no longer in school. If the university's claim is true, the number of players who graduate among the 20 should have the Binomial distribution with $n = 20$ and $p = 0.8$. What is the probability that exactly 11 out of 20 players graduate?

Finding Binomial probabilities: tables

The formula given on page 323 for Binomial probabilities is practical for hand calculations when n is small. However, in practice, you will rarely have to use this formula for calculations. Some calculators and most statistical software packages calculate Binomial probabilities. If you do not have suitable computing facilities, you can look up the probabilities for some values of n and p in Table C in the back of this book. The entries in the table are the probabilities $P(X = k)$ of individual outcomes for a Binomial random variable X .

EXAMPLE 5.10**Inspecting switches**

The quality engineer in Case 5.1 inspects an SRS of 10 switches from a large shipment of which 10% fail to conform to specifications. What is the probability that no more than 1 of the 10 switches in the sample fails inspection?

The count X of nonconforming switches in the sample has approximately the Binomial distribution with $n = 10$ and $p = 0.1$. Figure 5.6 is a probability histogram for this distribution. The distribution is strongly skewed. Although X can take any whole-number value from 0 to 10, the probabilities of values larger than 5 are so small that they do not appear in the histogram.

We want to calculate

$$P(X \leq 1) = P(X = 1) + P(X = 0)$$

when X is Binomial with $n = 10$ and $p = 0.1$. Your software may do this—look for the key word “Binomial.” To use Table C for this calculation, look opposite

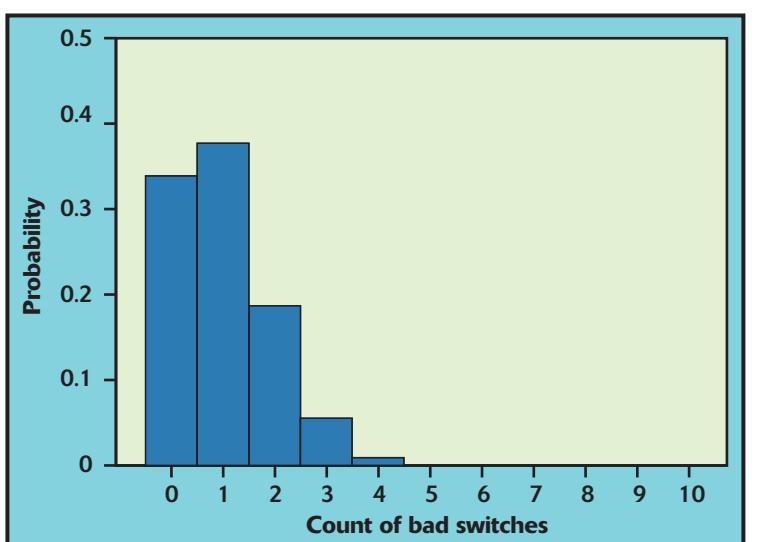


FIGURE 5.6 Probability histogram for the Binomial distribution with $n = 10$ and $p = 0.1$.

<i>n</i>	<i>k</i>	<i>p</i>
10	0	0.10
10	1	0.3487
10	2	0.3874
10	3	0.1937
10	4	0.0574
10	5	0.0112
10	6	0.0015
10	7	0.0001
10	8	0.0000
10	9	0.0000
10	10	0.0000

$n = 10$ and under $p = 0.10$. This part of the table appears at the left. The entry opposite each k is $P(X = k)$. We find

$$\begin{aligned} P(X \leq 1) &= P(X = 1) + P(X = 0) \\ &= 0.3874 + 0.3487 = 0.7361 \end{aligned}$$

About 74% of all samples will contain no more than 1 bad switch. This matches our calculation in Example 5.9.



The excerpt from Table C contains the full Binomial distribution for $n = 10$ and $p = 0.1$. The probabilities are rounded to four decimal places. Outcomes larger than 6 do not have probability exactly 0, but their probabilities are so small that the rounded values are 0.0000. Check that the sum of the probabilities given is 1, as it should be.

The values of p that appear in Table C are all 0.5 or smaller. When the probability of a success is greater than 0.5, restate the problem in terms of the number of failures. The probability of a failure is less than 0.5 when the probability of a success exceeds 0.5. When using the table, always stop to ask whether you must count successes or failures.

EXAMPLE 5.11

Free throws

Corinne is a basketball player who makes 75% of her free throws over the course of a season. In a key game, Corinne shoots 12 free throws and misses 5 of them. The fans think that she failed because she was nervous. Is it unusual for Corinne to perform this poorly?

To answer this question, assume that free throws are independent with probability 0.75 of a success on each shot. (Studies of long sequences of free throws have found no evidence that they are dependent, so this is a reasonable assumption.) Because the probability of making a free throw is greater than 0.5, we count misses in order to use Table C. The probability of a miss is $1 - 0.75$, or 0.25. The number X of misses in 12 attempts has the Binomial distribution with $n = 12$ and $p = 0.25$.

We want the probability of missing 5 or more. This is

$$\begin{aligned} P(X \geq 5) &= P(X = 5) + P(X = 6) + \cdots + P(X = 12) \\ &= 0.1032 + 0.0401 + \cdots + 0.0000 = 0.1576 \end{aligned}$$

Corinne will miss 5 or more out of 12 free throws about 16% of the time, or roughly one of every six games. While below her average level, her performance in this game was well within the range of the usual chance variation in her shooting.



- 5.28 Restaurant survey.** You operate a restaurant. You read that a sample survey by the National Restaurant Association shows that 40% of adults are committed to eating nutritious food when eating away from home. To help plan your menu, you decide to conduct a sample survey in your own area.

APPLY YOUR KNOWLEDGE

You will use random digit dialing to contact an SRS of 20 households by telephone.

- If the national result holds in your area, it is reasonable to use the Binomial distribution with $n = 20$ and $p = 0.4$ to describe the count X of respondents who seek nutritious food when eating out. Explain why.
- Ten of the 20 respondents say they are concerned about nutrition. Is this reason to believe that the percent in your area is higher than the national 40%? To answer this question, use software or Table C to find the probability that X is 10 or larger if $p = 0.4$ is true. If this probability is very small, that is reason to think that p is actually greater than 0.4.

Binomial mean and standard deviation

If a count X has the Binomial distribution based on n observations with probability p of success, what is its mean μ ? That is, in very many repetitions of the Binomial setting, what will be the average count of successes? We can guess the answer. If a basketball player makes 75% of her free throws, the mean number made in 12 tries should be 75% of 12, or 9. In general, the mean of a Binomial distribution should be $\mu = np$. Here are the facts.

BINOMIAL MEAN AND STANDARD DEVIATION

If a count X has the Binomial distribution with number of observations n and probability of success p , the **mean** and **standard deviation** of X are

$$\begin{aligned}\mu &= np \\ \sigma &= \sqrt{np(1-p)}\end{aligned}$$

Remember that these short formulas are good only for Binomial distributions. They can't be used for other distributions.

EXAMPLE 5.12

Inspecting switches

Continuing Case 5.1, the count X of bad switches is Binomial with $n = 10$ and $p = 0.1$. The mean and standard deviation of this Binomial distribution are



$$\begin{aligned}\mu &= np \\ &= (10)(0.1) = 1 \\ \sigma &= \sqrt{np(1-p)} \\ &= \sqrt{(10)(0.1)(0.9)} = \sqrt{0.9} = 0.9487\end{aligned}$$

In Figure 5.7, we have added the mean to the probability histogram of the distribution.

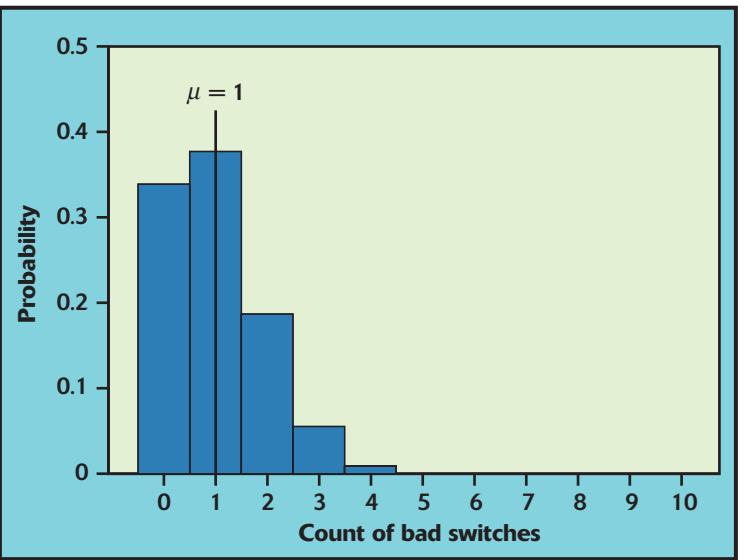


FIGURE 5.7 Probability histogram for the Binomial distribution with $n = 10$ and $p = 0.1$ and with the mean $\mu = 1$ marked.

APPLY YOUR KNOWLEDGE

- 5.29 **Restaurant survey.** As in Exercise 5.28, you ask an SRS of 20 adults from your restaurant's target area if they are concerned about nutrition when eating away from home. If the national proportion $p = 0.4$ holds in your area, what will be the mean number of "Yes" responses? What is the standard deviation of the count of "Yes" answers?
- 5.30 **Hispanic representation.**
- What is the mean number of Hispanics on randomly chosen committees of 15 workers in Exercise 5.26?
 - What is the standard deviation σ of the count X of Hispanic members?
 - Suppose that 10% of the factory workers were Hispanic. Then $p = 0.1$. What is σ in this case? What is σ if $p = 0.01$? What does your work show about the behavior of the standard deviation of a Binomial distribution as the probability of a success gets closer to 0?
- 5.31 **Do our athletes graduate?**
- Find the mean number of graduates out of 20 players in the setting of Exercise 5.27 if the university's claim is true.
 - Find the standard deviation σ of the count X .
 - Suppose that the 20 players came from a population of which $p = 0.9$ graduated. What is the standard deviation σ of the count of graduates? If $p = 0.99$, what is σ ? What does your work show about the behavior of the standard deviation of a Binomial distribution as the probability p of success gets closer to 1?

The Normal approximation to Binomial distributions

The Binomial probability formula and tables are practical only when the number of trials n is small. Even software and statistical calculators are

unable to handle calculations for very large n . Here is another alternative: *as the number of trials n gets larger, the Binomial distribution gets close to a Normal distribution.* When n is large, we can use Normal probability calculations to approximate hard-to-calculate Binomial probabilities. For an example, we return to the survey discussed in Case 3.1 (page 206).

EXAMPLE 5.13**Is clothes shopping frustrating?**

Sample surveys show that fewer people enjoy shopping than in the past. A recent survey asked a nationwide random sample of 2500 adults if they agreed or disagreed that “I like buying new clothes, but shopping is often frustrating and time-consuming.”⁵ The population that the poll wants to draw conclusions about is *all* U.S. residents aged 18 and over. Suppose that in fact 60% of *all* adult U.S. residents would say “agree” if asked the same question. What is the probability that 1520 or more of the sample agree?



Because there are almost 210 million adults, we can take the responses of 2500 randomly chosen adults to be independent. The number in our sample who agree that shopping is frustrating is a random variable X having the Binomial distribution with $n = 2500$ and $p = 0.6$. To find the probability that at least 1520 of the people in the sample find shopping frustrating, we must add the Binomial probabilities of all outcomes from $X = 1520$ to $X = 2500$. This isn’t practical. Here are three ways to do this problem:

1. Statistical software (but not the Excel spreadsheet program) can do the calculation. The result is

$$P(X \geq 1520) = 0.2131$$

2. We can simulate a large number of repetitions of the sample. Figure 5.8 displays a histogram of the counts X from 1000 samples of size 2500 when the truth about the population is $p = 0.6$. Because 221 of these 1000 samples have X at least 1520, the probability estimated from the simulation is

$$P(X \geq 1520) = \frac{221}{1000} = 0.221$$

3. Both of the previous methods require software. Instead, look at the Normal curve in Figure 5.8. This is the density curve of the Normal distribution with the same mean and standard deviation as the Binomial variable X :

$$\mu = np = (2500)(0.6) = 1500$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{(2500)(0.6)(0.4)} = 24.49$$

As the figure shows, this Normal distribution approximates the Binomial distribution quite well. So we can do a Normal calculation.

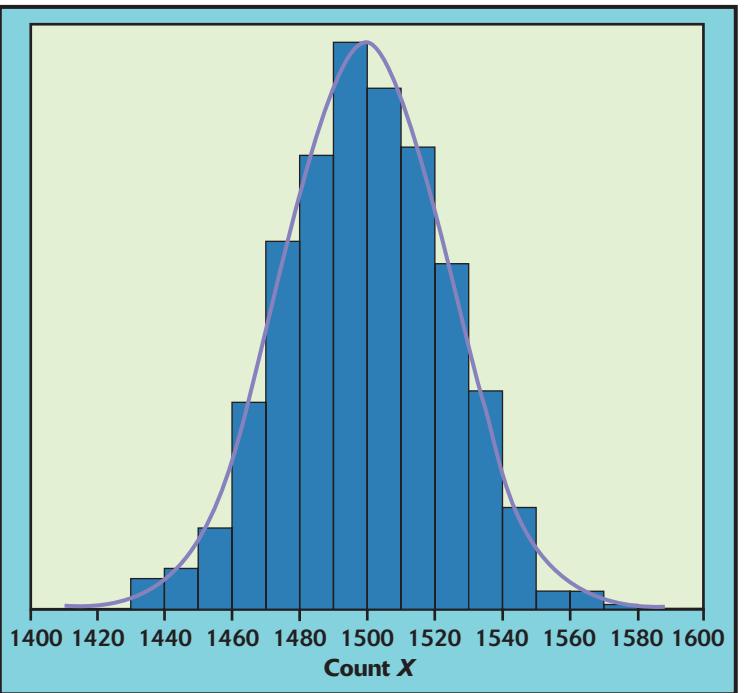


FIGURE 5.8 Histogram of 1000 Binomial counts ($n = 2500, p = 0.6$) and the Normal density curve that approximates this Binomial distribution.

EXAMPLE 5.14

Normal calculation of a Binomial probability

Act as though the count X had the $N(1500, 24.49)$ distribution. Here is the probability we want, using Table A:



$$\begin{aligned} P(X \geq 1520) &= P\left(\frac{X - 1500}{24.49} \geq \frac{1520 - 1500}{24.49}\right) \\ &= P(Z \geq 0.82) \\ &= 1 - 0.7939 = 0.2061 \end{aligned}$$

The Normal approximation 0.2061 differs from the software result 0.2131 by only 0.007.



NORMAL APPROXIMATION FOR BINOMIAL DISTRIBUTIONS

Suppose that a count X has the Binomial distribution with n trials and success probability p . When n is large, the distribution of X is approximately Normal, $N(np, \sqrt{np(1-p)})$.

As a rule of thumb, we will use the Normal approximation when n and p satisfy $np \geq 10$ and $n(1-p) \geq 10$.

The Normal approximation is easy to remember because it says that X is Normal with its Binomial mean and standard deviation. The accuracy of the Normal approximation improves as the sample size n increases. It is most accurate for any fixed n when p is close to 1/2, and least accurate when p is near 0 or 1. Whether or not you use the Normal approximation should depend on how accurate your calculations need to be. For most statistical purposes great accuracy is not required. Our “rule of thumb” for use of the Normal approximation reflects this judgment.

APPLY YOUR KNOWLEDGE

- 5.32 **Restaurant survey.** Return to the survey described in Exercise 5.28. You plan to use random digit dialing to contact an SRS of 200 households by telephone rather than just 20.
- What are the mean and standard deviation of the number of nutrition-conscious people in your sample if $p = 0.4$ is true?
 - What is the probability that X lies between 75 and 85? (Use the Normal approximation.)
- 5.33 **The effect of sample size.** The SRS of size 200 described in the previous exercise finds that 100 of the 200 respondents are concerned about nutrition. We wonder if this is reason to conclude that the percent in your area is higher than the national 40%.
- Find the probability that X is 100 or larger if $p = 0.4$ is true. If this probability is very small, that is reason to think that p is actually greater than 0.4.
 - In Exercise 5.28, you found $P(X \geq 10)$ for a sample of size 20. In (a), you have found $P(X \geq 100)$ for a sample of size 200 from the same population. Both of these probabilities answer the question “How likely is a sample with at least 50% successes when the population has 40% successes?” What does comparing these probabilities suggest about the importance of sample size?

SECTION 5.2 SUMMARY

- A count X of successes has a **Binomial distribution** in the **Binomial setting**: the number of observations n is fixed in advance; the observations are independent of each other; each observation results in a success or a failure; and each observation has the same probability p of a success.
- The Binomial distribution with n observations and probability p of success gives a good approximation to the sampling distribution of the count of successes in an SRS of size n from a large population containing proportion p of successes.
- If X has the Binomial distribution with parameters n and p , the possible values of X are the whole numbers $0, 1, 2, \dots, n$. The **Binomial probability** that X takes any value is

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Binomial probabilities are most easily found by software. This formula is practical for calculations when n is small. Table C contains Binomial probabilities for some values of n and p . For large n , you can use the Normal approximation.

■ **The Binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

counts the number of ways k successes can be arranged among n observations. Here the factorial $n!$ is

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

for positive whole numbers n , and $0! = 1$.

■ **The mean and standard deviation** of a Binomial count X are

$$\mu = np$$

$$\sigma = \sqrt{np(1-p)}$$

■ **The Normal approximation** to the Binomial distribution says that if X is a count having the Binomial distribution with parameters n and p , then when n is large, X is approximately $N(np, \sqrt{np(1-p)})$. We will use this approximation when $np \geq 10$ and $n(1-p) \geq 10$.

SECTION 5.2 EXERCISES

All of the Binomial probability calculations required in these exercises can be done by using Table C or the Normal approximation. Your instructor may request that you use the Binomial probability formula or software.

- 5.34 **Binomial setting?** In each situation below, is it reasonable to use a Binomial distribution for the random variable X ? Give reasons for your answer in each case.
- (a) An auto manufacturer chooses one car from each hour's production for a detailed quality inspection. One variable recorded is the count X of finish defects (dimples, ripples, etc.) in the car's paint.
 - (b) Joe buys a ticket in his state's "Pick 3" lottery game every week; X is the number of times in a year that he wins a prize.
- 5.35 **Binomial setting?** In each of the following cases, decide whether or not a Binomial distribution is an appropriate model, and give your reasons.
- (a) A firm uses a computer-based training module to prepare 20 machinists to use new numerically controlled lathes. The module contains a test at the end of the course; X is the number who perform satisfactorily on the test.
 - (b) The list of potential product testers for a new product contains 100 persons chosen at random from the adult residents of a large city. Each person on the list is asked whether he or she would participate in the study if given the chance; X is the number who say "Yes."

- 5.36 **Random digits.** Each entry in a table of random digits like Table B has probability 0.1 of being a 0, and digits are independent of each other.
- What is the probability that a group of five digits from the table will contain at least one 0?
 - What is the mean number of 0s in lines 40 digits long?
- 5.37 **Unmarried women.** Among employed women, 25% have never been married. Select 10 employed women at random.
- The number in your sample who have never been married has a Binomial distribution. What are n and p ?
 - What is the probability that exactly 2 of the 10 women in your sample have never been married?
 - What is the probability that 2 or fewer have never been married?
 - What is the mean number of women in such samples who have never been married? What is the standard deviation?
- 5.38 **Generic brand soda.** In a taste test of a generic soda versus a brand name soda, 25% of tasters can distinguish between the colas. Twenty tasters are asked to take the taste test and guess which cup contains the brand name soda. The tests are done independently in separate locations, so that the tasters do not interact with each other during the test.
- The count of correct guesses in 20 taste tests has a Binomial distribution. What are n and p ?
 - What is the mean number of correct guesses in many repetitions?
 - What is the probability of exactly 5 correct guesses?
- 5.39 **Random stock prices.** A believer in the “random walk” theory of stock markets thinks that an index of stock prices has probability 0.65 of increasing in any year. Moreover, the change in the index in any given year is not influenced by whether it rose or fell in earlier years. Let X be the number of years among the next 5 years in which the index rises.
- X has a Binomial distribution. What are n and p ?
 - What are the possible values that X can take?
 - Find the probability of each value of X . Draw a probability histogram for the distribution of X .
 - What are the mean and standard deviation of this distribution? Mark the location of the mean on your histogram.
- 5.40 **Lie detectors.** A federal report finds that lie detector tests given to truthful persons have probability about 0.2 of suggesting that the person is deceptive.⁶
- A company asks 12 job applicants about thefts from previous employers, using a lie detector to assess their truthfulness. Suppose that all 12 answer truthfully. What is the probability that the lie detector says all 12 are truthful? What is the probability that the lie detector says at least 1 is deceptive?
 - What is the mean number among 12 truthful persons who will be classified as deceptive? What is the standard deviation of this number?
 - What is the probability that the number classified as deceptive is less than the mean?

- 5.41 **Multiple-choice tests.** Here is a simple probability model for multiple-choice tests. Suppose that each student has probability p of correctly answering a question chosen at random from a universe of possible questions. (A strong student has a higher p than a weak student.) Answers to different questions are independent. Jodi is a good student for whom $p = 0.75$.
- Use the Normal approximation to find the probability that Jodi scores 70% or lower on a 100-question test.
 - If the test contains 250 questions, what is the probability that Jodi will score 70% or lower?
- 5.42 **Mark McGwire's home runs.** In 1998, Mark McGwire of the St. Louis Cardinals hit 70 home runs, a new major-league record. Was this feat as surprising as most of us thought? In the three seasons before 1998, McGwire hit a home run in 11.6% of his times at bat. He went to bat 509 times in 1998. McGwire's home-run count in 509 times at bat has approximately the Binomial distribution with $n = 509$ and $p = 0.116$.
- What is the mean number of home runs he will hit in 509 times at bat?
 - What is the probability of 70 or more home runs? Use the Normal approximation.
 - Compare your answer in (b) to the actual probability of 0.0764 found using software.

<>	A	B	C
1	$1 - \text{BINOMDIST}(69, 509, 0.116, 1) =$	0.0764	
2			

- 5.43 **Planning a survey.** You are planning a sample survey of small businesses in your area. You will choose an SRS of businesses listed in the telephone book's Yellow Pages. Experience shows that only about half the businesses you contact will respond.
- If you contact 150 businesses, it is reasonable to use the Binomial distribution with $n = 150$ and $p = 0.5$ for the number X who respond. Explain why.
 - What is the expected number (the mean) who will respond?
 - What is the probability that 70 or fewer will respond? (Use the Normal approximation.)
 - How large a sample must you take to increase the mean number of respondents to 100?
- 5.44 **Are we shipping on time?** Your mail-order company advertises that it ships 90% of its orders within three working days. You select an SRS of 100 of the 5000 orders received in the past week for an audit. The audit reveals that 86 of these orders were shipped on time.
- If the company really ships 90% of its orders on time, what is the probability that 86 or fewer in an SRS of 100 orders are shipped on time?
 - A critic says, "Aha! You claim 90%, but in your sample the on-time percentage is only 86%. So the 90% claim is wrong." Explain in simple language why your probability calculation in (a) shows that the result of the sample does not refute the 90% claim.

- 5.45 **Checking for survey errors.** One way of checking the effect of undercoverage, nonresponse, and other sources of error in a sample survey is to compare the sample with known facts about the population. About 12% of American adults are black. The number X of blacks in a random sample of 1500 adults should therefore vary with the Binomial ($n = 1500, p = 0.12$) distribution.
- What are the mean and standard deviation of X ?
 - Use the Normal approximation to find the probability that the sample will contain 170 or fewer blacks. Be sure to check that you can safely use the approximation.

5.3 The Poisson Distributions

Not all counts have Binomial distributions. It is common to meet counts that are open-ended, that is, that do not have the fixed number of observations required by the Binomial model. Count the number of finish defects in the sheet metal of a car or a refrigerator: the count could be 0, 1, 2, 3, and so on indefinitely. A bank counts the number of automatic teller machine (ATM) customers arriving at a particular ATM between 2:00 p.m. and 4:00 p.m. A railyard counts the number of work injuries that happen in a month. All of these count examples share common characteristics.

The Poisson setting

The Poisson distribution is another distribution for counting random variables. Count the number of events (call them “successes”) that occur in some fixed unit of measure such as an area of sheet metal, a period of time, or a length of cable. The Poisson distribution is appropriate in the following situation.

THE POISSON SETTING

- The number of successes that occur in any unit of measure is **independent** of the number of successes that occur in any nonoverlapping unit of measure.
- The probability that a success will occur in a unit of measure is the same for all units of equal size and is proportional to the size of the unit.
- The probability that 2 or more successes will occur in a unit approaches 0 as the size of the unit becomes smaller.

For Binomial distributions, the important quantities were n , the fixed number of observations, and p , the probability of success on any given observation. The quantity important in specifying Poisson distributions is the mean number of successes occurring per unit of measure.

POISSON DISTRIBUTION

The distribution of the count X of successes in the Poisson setting is the **Poisson distribution** with mean μ . The parameter μ is the mean number of successes per unit of measure. The possible values of X are the whole numbers $0, 1, 2, 3, \dots$. If k is any whole number 0 or greater, then*

$$P(X = k) = \frac{e^{-\mu} \mu^k}{k!}$$

The standard deviation of the distribution is $\sqrt{\mu}$.

EXAMPLE 5.15**Flaws in carpets**

A carpet manufacturer knows that the number of flaws per square yard in a type of carpet material varies with an average of 1.6 flaws per square yard. The count X of flaws per square yard can be modeled by the Poisson distribution with $\mu = 1.6$. The unit of measure is a square yard of carpet material. What is the probability of no more than 2 defects in a randomly chosen square yard of this material?

We will calculate $P(X \leq 2)$ in two ways:

1. Software can do the calculation:

	A	B	C
1	Poisson(0, 1.6, 0)=	0.2019	
2	Poisson(0, 1.6, 0)=	0.3230	
3	Poisson(2, 1.6, 0)=	0.2584	
4	SUM(B1 :B3)=	0.7834	
5			

2. We can use the Poisson probability formula:

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \frac{e^{-1.6}(1.6)^0}{0!} + \frac{e^{-1.6}(1.6)^1}{1!} + \frac{e^{-1.6}(1.6)^2}{2!} \\ &= 0.2019 + 0.3230 + 0.2584 = 0.7833 \end{aligned}$$

The software answer and the hand calculation differ by 0.0001 due to roundoff error in the hand calculation. The software calculates the individual probabilities to many significant digits even though it displays only four significant digits.

cumulative probability

Recall that Table A gives cumulative probabilities of the form $P(X \leq k)$ for the standard Normal distribution. Most software will calculate cumulative

*The quantity e in the Poisson probability formula is a mathematical constant, $e = 2.71828$ to six significant digits. Many calculators have an e^x function.

probabilities for other distributions, including the Poisson family. Cumulative probability calculations make solving many problems less tedious.

EXAMPLE 5.16**Counting ATM customers**

Suppose the number of persons using an ATM in any given hour can be modeled by a Poisson distribution with $\mu = 5.5$. What is the probability of more than 8 persons using the machine during the next hour? Calculating this probability requires two steps:

1. Write $P(X > 8)$ as an expression involving a cumulative probability:

$$P(X > 8) = 1 - P(X \leq 8)$$

2. Calculate $P(X \leq 8)$ and subtract the value from 1.

◊	A	B	C
1	1-Poisson(8,5.5,1)=	0.1056	
2			

This is quicker and less prone to error than the method of Example 5.15, which would require specifying nine individual probabilities and summing their values.

**The Poisson model**

If we add counts of successes in nonoverlapping areas of space or time, we are just counting the successes in a larger area. That count still meets the conditions of the Poisson setting. Put more formally, if X is a Poisson random variable with mean μ_X and Y is a Poisson random variable with mean μ_Y and Y is independent of X , then $X + Y$ is a Poisson random variable with mean $\mu_X + \mu_Y$. This fact is important in using Poisson models. We can combine areas or look at just a portion of an area and still use Poisson distributions for counts of successes.

EXAMPLE 5.17**Paint finish flaws**

Auto bodies are painted during manufacture by robots programmed to move in such a way that the paint is uniform in thickness and quality. You are testing a newly programmed robot by counting paint sags caused by small areas receiving too much paint. Sags are more common on vertical surfaces. Suppose that counts of sags on the roof follow the Poisson model with mean 0.7 sags per square yard and that counts on the side panels of the auto body follow the Poisson model with mean 1.4 sags per square yard. Counts in nonoverlapping areas are independent. Then

- The number of sags in 2 square yards of roof is a Poisson random variable with mean $0.7 + 0.7 = 1.4$.
- The total roof area of the auto body is 4.8 square yards. The number of paint sags on a roof is a Poisson random variable with mean $4.8 \times 0.7 = 3.36$.

- A square foot is $1/9$ square yard. The number of paint sags in a square foot of roof is a Poisson random variable with mean $1/9 \times 0.7 = 0.078$.
- If we examine 1 square yard of roof and 1 square yard of side panel, the number of sags is a Poisson random variable with mean $0.7 + 1.4 = 2.1$.

APPLY YOUR KNOWLEDGE

- 5.46 **Industrial accidents.** A large manufacturing plant has averaged 7 “reportable accidents” per month. Suppose that accident counts over time follow a Poisson distribution with mean 7 per month.
- What is the probability of exactly 7 accidents in a month?
 - What is the probability of 7 or fewer accidents in a month?
- 5.47 **A safety initiative.** This year, a “safety culture change” initiative attempts to reduce the number of accidents at the plant described in the previous exercise. There are 66 reportable accidents during the year. Suppose that the Poisson distribution of the previous exercise continues to apply.
- What is the distribution of the number of reportable accidents in a year?
 - What is the probability of 66 or fewer accidents in a year? (Use software.) The probability is small, which is evidence that the initiative did reduce the accident rate.

BEYOND THE BASICS: MORE DISTRIBUTION APPROXIMATIONS

In Section 5.2, we observed that the Normal distribution could be used to calculate Binomial probabilities when n , the number of trials, is large (page 330). When n is large, the Binomial probability histogram has the familiar mound shape of the Normal density curve. This fact allows us to use Normal probabilities and to avoid tedious hand calculations or the need to use software to calculate Binomial probabilities.

Using the Normal distribution to approximate the Binomial distribution is just one example of using one distribution to approximate another to make probability calculations more convenient. With the distributions we have studied, two more approximations are common:

- **Normal approximation to the Poisson.** The Excel spreadsheet program returns an error when asked to calculate $P(X \leq 142)$ for a Poisson random variable X with mean $\mu = 150$. What can we do if our software cannot handle Poisson distributions with large means? Fortunately, when μ is large, Poisson probabilities can be approximated using the Normal distribution with mean μ and standard deviation $\sqrt{\mu}$. The following table compares $P(X \leq k)$ for a Poisson random variable X with $\mu = 150$ with approximations using the $N(150, 12.247)$ distribution.

k	Poisson	Normal
142	0.2730	0.2568
150	0.5217	0.5000
160	0.8054	0.7929

The Normal approximation is adequate for many practical purposes, but we recommend statistical software that can give exact Poisson probabilities.

- **Poisson approximation to the Binomial.** We recommend using the Normal approximation to a Binomial distribution only when n and p satisfy $np \geq 10$ and $n(1-p) \geq 10$. In cases where p is so small that $np < 10$, using the Poisson distribution with $\mu = np$ to calculate Binomial probabilities yields more accurate results. The following table compares $P(X \leq k)$ for a Binomial distribution with $n = 1000$ and $p = 0.001$ with Poisson probabilities calculated using $\mu = np = (1000)(0.001) = 1$.

k	Binomial	Poisson
0	0.3677	0.3679
1	0.7358	0.7358
2	0.9198	0.9197
3	0.9811	0.9810
4	0.9964	0.9963
5	0.9994	0.9994
6	0.9999	0.9999
7	1.0000	1.0000

The Poisson approximation gives very accurate probability calculations for the Binomial distribution with $n = 1000$ and $p = 0.001$.

Even statistical software has its limits, and some Binomial and Poisson probability calculations can exceed those limits. In many cases, however, one of the approximations we have discussed will make calculations possible.

SECTION 5.3 SUMMARY

- A count X of successes has a **Poisson distribution in the Poisson setting**: the number of successes in any unit of measure is independent of the number of successes in any other nonoverlapping unit; the probability of a success in a unit of measure is the same for all units of equal size and is proportional to the size of the unit; the probability of 2 or more successes in a unit approaches 0 as the size of the unit becomes smaller.
- If X has the Poisson distribution with mean μ , then the standard deviation of X is $\sqrt{\mu}$, and the possible values of X are all the whole numbers 0, 1, 2, 3, and so on. The **Poisson probability** that X takes any one of these values is

$$P(X = k) = \frac{e^{-\mu} \mu^k}{k!} \quad k = 0, 1, 2, 3, \dots$$

- Poisson probabilities are most easily found by software. The formula above is practical when only a small number of probabilities is needed and k is not large.

- Sums of independent Poisson random variables also have the Poisson distribution. In a Poisson model with mean μ per unit of space or time, the count of successes in a units is a Poisson random variable with mean $a\mu$.

SECTION 5.3 EXERCISES

Use software to calculate the Poisson probabilities in the following exercises.

- 5.48 **Too much email?** According to email logs, one employee at your company receives an average of 110 emails per week. Suppose the count of emails received can be adequately modeled as a Poisson random variable.
- What is the probability of this employee receiving exactly 110 emails in a given week?
 - What is the probability of receiving 100 or fewer emails in a given week?
 - What is the probability of receiving more than 125 emails in a given week?
 - What is the probability of receiving 125 or more emails in a given week?
(Be careful: this is not the same event as in part (c).)
- 5.49 **Traffic model.** The number of vehicles passing a particular mile marker during 15-minute units of time can be modeled as a Poisson random variable. Counting devices show that the average number of vehicles passing the mile marker per 15 minutes is 48.7.
- What is the probability of 50 or more vehicles passing the marker during a 15-minute time period?
 - What is the standard deviation of the number of vehicles passing the marker in a 15-minute time period? A 30-minute time period?
 - What is the probability of 100 or more vehicles passing the marker during a 30-minute time period?
- 5.50 **Too much email?** According to email logs, one employee at your company receives an average of 110 emails per week. Suppose the count of emails received can be adequately modeled as a Poisson random variable.
- What is the distribution of the number of emails in a two-week period?
 - What is the probability of receiving 200 or fewer emails in a two-week period?
- 5.51 **Work-related deaths.** Work-related deaths in the United States have a mean of 17 per day. Suppose the count of work-related deaths per day follows an approximate Poisson distribution.
- What is the standard deviation for daily work-related deaths?
 - What is the probability of 10 or fewer work-related deaths in one day?
 - What is the probability of more than 30 work-related deaths in one day?
- 5.52 **Flaws in carpets.** Flaws in carpet material follow the Poisson model with mean 1.6 flaws per square yard. An inspector examines 100 randomly selected square yard specimens of the material, records the number of flaws found in each specimen, and calculates \bar{x} , the average number of flaws per square yard inspected.
- The total number of flaws $100\bar{x}$ is a Poisson random variable. What is its mean?

- (b) What is the probability that the total number of flaws $100\bar{x}$ exceeds 110?
- (c) We can use the central limit theorem (page 292) to calculate the same probability as in part (b) by realizing that $P(100\bar{x} > 110) = P(\bar{x} > 110/100)$. What is the probability found using the central limit theorem?
- (d) Compare your answers to (b) and (c). How close are the two answers? Which one is more accurate and why?
- 5.53 **Calling tech support.** The number of calls received between 8 a.m. and 9 a.m. by a software developer's technical support line has a Poisson distribution with a mean of 14.
- (a) What is the probability of at least 5 calls between 8 a.m. and 9 a.m.?
 - (b) What is the probability of at least 5 calls between 8:15 a.m. and 8:45 a.m.?
 - (c) What is the probability of at least 5 calls between 8:15 a.m. and 8:30 a.m.?
- 5.54 **Web site hits.** A “hit” for a Web site is a request for a file from the Web site's server computer. Some popular Web sites have thousands of hits per minute. One popular Web site boasts an average of 6500 hits per minute between the hours of 9 a.m. and 6 p.m. Some weaker software packages will have trouble calculating Poisson probabilities with such a large value of μ .
- (a) Try calculating the probability of 6400 hits or more during the minute beginning at 10:05 a.m. using the software that you have available. Did you get an answer? If not, how did the software respond?
 - (b) Now, use the central limit theorem to calculate the probability of 6400 hits or more during the minute beginning at 10:05 a.m. To do this, think of the number of hits in this minute as the sum of the number of hits for each of the 60 seconds in this minute. We can express $P(\text{sum of hits for each of the 60 seconds} \geq 6400)$ as $P(\bar{x} \geq 6400/60)$ where \bar{x} is the average number of hits per second for the 60 seconds in the minute of interest.
- 5.55 **Credit card manufacturing.** Large sheets of plastic are cut into smaller pieces to be pressed into credit cards. One manufacturer uses sheets of plastic known to have approximately 2.3 defects per square yard. The number of defects can be modeled as a Poisson random variable X .
- (a) What is the standard deviation of the number of defects per square yard?
 - (b) What is the probability of an inspector finding more than 5 defects in a randomly chosen square yard?
 - (c) Using trial and error with your software, find the largest value k such that $P(X > k) \geq 0.15$.
- 5.56 **Initial public offerings.** The number of companies making their initial public offering of stock (IPO) can be modeled by a Poisson distribution with a mean of 15 per month.
- (a) What is the probability of fewer than 3 IPOs in a month?
 - (b) What is the probability of fewer than 15 IPOs in a month?
 - (c) What is the probability of fewer than 30 IPOs in a two-month period?
 - (d) What is the probability of fewer than 180 IPOs in a year?

5.4 Conditional Probability

In Section 2.5 we met the idea of a *conditional distribution*, the distribution of a variable given that a condition is satisfied. Now we will introduce the probability language for this idea.

EXAMPLE 5.18

Employment revisited



The discussion of unemployment rates in Case 1.1 (page 9) pointed out that the government has very specific definitions of terms like “in the labor force” and “unemployed.” Using those definitions, the following table contains counts (in thousands) of persons aged 16 to 24 who are enrolled in school classified by gender and employment status:

	Employed	Unemployed	Not in labor force	Total
Male	3,927	520	4,611	9,058
Female	4,313	446	4,357	9,116
Total	8,240	966	8,968	18,174

Randomly choose a person aged 16 to 24 who is enrolled in school. What is the probability that the person is employed? Because “choose at random” gives all 18,174,000 such persons the same chance, the probability is just the proportion that are employed. In thousands,

$$P(\text{employed}) = \frac{8,240}{18,174} = 0.4534$$

Now we are told that the person chosen is female. The probability the person is employed, *given the information that the person is female*, is

$$P(\text{employed} | \text{female}) = \frac{4,313}{9,116} = 0.4731$$

conditional probability

This is a conditional probability.



The conditional probability 0.4731 in Example 5.18 gives the probability of one event (the person chosen is employed) under the condition that we know another event (the person is female). You can read the bar | as “given the information that.” We found the conditional probability by applying common sense to the two-way table.

We want to turn this common sense into something more general. To do this, we reason as follows. To find the proportion of 16- to 24-year-olds enrolled in school who are *both* female *and* employed, first find the proportion of females in the group of interest (16- to 24-year-olds enrolled in school). Then multiply by the proportion of these females who are employed. If 20%

are female and half of these are employed, then half of 20%, or 10%, are females who are employed. The actual proportions from Example 5.18 are

$$\begin{aligned} P(\text{female and employed}) &= P(\text{female}) \times P(\text{employed} | \text{female}) \\ &= (0.5016)(0.4731) = 0.2373 \end{aligned}$$

You can check that this is right: the probability that a randomly chosen person from this group is a female who is employed is

$$P(\text{female and employed}) = \frac{4,313}{18,174} = 0.2373$$

Try to think your way through this in words before looking at the formal notation. We have just discovered the general multiplication rule of probability.

GENERAL MULTIPLICATION RULE FOR ANY TWO EVENTS

The probability that both of two events A and B happen together can be found by

$$P(A \text{ and } B) = P(A)P(B | A)$$

Here $P(B | A)$ is the conditional probability that B occurs given the information that A occurs.

In words, this rule says that for both of two events to occur, first one must occur and then, given that the first event has occurred, the second must occur.

EXAMPLE 5.19

Focus group probabilities

A focus group of 10 consumers has been selected to view a new TV commercial. After the viewing, 2 members of the focus group will be randomly selected and asked to answer detailed questions about the commercial. The group contains 4 men and 6 women. What is the probability that the 2 chosen to answer questions will both be women?

To find the probability of randomly selecting 2 women, first calculate

$$P(\text{first person is female}) = \frac{6}{10}$$

$$P(\text{second person is female} | \text{first person is female}) = \frac{5}{9}$$

Both probabilities are found by counting group members. The probability that the first person selected is a female is $6/10$ because 6 of the 10 group members are female. If the first person is a female, that leaves 5 females among the 9 remaining people. So the *conditional* probability of another female is $5/9$. The multiplication rule now says that

$$P(\text{both people are female}) = \frac{6}{10} \times \frac{5}{9} = \frac{1}{3} = 0.3333$$

One-third of the time, randomly picking 2 people from a group of 4 males and 6 females will result in a pair of females.



Remember that events A and B play different roles in the conditional probability $P(B | A)$. Event A represents the information we are given, and B is the event whose probability we are computing.

EXAMPLE 5.20

Internet users

About 20% of all Web surfers use Macintosh computers. About 90% of all Macintosh users surf the Web. If you know someone who uses a Macintosh computer, then the probability that that person surfs the Web is

$$P(\text{surfs the Web} | \text{Macintosh user}) = 0.90$$

The 20% is a different conditional probability that does not apply when you are considering someone who you know uses a Macintosh computer.



The general multiplication rule also extends to the probability that all of several events occur. The key is to condition each event on the occurrence of *all* of the preceding events. For example, for three events A , B , and C ,

$$P(A \text{ and } B \text{ and } C) = P(A)P(B | A)P(C | A \text{ and } B)$$

EXAMPLE 5.21

The future of high school athletes

Only 5% of male high school basketball, baseball, and football players go on to play at the college level. Of these, only 1.7% enter major league professional sports. About 40% of the athletes who compete in college and then reach the pros have a career of more than 3 years.⁷ Define these events:

$$A = \{\text{competes in college}\}$$

$$B = \{\text{competes professionally}\}$$

$$C = \{\text{pro career longer than 3 years}\}$$

What is the probability that a high school athlete competes in college and then goes on to have a pro career of more than 3 years? We know that

$$P(A) = 0.05$$

$$P(B | A) = 0.017$$

$$P(C | A \text{ and } B) = 0.4$$

The probability we want is therefore

$$\begin{aligned} P(A \text{ and } B \text{ and } C) &= P(A)P(B | A)P(C | A \text{ and } B) \\ &= 0.05 \times 0.017 \times 0.40 = 0.00034 \end{aligned}$$

Only about 3 of every 10,000 high school athletes can expect to compete in college and have a professional career of more than 3 years. High school athletes would be wise to concentrate on studies rather than on unrealistic hopes of fortune from pro sports.

**APPLY YOUR KNOWLEDGE**

- 5.57 **Woman managers.** Choose an employed person at random. Let A be the event that the person chosen is a woman, and B the event that the person holds a managerial or professional job. Government data tell us that $P(A) = 0.46$ and the probability of managerial and professional jobs among women is $P(B | A) = 0.32$. Find the probability that a randomly chosen employed person is a woman holding a managerial or professional position.
- 5.58 **Buying from Japan.** Functional Robotics Corporation buys electrical controllers from a Japanese supplier. The company's treasurer thinks that there is probability 0.4 that the dollar will fall in value against the Japanese yen in the next month. The treasurer also believes that if the dollar falls there is probability 0.8 that the supplier will demand renegotiation of the contract. What probability has the treasurer assigned to the event that the dollar falls and the supplier demands renegotiation?
- 5.59 **Employment revisited.** Use the two-way table in Example 5.18 to find these conditional probabilities.
- (a) $P(\text{employed} | \text{male})$
 - (b) $P(\text{male} | \text{employed})$
 - (c) $P(\text{female} | \text{unemployed})$
 - (d) $P(\text{unemployed} | \text{female})$

**Conditional probability and independence**

If we know $P(A)$ and $P(A \text{ and } B)$, we can rearrange the multiplication rule to produce a *definition* of the conditional probability $P(B | A)$ in terms of unconditional probabilities.

DEFINITION OF CONDITIONAL PROBABILITY

When $P(A) > 0$, the conditional probability of B given A is

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)}$$

The conditional probability $P(B | A)$ makes no sense if the event A can never occur, so we require that $P(A) > 0$ whenever we talk about $P(B | A)$. The definition of conditional probability reminds us that in principle all probabilities, including conditional probabilities, can be found from the assignment of probabilities to events that describe a random phenomenon. More often, as in Examples 5.18 and 5.19, conditional probabilities are part

of the information given to us in a probability model, and the multiplication rule is used to compute $P(A \text{ and } B)$.

The conditional probability $P(B | A)$ is generally not equal to the unconditional probability $P(B)$. That is because the occurrence of event A generally gives us some additional information about whether or not event B occurs. If knowing that A occurs gives no additional information about B , then A and B are independent events. The precise definition of independence is expressed in terms of conditional probability.

INDEPENDENT EVENTS

Two events A and B that both have positive probability are independent if

$$P(B | A) = P(B)$$

This definition makes precise the informal description of independence given in Section 5.1. We now see that the multiplication rule for independent events, $P(A \text{ and } B) = P(A)P(B)$, is a special case of the general multiplication rule, $P(A \text{ and } B) = P(A)P(B | A)$, just as the addition rule for disjoint events is a special case of the general addition rule. We will rarely use the definition of independence, because most often independence is part of the information given to us in a probability model.

APPLY YOUR KNOWLEDGE

- 5.60 **College degrees.** Here are the counts (in thousands) of earned degrees in the United States in the 2001–2002 academic year, classified by level and by the gender of the degree recipient:⁸

	Bachelor's	Master's	Professional	Doctorate	Total
Female	645	227	32	18	922
Male	505	161	40	26	732
Total	1150	388	72	44	1654

- (a) If you choose a degree recipient at random, what is the probability that the person you choose is a woman?
 - (b) What is the conditional probability that you choose a woman, given that the person chosen received a professional degree?
 - (c) Are the events “choose a woman” and “choose a professional degree recipient” independent? How do you know?
- 5.61 **Prosperity and education.** Call a household prosperous if its income exceeds \$100,000. Call the household educated if the householder completed college. Select an American household at random, and let A be the event that the selected household is prosperous and B the event that it is educated. According to the Current Population Survey, $P(A) = 0.134$, $P(B) = 0.254$, and the probability that a household is both prosperous and educated is $P(A \text{ and } B) = 0.080$.

- Find the conditional probability that a household is educated, given that it is prosperous.
- Find the conditional probability that a household is prosperous, given that it is educated.
- Are events A and B independent? How do you know?

Tree diagrams and Bayes's rule

Probability problems often require us to combine several of the basic rules into a more elaborate calculation. Here is an example that illustrates how to solve problems that have several stages.

EXAMPLE 5.22

How many go pro?

What is the probability that a high school athlete will go on to professional sports? In the notation of Example 5.21, this is $P(B)$. To find $P(B)$ from the information in Example 5.21, use the tree diagram in Figure 5.9 to organize your thinking.

Each segment in the tree is one stage of the problem. Each complete branch shows a path that an athlete can take. The probability written on each segment is the conditional probability that an athlete follows that segment given that he has reached the point from which it branches. Starting at the left, high school athletes either do or do not compete in college. We know that the probability of competing in college is $P(A) = 0.05$, so the probability of not competing is $P(A^c) = 0.95$. These probabilities mark the leftmost branches in the tree.

Conditional on competing in college, the probability of playing professionally is $P(B | A) = 0.017$. So the conditional probability of *not* playing professionally is

$$P(B^c | A) = 1 - P(B | A) = 1 - 0.017 = 0.983$$

These conditional probabilities mark the paths branching out from A in Figure 5.9.

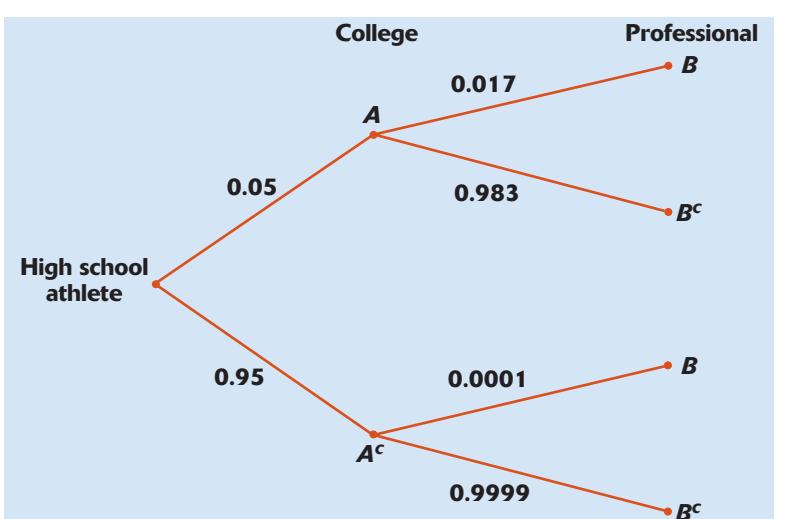


FIGURE 5.9 Tree diagram for Example 5.22.

The lower half of the tree diagram describes athletes who do not compete in college (A^c). It is unusual for these athletes to play professionally, but a few go straight from high school to professional leagues. Suppose that the conditional probability that a high school athlete reaches professional play given that he does not compete in college is $P(B | A^c) = 0.0001$. We can now mark the two paths branching from A^c in Figure 5.9.

There are two disjoint paths to B (professional play). By the addition rule, $P(B)$ is the sum of their probabilities. The probability of reaching B through college (top half of the tree) is

$$\begin{aligned} P(B \text{ and } A) &= P(A)P(B | A) \\ &= 0.05 \times 0.017 = 0.00085 \end{aligned}$$

The probability of reaching B without college is

$$\begin{aligned} P(B \text{ and } A^c) &= P(A^c)P(B | A^c) \\ &= 0.95 \times 0.0001 = 0.000095 \end{aligned}$$

The final result is

$$P(B) = 0.00085 + 0.000095 = 0.000945$$

About 9 high school athletes out of 10,000 will play professional sports.



Tree diagrams combine the addition and multiplication rules. The multiplication rule says that the probability of reaching the end of any complete branch is the product of the probabilities written on its segments. The probability of any outcome, such as the event B that an athlete reaches professional sports, is then found by adding the probabilities of all branches that are part of that event.

There is another kind of probability question that we might ask in the context of studies of athletes. Our earlier calculations look forward toward professional sports as the final stage of an athlete's career. Now let's concentrate on professional athletes and look back at their earlier careers.

EXAMPLE 5.23

Professional athletes' past

What proportion of professional athletes competed in college? In the notation of Example 5.21, this is the conditional probability $P(A | B)$. We start from the definition of conditional probability:

$$\begin{aligned} P(A | B) &= \frac{P(A \text{ and } B)}{P(B)} \\ &= \frac{0.00085}{0.000945} = 0.8995 \end{aligned}$$

Almost 90% of professional athletes competed in college.



We know the probabilities $P(A)$ and $P(A^c)$ that a high school athlete does and does not compete in college. We also know the conditional probabilities $P(B | A)$ and $P(B | A^c)$ that an athlete from each group reaches professional sports. Example 5.22 shows how to use this information to calculate $P(B)$. The method can be summarized in a single expression that adds the probabilities of the two paths to B in the tree diagram:

$$P(B) = P(A)P(B | A) + P(A^c)P(B | A^c)$$

In Example 5.23 we calculated the “reverse” conditional probability $P(A | B)$. The denominator 0.000945 in that example came from the expression just above. Put in this general notation, we have another probability law.

BAYES'S RULE

If A and B are any events whose probabilities are not 0 or 1,

$$P(A | B) = \frac{P(B | A)P(A)}{P(B | A)P(A) + P(B | A^c)P(A^c)}$$

Bayes's rule is named after Thomas Bayes, who wrestled with arguing from outcomes like B back to antecedents like A in a book published in 1763. It is far better to think your way through problems like Examples 5.22 and 5.23 rather than memorize these formal expressions.

APPLY YOUR KNOWLEDGE

- 5.62 **Where to manufacture?** Zipdrive, Inc., has developed a new disk drive for small computers. The demand for the new product is uncertain but can be described as “high” or “low” in any one year. After 4 years, the product is expected to be obsolete. Management must decide whether to build a plant or to contract with a factory in Hong Kong to manufacture the new drive. Building a plant will be profitable if demand remains high but could lead to a loss if demand drops in future years.

After careful study of the market and of all relevant costs, Zipdrive's planning office provides the following information. Let A be the event that the first year's demand is high, and B be the event that the following 3 years' demand is high. The planning office's best estimate of the probabilities is

$$P(A) = 0.9$$

$$P(B | A) = 0.36$$

$$P(B | A^c) = 0$$

The probability that building a plant is more profitable than contracting the production to Hong Kong is 0.95 if demand is high all 4 years, 0.3 if demand is high only in the first year, and 0.1 if demand is low all 4 years.

Draw a tree diagram that organizes this information. The tree will have three stages: first year's demand, next 3 years' demand, and whether building or contracting is more profitable. Which decision has the higher probability of being more profitable? (When probability analysis is used for investment decisions like this, firms usually compare the mean profits rather than the probability of a profit. We ignore this complication.)

- 5.63 **PDA screens.** A manufacturer of Personal Digital Assistants (PDAs) purchases screens from two different suppliers. The company receives 55% of its screens from Screensource and the remaining screens from Brightscreens. The quality of the screens varies between the suppliers: Screensource supplies 1% unsatisfactory screens while 4% of the screens from Brightscreens are unsatisfactory. Given that a randomly chosen screen is unsatisfactory, what is the probability it came from Brightscreens? (Hint: In the notation of this section, take A to be the event that the screen came from Brightscreens, and let B be the event that a randomly chosen screen is unsatisfactory.)

SECTION 5.4 SUMMARY

- The conditional probability $P(B | A)$ of an event B given an event A is defined by

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)}$$

when $P(A) > 0$. In practice, we most often find conditional probabilities from directly available information rather than from the definition.

- Any assignment of probability obeys the **general multiplication rule** $P(A \text{ and } B) = P(A)P(B | A)$. This rule is often used along with **tree diagrams** in calculating probabilities in settings with several stages.
- A and B are **independent** when $P(B | A) = P(B)$. The multiplication rule then becomes $P(A \text{ and } B) = P(A)P(B)$.
- When $P(A)$, $P(B | A)$, and $P(B | A^c)$ are known, **Bayes's rule** can be used to calculate $P(A | B)$ as follows:

$$P(A | B) = \frac{P(B | A)P(A)}{P(B | A)P(A) + P(B | A^c)P(A^c)}$$

SECTION 5.4 EXERCISES

- 5.64 **Income tax returns.** In 1999, the Internal Revenue Service received 127,075,145 individual tax returns. Of these, 9,534,653 reported an adjusted gross income of at least \$100,000, and 205,124 reported at least \$1 million.
- What is the probability that a randomly chosen individual tax return reports an income of at least \$100,000? At least \$1 million?
 - If you know that the return chosen shows an income of \$100,000 or more, what is the conditional probability that the income is at least \$1 million?
- 5.65 **Tastes in music.** Musical styles other than rock and pop are becoming more popular. A survey of college students finds that 40% like country music, 30% like gospel music, and 10% like both.
- What is the conditional probability that a student likes gospel music if we know that he or she likes country music?
 - What is the conditional probability that a student who does not like country music likes gospel music? (A Venn diagram may help you.)

- 5.66 **College degrees.** Exercise 5.60 (page 346) gives the counts (in thousands) of earned degrees in the United States in a recent year. Use these data to answer the following questions.
- What is the probability that a randomly chosen degree recipient is a man?
 - What is the conditional probability that the person chosen received a bachelor's degree, given that he is a man?
 - Use the multiplication rule to find the probability of choosing a male bachelor's degree recipient. Check your result by finding this probability directly from the table of counts.
- 5.67 **Geometric probability.** Choose a point at random in the square with sides $0 \leq x \leq 1$ and $0 \leq y \leq 1$. This means that the probability that the point falls in any region within the square is equal to the area of that region. Let X be the x coordinate and Y the y coordinate of the point chosen. Find the conditional probability $P(Y < 1/2 | Y > X)$. (Hint: Draw a diagram of the square and the events $Y < 1/2$ and $Y > X$.)
- 5.68 **Classifying occupations.** Exercise 4.105 (page 300) gives the probability distribution of the gender and occupation of a randomly chosen American worker. Use this distribution to answer the following questions.
- Given that the worker chosen holds a managerial (Class A) job, what is the conditional probability that the worker is female?
 - Classes D and E include most mechanical and factory jobs. What is the conditional probability that a worker is female, given that a worker holds a job in one of these classes?
 - Are gender and job type independent? How do you know?
- 5.69 **Preparing for the GMAT.** A company that offers courses to prepare would-be MBA students for the GMAT examination finds that 40% of its customers are currently undergraduate students and 60% are college graduates. After completing the course, 50% of the undergraduates and 70% of the graduates achieve scores of at least 600 on the GMAT. Use a tree diagram to organize this information.
- What percent of customers are undergraduates *and* score at least 600?
What percent of customers are graduates *and* score at least 600?
 - What percent of all customers score at least 600 on the GMAT?
- 5.70 **Telemarketing.** A telemarketing company calls telephone numbers chosen at random. It finds that 70% of calls are not completed (the party does not answer or refuses to talk), that 20% result in talking to a woman, and that 10% result in talking to a man. After that point, 30% of the women and 20% of the men actually buy something. What percent of calls result in a sale? (Draw a tree diagram.)
- 5.71 **Success on the GMAT.** In the setting of Exercise 5.69, what percent of the customers who score at least 600 on the GMAT are undergraduates? (Write this as a conditional probability.)
- 5.72 **Sales to women.** In the setting of Exercise 5.70, what percent of sales are made to women? (Write this as a conditional probability.)
- 5.73 **Credit card defaults.** The credit manager for a local department store discovers that 88% of all the store's credit card holders who defaulted on

their payments were late (by a week or more) with two or more of their monthly payments before failing to pay entirely (defaulting). This prompts the manager to suggest that future credit be denied to any customer who is late with two monthly payments. Further study shows that 3% of all credit customers default on their payments and 40% of those who have not defaulted have had at least two late monthly payments in the past.

- (a) What is the probability that a customer who has two or more late payments will default?
- (b) Under the credit manager's policy, in a group of 100 customers who have their future credit denied, how many would we expect *not* to default on their payments?
- (c) Does the credit manager's policy seem reasonable? Explain your response.

5.74 Successful bids. Consolidated Builders has bid on two large construction projects. The company president believes that the probability of winning the first contract (event A) is 0.6, that the probability of winning the second (event B) is 0.5, and that the probability of winning both jobs (event $\{A \text{ and } B\}$) is 0.3. What is the probability of the event $\{A \text{ or } B\}$ that Consolidated will win at least one of the jobs?

5.75 Independence? In the setting of the previous exercise, are events A and B independent? Do a calculation that proves your answer.

5.76 Successful bids, continued. Draw a Venn diagram that illustrates the relation between events A and B in Exercise 5.74. Write each of the following events in terms of A , B , A^c , and B^c . Indicate the events on your diagram and use the information in Exercise 5.74 to calculate the probability of each.

- (a) Consolidated wins both jobs.
- (b) Consolidated wins the first job but not the second.
- (c) Consolidated does not win the first job but does win the second.
- (d) Consolidated does not win either job.

5.77 Inspecting final products. Final products are sometimes selected to go through a complete inspection before leaving the production facility. Suppose that 8% of all products made at a particular facility fail to conform to specifications. Furthermore, 55% of all nonconforming items are selected for complete inspection while 20% of all conforming items are selected for complete inspection. Given that a randomly chosen item has gone through a complete inspection, what is the probability the item is nonconforming?

STATISTICS IN SUMMARY

This chapter concerns some further facts about probability that are useful in modeling but are not needed in our study of statistics. Section 5.1 discusses general rules that all probability models must obey, including the important multiplication rule for independent events. There are many specific probability models for specific situations. Section 5.2 uses the multiplication rule to obtain one of the most important probability models, the Binomial distribution for counts. Remember that not all counts have a Binomial distribution, just as not all measured variables have a Normal distribution.

In Section 5.3 we considered the Poisson distribution, an alternative model for counts. When events are not independent, we need the idea of conditional probability. That is the topic of Section 5.4. At this point, we finally reach the fully general form of the basic rules of probability. Here is a review list of the most important skills you should have acquired from your study of this chapter.

A. PROBABILITY RULES

1. Use Venn diagrams to picture relationships among several events.
2. Use the general addition rule to find probabilities that involve overlapping events.
3. Understand the idea of independence. Judge when it is reasonable to assume independence as part of a probability model.
4. Use the multiplication rule for independent events to find the probability that all of several independent events occur.
5. Use the multiplication rule for independent events in combination with other probability rules to find the probabilities of complex events.

B. BINOMIAL DISTRIBUTIONS

1. Recognize the Binomial setting: we have a fixed number n of independent success-failure trials with the same probability p of success on each trial.
2. Recognize and use the Binomial distribution of the count of successes in a Binomial setting.
3. (Optional.) Use the Binomial probability formula to find probabilities of events involving the count X of successes in a Binomial setting for small values of n .
4. Use Binomial tables to find Binomial probabilities.
5. Find the mean and standard deviation of a Binomial count X .
6. Recognize when you can use the Normal approximation to a Binomial distribution. Use the Normal approximation to calculate probabilities that concern a Binomial count X .

C. POISSON DISTRIBUTIONS

1. Recognize the Poisson setting: we are counting the number of successes in a fixed unit of measure (time, area, volume, or length).
2. Given a Poisson model with stated mean count per unit, find the Poisson distribution for the count in a multiple or a fractional number of units.
3. Use software to calculate Poisson probabilities.
4. Find the mean and standard deviation of a Poisson count X .
5. Use the central limit theorem to approximate Poisson probabilities when μ is too large for your software by dividing the basic unit of

measure into many smaller units of measure and viewing the Poisson random variable as the sum of many independent Poisson random variables.

D. CONDITIONAL PROBABILITY

1. Understand the idea of conditional probability. Identify the two events required from a verbal description of conditional probability. Find conditional probabilities for individuals chosen at random from a two-way table of counts of outcomes.
2. Use the general multiplication rule to find $P(A \text{ and } B)$ from $P(A)$ and the conditional probability $P(B | A)$.
3. Use a tree diagram to organize several-stage probability models.
4. Use Bayes's rule to calculate conditional probabilities when given other "reverse" conditional probabilities.

CHAPTER 5 REVIEW EXERCISES

- 5.78 **Playing the slots.** Slot machines are now video games, with winning determined by electronic random number generators. In the old days, slot machines worked like this: you pull the lever to spin three wheels; each wheel has 20 symbols, all equally likely to show when the wheel stops spinning; the three wheels are independent of each other. Suppose that the middle wheel has 9 bells among its 20 symbols, and the left and right wheels have 1 bell each.
- (a) You win the jackpot if all three wheels show bells. What is the probability of winning the jackpot?
 - (b) What is the probability that the wheels stop with exactly 2 bells showing?
- 5.79 **Leaking gas tanks.** Leakage from underground gasoline tanks at service stations can damage the environment. It is estimated that 25% of these tanks leak. You examine 15 tanks chosen at random, independently of each other.
- (a) What is the mean number of leaking tanks in such samples of 15?
 - (b) What is the probability that 10 or more of the 15 tanks leak?
 - (c) Now you do a larger study, examining a random sample of 1000 tanks nationally. What is the probability that at least 275 of these tanks are leaking?
- 5.80 **Environmental credits.** An opinion poll asks an SRS of 500 adults whether they favor tax credits for companies that demonstrate a commitment to preserving the environment. Suppose that in fact 45% of the population favor this idea. What is the probability that more than half of the sample are in favor?
- 5.81 **Computer training.** Macintosh users make up about 5% of all computer users. A computer training school that wants to attract Macintosh users mails an advertising flyer to 25,000 computer users.
- (a) If the mailing list can be considered a random sample of the population, what is the mean number of Macintosh users who will receive the flyer?
 - (b) What is the probability that at least 1245 Macintosh users will receive the flyer?

- 5.82 **Is this coin balanced?** While he was a prisoner of the Germans during World War II, John Kerrich tossed a coin 10,000 times. He got 5067 heads. Take Kerrich's tosses to be an SRS from the population of all possible tosses of his coin. If the coin is perfectly balanced, $p = 0.5$. Is there reason to think that Kerrich's coin gave too many heads to be balanced? To answer this question, find the probability that a balanced coin would give 5067 or more heads in 10,000 tosses. What do you conclude?
- 5.83 **Who is driving?** A sociology professor asks her class to observe cars having a man and a woman in the front seat and record which of the two is the driver.
- Explain why it is reasonable to use the Binomial distribution for the number of male drivers in n cars if all observations are made in the same location at the same time of day.
 - Explain why the Binomial model may not apply if half the observations are made outside a church on Sunday morning and half are made on campus after a dance.
 - The professor requires students to observe 10 cars during business hours in a retail district close to campus. Past observations have shown that the man is driving about 85% of cars in this location. What is the probability that the man is driving 8 or fewer of the 10 cars?
 - The class has 10 students, who will observe 100 cars in all. What is the probability that the man is driving 80 or fewer of these?

- 5.84 **Income and savings.** A sample survey chooses a sample of households and measures their annual income and their savings. Some events of interest are

A = the household chosen has income at least \$100,000

C = the household chosen has at least \$50,000 in savings

Based on this sample survey, we estimate that $P(A) = 0.13$ and $P(C) = 0.25$.

- We want to find the probability that a household either has income at least \$100,000 *or* savings at least \$50,000. Explain why we do not have enough information to find this probability. What additional information is needed?
- We want to find the probability that a household has income at least \$100,000 *and* savings at least \$50,000. Explain why we do not have enough information to find this probability. What additional information is needed?

- 5.85 **Medical risks.** You have torn a tendon and are facing surgery to repair it. The surgeon explains the risks to you: infection occurs in 3% of such operations, the repair fails in 14%, and both infection and failure occur together in 1%. What percent of these operations succeed and are free from infection?

Working. In the language of government statistics, you are “in the labor force” if you are available for work and either working or actively seeking work. The unemployment rate is the proportion of the labor force (not of the entire population) who are unemployed. Here are data from the Current Population Survey for the civilian population aged 25 years and over. The table entries are counts in thousands of people. Exercises 5.86 to 5.88 concern these data.

Highest education	Total population	In labor force	Employed
Did not finish high school	27,325	12,073	11,139
High school but no college	57,221	36,855	35,137
Less than bachelor's degree	45,471	33,331	31,975
College graduate	47,371	37,281	36,259

- 5.86 **Unemployment rates.** Find the unemployment rate for people with each level of education. How does the unemployment rate change with education? Explain carefully why your results show that level of education and being employed are not independent.
- 5.87 **Education and work.**
- What is the probability that a randomly chosen person 25 years of age or older is in the labor force?
 - If you know that the person chosen is a college graduate, what is the conditional probability that he or she is in the labor force?
 - Are the events “in the labor force” and “college graduate” independent? How do you know?
- 5.88 **Education and work, continued.** You know that a person is employed. What is the conditional probability that he or she is a college graduate? You know that a second person is a college graduate. What is the conditional probability that he or she is employed?
- 5.89 **Testing for HIV.** Enzyme immunoassay (EIA) tests are used to screen blood specimens for the presence of antibodies to HIV, the virus that causes AIDS. Antibodies indicate the presence of the virus. The test is quite accurate but is not always correct. Here are approximate probabilities of positive and negative EIA outcomes when the blood tested does and does not actually contain antibodies to HIV:⁹

	Test result	
	+	-
Antibodies present	0.9985	0.0015
Antibodies absent	0.006	0.994

Suppose that 1% of a large population carries antibodies to HIV in their blood.

- Draw a tree diagram for selecting a person from this population (outcomes: antibodies present or absent) and for testing his or her blood (outcomes: EIA positive or negative).
- What is the probability that the EIA is positive for a randomly chosen person from this population?
- What is the probability that a person has the antibody given that the EIA test is positive?

(This exercise illustrates a fact that is important when considering proposals for widespread testing for HIV, illegal drugs, or agents of biological warfare:

if the condition being tested is uncommon in the population, many positives will be false positives.)

- 5.90 **Testing for HIV, continued.** The previous exercise gives data on the results of EIA tests for the presence of antibodies to HIV. Repeat part (c) of this exercise for two different populations:
- Blood donors are prescreened for HIV risk factors, so perhaps only 0.1% (0.001) of this population carries HIV antibodies.
 - Clients of a drug rehab clinic are a high-risk group, so perhaps 10% of this population carries HIV antibodies.
 - What general lesson do your calculations illustrate?
- 5.91 **The Geometric distributions.** You are tossing a balanced die that has probability $1/6$ of coming up 1 on each toss. Tosses are independent. We are interested in how long we must wait to get the first 1.
- The probability of a 1 on the first toss is $1/6$. What is the probability that the first toss is not a 1 and the second toss is a 1?
 - What is the probability that the first two tosses are not 1s and the third toss is a 1? This is the probability that the first 1 occurs on the third toss.
 - Now you see the pattern. What is the probability that the first 1 occurs on the fourth toss? On the fifth toss? Give the general result: what is the probability that the first 1 occurs on the k th toss?

Geometric distribution

Comment: The distribution of the number of trials to the first success is called a **Geometric distribution**. In this problem you have found Geometric distribution probabilities when the probability of a success on each trial is $p = 1/6$. The same idea works for any p .

- 5.92 **Teenage drivers.** An insurance company has the following information about drivers aged 16 to 18 years: 20% are involved in accidents each year; 10% in this age group are A students; among those involved in an accident, 5% are A students.
- Let A be the event that a young driver is an A student and C the event that a young driver is involved in an accident this year. State the information given in terms of probabilities and conditional probabilities for the events A and C .
 - What is the probability that a randomly chosen young driver is an A student and is involved in an accident?
- 5.93 **Race and ethnicity.** The 2000 census allowed each person to choose from a long list of races. That is, in the eyes of the Census Bureau, you belong to whatever race you say you belong to. “Hispanic/Latino” is a separate category; Hispanics may be of any race. If we choose a resident of the United States at random, the 2000 census gives these probabilities:

	Hispanic	Not Hispanic
Asian	0.000	0.036
Black	0.003	0.121
White	0.060	0.691
Other	0.062	0.027

- (a) What is the probability that a randomly chosen person is white?
(b) You know that the person chosen is Hispanic. What is the conditional probability that this person is white?
- 5.94 **More on teenage drivers.** Use your work from Exercise 5.92 to find the percent of A students who are involved in accidents. (Start by expressing this as a conditional probability.)
- 5.95 **More on race and ethnicity.** Use the information in Exercise 5.93 to answer these questions.
(a) What is the probability that a randomly chosen American is Hispanic?
(b) You know that the person chosen is black. What is the conditional probability that this person is Hispanic?
- 5.96 **Screening job applicants.** A company retains a psychologist to assess whether job applicants are suited for assembly-line work. The psychologist classifies applicants as A (well suited), B (marginal), or C (not suited). The company is concerned about event D: an employee leaves the company within a year of being hired. Data on all people hired in the past five years gives these probabilities:

$$P(A) = 0.4$$

$$P(B) = 0.3$$

$$P(C) = 0.3$$

$$P(A \text{ and } D) = 0.1$$

$$P(B \text{ and } D) = 0.1$$

$$P(C \text{ and } D) = 0.2$$

Sketch a Venn diagram of the events A, B, C, and D and mark on your diagram the probabilities of all combinations of psychological assessment and leaving (or not) within a year. What is $P(D)$, the probability that an employee leaves within a year?

- 5.97 **Who buys iMacs?** The iMac computer was introduced by Apple Computer in the fall of 1998 and quickly became one of the company's best-selling products. The iMac was particularly aimed at first-time computer buyers. Approximately 5 months after the introduction of the iMac, Apple reported that 32% of iMac buyers were first-time computer buyers. At this same time, approximately 5% of all computer sales were of iMacs.¹⁰ Of buyers who did not purchase an iMac, approximately 40% were first-time computer buyers. Among first-time computer buyers during this time, what percent bought iMacs?
- 5.98 **Stealing software.** Employees sometimes install on their home computers software that was purchased by their employer for use on their work computers. For most commercial software packages, this is illegal. Suppose that 5% of all employees at a large corporation have illegally installed corporate software on their home computers knowing the act is illegal and an additional 2% have installed corporate software on their home computers not realizing that this is illegal. Of the 5% aware that the home installation is illegal, 80% will deny that they knew the act was illegal if confronted by a "software auditor." If an employee who has illegally installed software at home is confronted and denies knowing it was an illegal act, what is the probability that the employee knew the home installation was illegal?



CHAPTER 5 CASE STUDY EXERCISES

CASE STUDY 5.1: The Pentium FDIV bug. The Pentium FDIV bug was described in the Prelude to this chapter (page 306). The probability of one or more errors in 365 days is calculated using the multiplication rule as illustrated in Example 5.3 (page 312) and Example 5.4 (page 313). The formula can be expressed as

$$P(\text{one or more errors in 365 days}) = 1 - (1 - a)^{365 \times b}$$

where a is the $P(\text{error for a single division})$ and b is the assumed number of divisions per day for a typical user.

- A. **Intel's estimates.** Using Intel's estimates for a and b as described in the Prelude, calculate the probability of one or more errors in 365 days. You will need to use statistical or mathematical software to do this calculation. Verify the probability stated in the Prelude.
- B. **IBM's estimates.** Using IBM's estimates for a and b as described in the Prelude, calculate the probability of one or more errors in 365 days. You will need to use statistical or mathematical software to do this calculation. Verify the probability stated in the Prelude.

CASE STUDY 5.2: More on the Pentium FDIV bug.

- A. **More with IBM's estimates.** Using IBM's estimates for a and b as described in the Prelude, calculate the probability of one or more errors in 24 days. You will need to use statistical or mathematical software to do this calculation. In an IBM report dated December 1994, the authors of the report state that a typical user could make a mistake every 24 days. Do you agree with this statement? Explain your reasoning and supporting calculations.
- B. **Even more with IBM's estimates.** Using IBM's estimates for a and b as described in the Prelude, calculate the probability of one or more errors in a single day (round your answer to 5 decimal places). For 100,000 typical users, how many errors would you expect in a single day? In the IBM report dated December 1994, the authors of the report state that 100,000 Pentium users could expect 4000 errors to occur each day. Do you agree with this statement? Explain your reasoning and supporting calculations.

Prelude

Reacting to numbers

Ivan is the product manager in charge of blenders for a manufacturer of household appliances. Last month's sales report showed a 4% increase in blender sales to retailers, so Ivan treated the sales staff to lunch. This month, blender sales decline by 3%. Ivan is mystified—he can find no reason why sales should fall. Perhaps the sales force is getting lazy. Ivan expresses his disappointment, the sales staff are defensive, morale declines . . .

Then Caroline, a student intern who has taken a statistics class, asks about the source of the sales data. They are, she is told, estimates based on a sample of customer orders. Complete sales data aren't available until orders are filled and paid for. Caroline looks at the month-to-month variation in past estimates and at the relationship of the estimates to actual orders in the same month from all customers. Her report notes many reasons why the estimates will vary, including both normal month-to-month variation in blender orders and the additional variation due to using data on just a sample. Her report shows that changes of 3% or 4% up or down often occur simply because the data look only at a sample of orders. She even describes how large a change in the estimates provides good evidence that actual orders are really different from last month's total.

Ivan took every number as fixed and solid. He failed to grasp a statistical truism: we expect data to vary "just by chance."

Only variation larger than typical chance variation is good evidence of a real change.

Introduction to Inference

- 6.1 Estimating with Confidence**
- 6.2 Tests of Significance**
- 6.3 Using Significance Tests**
- 6.4 Power and Inference as a Decision**



Introduction

The purpose of statistical inference is to draw conclusions from data, conclusions that take into account the natural variability in the data. To do this, formal inference relies on probability to describe chance variation. We can then correct our “eyeball” judgment by calculation.

EXAMPLE 6.1

What are the probabilities?

Suppose we show a new TV commercial and the present commercial to 20 consumers each. Twelve of those who see the new ad declare an interest in buying the product, versus only 8 who watch the current version. Is the new commercial more effective? Perhaps, but a difference this large or larger between the results in the two groups would occur about one time in five simply because of chance variation. An effect that could so easily be just chance is not convincing.

In this chapter we introduce the two most prominent types of formal statistical inference. Section 6.1 concerns *confidence intervals* for estimating the value of a population parameter. Section 6.2 presents *tests of significance*, which assess the evidence for a claim. Both types of inference are based on the sampling distributions of statistics. That is, both report probabilities that state *what would happen if we used the inference method many times*. This kind of probability statement is characteristic of standard statistical inference. Users of statistics must understand the nature of the reasoning employed and the meaning of the probability statements that appear, for example, on computer output for statistical procedures.

Because the methods of formal inference are based on sampling distributions, they require a probability model for the data. Trustworthy probability models can arise in many ways, but the model is most secure and inference is most reliable when the data are produced by a properly randomized design. *When you use statistical inference you are acting as if the data come from a random sample or a randomized experiment.* If this is not true, your conclusions may be open to challenge. Do not be overly impressed by the complex details of formal inference. This elaborate machinery cannot remedy basic flaws in producing the data such as voluntary response samples and uncontrolled experiments. Use the common sense developed in your study of the first three chapters of this book, and proceed to detailed formal inference only when you are satisfied that the data deserve such analysis.

This chapter introduces the reasoning of statistical inference. We will discuss only a few specific techniques—for inference about the unknown mean μ of a population. Moreover, we will temporarily make an unrealistic assumption: that we know the standard deviation σ of the population. Later chapters will present inference methods for use in most of the settings we met in learning to explore data. There are libraries—both of books and of computer software—full of more elaborate statistical techniques. Informed use of any of these methods requires an understanding of the underlying

reasoning. A computer will do the arithmetic, but *you* must still exercise judgment based on understanding.

6.1 Estimating with Confidence

One way to characterize a collection of businesses is to determine the average of some measure of size. Total assets is one commonly used measure. If the collection of businesses is large, we generally take a sample and use the information gathered to make an inference about the entire collection. We use the term **population** to refer to the entire collection of interest.



CASE 6.1 COMMUNITY BANKS

Community banks are banks with less than a billion dollars of assets. There are approximately 7500 such banks in the United States. In many studies of the industry these banks are considered separately from banks that have more than a billion dollars of assets. The latter banks are called “large institutions.”

The Community Bankers Council of the American Bankers Association (ABA) conducts an annual survey of community banks.¹ For the 110 banks that make up the sample in a recent survey, the mean assets are $\bar{x} = 220$ (in millions of dollars). What can we say about μ , the mean assets of all community banks?



The sample mean \bar{x} is the natural estimator of the unknown population mean μ . We know that \bar{x} is an unbiased estimator of μ . More important, the law of large numbers says that the sample mean must approach the population mean as the size of the sample grows. The value $\bar{x} = 220$ therefore appears to be a reasonable estimate of the mean assets μ for all community banks. But how reliable is this estimate? A second sample would surely not give 220 again. Unbiasedness says only that there is no systematic tendency to underestimate or overestimate the truth. Could we plausibly get a sample mean of 250 or 200 on repeated samples? An estimate without an indication of its variability is of limited value.

Statistical confidence

Just as unbiasedness of an estimator concerns the center of its sampling distribution, questions about variation are answered by looking at the spread. The central limit theorem tells us that if the entire population of community bank assets has mean μ and standard deviation σ , then in repeated samples of size 110 the sample mean \bar{x} approximately follows the $N(\mu, \sigma/\sqrt{110})$ distribution. Suppose that the true standard deviation σ is equal to the sample standard deviation $s = 161$. This is not realistic,

although it will give reasonably accurate results for samples as large as 110. In the next chapter we will learn how to proceed when σ is not known. But for now, we are more interested in statistical reasoning than in such details of our methods. In repeated sampling the sample mean \bar{x} is approximately Normal, centered at the unknown population mean μ , with standard deviation

$$\sigma_{\bar{x}} = \frac{161}{\sqrt{110}} = 15 \text{ millions of dollars}$$

Now we can talk about estimating μ . Consider this line of thought, which is illustrated by Figure 6.1:

- The 68–95–99.7 rule says that the probability is about 0.95 that \bar{x} is within 30 (two standard deviations of \bar{x}) of the population mean assets μ .
- To say that \bar{x} lies within 30 of μ is the same as saying that μ is within 30 of \bar{x} .
- So 95% of all samples will capture the true μ in the interval from $\bar{x} - 30$ to $\bar{x} + 30$.

We have simply restated a fact about the sampling distribution of \bar{x} . *The language of statistical inference uses this fact about what would happen in the long run to express our confidence in the results of any one sample.* Our sample gave $\bar{x} = 220$. We say that we are 95% confident that the unknown mean assets for all community banks lie between

$$\bar{x} - 30 = 220 - 30 = 190$$

and

$$\bar{x} + 30 = 220 + 30 = 250$$

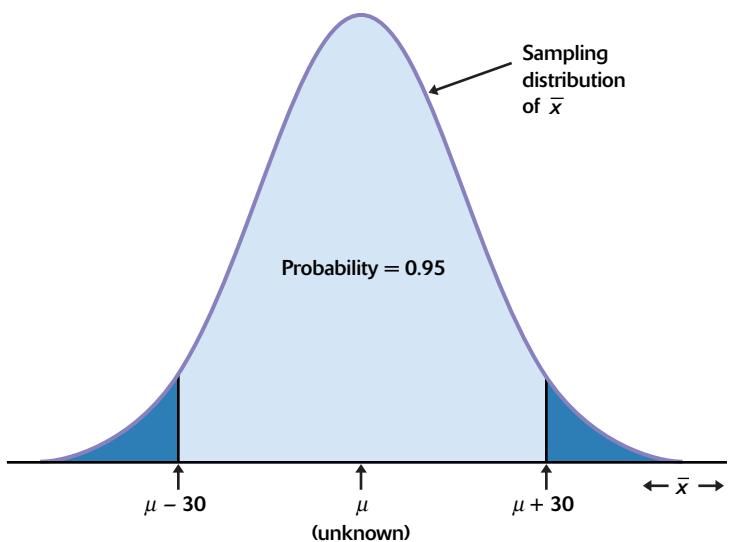


FIGURE 6.1 In 95% of all samples, \bar{x} lies within ± 30 of μ . So μ also lies within ± 30 of \bar{x} in those samples.

Be sure that you understand the grounds for our confidence. There are only two possibilities:

1. The interval between 190 and 250 contains the true μ .
2. Our SRS was one of the few samples for which \bar{x} is not within 30 of the true μ . Only 5% of all samples give such inaccurate results.

We cannot know whether our sample is one of the 95% for which the interval $\bar{x} \pm 30$ catches μ or one of the unlucky 5%. The statement that we are 95% confident that the unknown μ lies between 190 and 250 is shorthand for saying, “We arrived at these numbers by a process that gives correct results 95% of the time.”

APPLY YOUR KNOWLEDGE

- 6.1 **Company invoices.** The mean amount μ for all of the invoices for your company last month is not known. Based on your past experience, you are willing to assume that the standard deviation of invoice amounts is about \$200. If you take a random sample of 100 invoices, what is the value of the standard deviation for \bar{x} ?
- 6.2 In the setting of the previous exercise, the 68–95–99.7 rule says that the probability is about 0.95 that \bar{x} is within _____ of the population mean μ . Fill in the blank.
- 6.3 In the setting of the previous two exercises, about 95% of all samples will capture the true mean of all of the invoices in the interval \bar{x} plus or minus _____. Fill in the blank.

Confidence intervals

The interval of numbers between the values $\bar{x} \pm 30$ is called a *95% confidence interval* for μ . Like most confidence intervals we will meet, this one has the form

$$\text{estimate} \pm \text{margin of error}$$

margin of error

The estimate (\bar{x} in this case) is our guess for the value of the unknown parameter. The **margin of error** 30 shows how precise we believe our guess is, based on the variability of the estimate. This is a 95% confidence interval because it catches the unknown μ in 95% of all possible samples.

CONFIDENCE INTERVAL

- A level C confidence interval for a parameter has two parts:
- An interval calculated from the data, usually of the form
- $$\text{estimate} \pm \text{margin of error}$$
- A confidence level C, which gives the probability that the interval will capture the true parameter value in repeated samples.

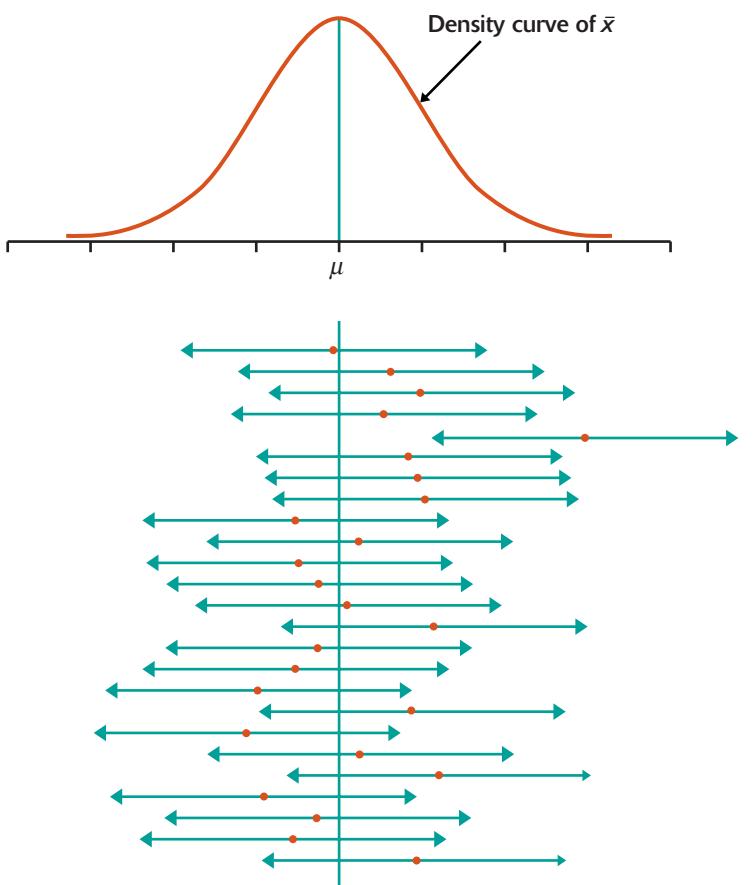


FIGURE 6.2 Twenty-five samples from the same population gave these 95% confidence intervals. In the long run, 95% of all samples give an interval that covers μ .

Figure 6.2 illustrates the behavior of 95% confidence intervals in repeated sampling. The center of each interval is at \bar{x} and therefore varies from sample to sample. The sampling distribution of \bar{x} appears at the top of the figure. The 95% confidence intervals $\bar{x} \pm 3\sigma$ from 25 SRSs appear below. The center \bar{x} of each interval is marked by a dot. The arrows on either side of the dot span the confidence interval. All except one of the 25 intervals cover the true value of μ . In a very large number of samples, 95% of the confidence intervals would contain μ . You can choose the confidence level. Common practice is to choose 95%, but 90% and 99% are also popular.

The *Confidence Intervals* applet at www.whfreeman.com/pbs animates Figure 6.2 and allows you to choose among several levels of confidence. This interactive applet is an excellent way to grasp the idea of a confidence interval.



APPLY YOUR KNOWLEDGE

- 6.4 80% confidence intervals. The idea of an 80% confidence interval is that the interval captures the true parameter value in 80% of all samples. That's not high enough confidence for practical use, but 80% hits and 20% misses



make it easy to see how a confidence interval behaves in repeated samples from the same population.

- (a) Set the confidence level in the *Confidence Intervals* applet to 80%. Click “Sample” to choose an SRS and calculate the confidence interval. Do this 10 times to simulate 10 SRSs with their 10 confidence intervals. How many of the 10 intervals captured the true mean μ ? How many missed?
- (b) You see that we can’t predict whether the next sample will hit or miss. The confidence level, however, tells us what percent will hit in the long run. Reset the applet and click “Sample 50” to get the confidence intervals from 50 SRSs. How many hit? Keep clicking “Sample 50” and record the percent of hits among 100, 200, 300, 400, 500, 600, 700, 800, and 1000 SRSs. Even 1000 samples is not truly “the long run,” but we expect the percent of hits in 1000 samples to be fairly close to the confidence level, 80%.

Confidence interval for a population mean

We use the sampling distribution of the sample mean \bar{x} to construct a level C confidence interval for the mean μ of a population. We assume that the data are an SRS of size n . The sampling distribution is exactly $N(\mu, \sigma/\sqrt{n})$ when the population has the $N(\mu, \sigma)$ distribution. The central limit theorem says that this same sampling distribution is approximately correct for large samples whenever the population mean and standard deviation are μ and σ .

Our construction of a 95% confidence interval for the mean assets of community banks began by noting that any Normal distribution has probability about 0.95 within ± 2 standard deviations of its mean. To construct a level C confidence interval, we first catch the central C area under a Normal curve. Because all Normal distributions are the same in the standard scale, we can obtain everything we need from the standard Normal curve. Figure 6.3 shows the relationship between the central area C and the

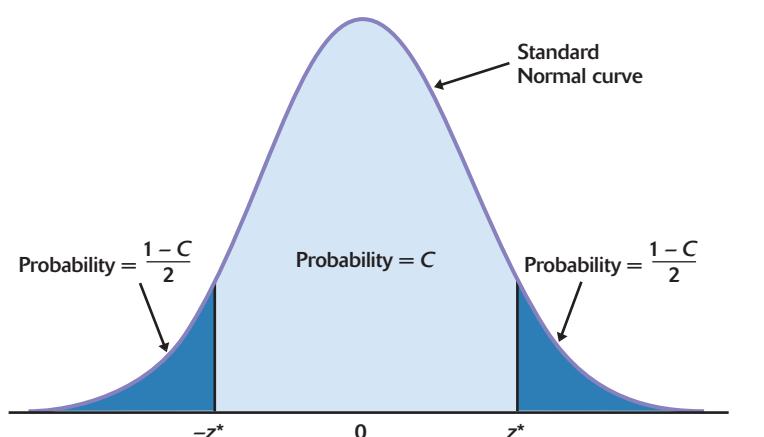


FIGURE 6.3 The area between the critical values $-z^*$ and z^* under the standard Normal curve is C .

points z^* that mark off this area. Values of z^* for many choices of C appear in the row labeled z^* in Table D at the back of the book. Here are the most important entries from that part of the table:

z^*	1.645	1.960	2.576
C	90%	95%	99%

critical value

Values z^* that mark off specified areas are called **critical values** of the standard Normal distribution. Notice that for $C = 95\%$ the table gives $z^* = 1.960$. This is slightly more precise than the value $z^* = 2$ based on the 68–95–99.7 rule.

Any Normal curve has probability C between the point z^* standard deviations below the mean and the point z^* above the mean, as Figure 6.3 reminds us. The sample mean \bar{x} has the Normal distribution with mean μ and standard deviation σ/\sqrt{n} . So there is probability C that \bar{x} lies between

$$\mu - z^* \frac{\sigma}{\sqrt{n}} \text{ and } \mu + z^* \frac{\sigma}{\sqrt{n}}$$

This is exactly the same as saying that the unknown population mean μ lies between

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} \text{ and } \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

So, the probability that the interval

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

contains μ is C . This is our confidence interval. The estimate of the unknown μ is \bar{x} , and the margin of error is $z^* \sigma/\sqrt{n}$.

CONFIDENCE INTERVAL FOR A POPULATION MEAN

Choose an SRS of size n from a population having unknown mean μ and known standard deviation σ . A level C confidence interval for μ is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

Here z^* is the **critical value** with area C between $-z^*$ and z^* under the standard Normal curve. The quantity

$$z^* \sigma/\sqrt{n}$$

is the **margin of error**. The interval is exact when the population distribution is Normal and is approximately correct when n is large in other cases.

EXAMPLE 6.2**Banks' loan-to-deposit ratio**

The ABA survey of community banks also asked about the loan-to-deposit ratio (LTDR), a bank's total loans as a percent of its total deposits. The mean LTDR for the 110 banks in the sample is $\bar{x} = 76.7$ and the standard deviation is $s = 12.3$. This sample is sufficiently large for us to use s as the population σ here. Give a 95% confidence interval for the mean LTDR for community banks.

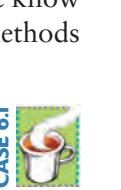
For 95% confidence, we see from Table D that $z^* = 1.96$. The margin of error is

$$z^* \frac{\sigma}{\sqrt{n}} = 1.96 \frac{12.3}{\sqrt{110}} = 2.3$$

A 95% confidence interval for μ is therefore

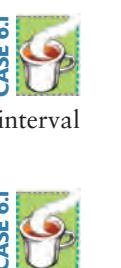
$$\begin{aligned}\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} &= 76.7 \pm 2.3 \\ &= (74.4, 79.0)\end{aligned}$$

We are 95% confident that the mean LTDR is between 74.4% and 79.0%.



In this example we used an estimate for σ based on a large sample. Sometimes we may know σ quite accurately from past experience with similar data. When we have few data, past experience may be better than estimation from our sample. In general, however, we can't act as if we know the population standard deviation σ . The next chapter presents methods that don't assume that we know σ .

- APPLY YOUR KNOWLEDGE**
- 6.5 **Bank assets.** The 110 banks in the ABA survey had mean assets of 220 million dollars. The standard deviation of their assets was 161. Assume that the sample standard deviation can be used in place of the population standard deviation. Give a 99% confidence interval for μ , the mean assets for all community banks.
- 6.6 In the setting of the previous exercise, would the margin of error for 90% confidence be larger or smaller? Verify your answer by performing the calculations.

**How confidence intervals behave**

The margin of error $z^* \sigma / \sqrt{n}$ for estimating the mean of a Normal population illustrates several important properties that are shared by all confidence intervals in common use. In particular, a higher confidence level increases z^* and so increases the margin of error for intervals based on the same data. We would like high confidence and a small margin of error, but improving one degrades the other. If a confidence interval's margin of error is too large, there are only three ways to reduce it:

- Use a lower level of confidence (smaller C , hence smaller z^*).
- Reduce σ .
- Increase the sample size (larger n).

The common choices of confidence level are 99%, 95%, and 90%. The critical values z^* for these levels are 2.576, 1.960, and 1.645, decreasing as the confidence level drops. Figure 6.3 makes it clear why z^* is smaller for lower confidence (smaller C). If n and σ are unchanged, settling for lower confidence will reduce the margin of error.

The standard deviation σ measures variation in the population. Think of the variation among individuals in the population as noise that obscures the average value μ . It is harder to pin down the mean μ of a highly variable population. This is why the margin of error of a confidence interval increases with σ . Sometimes we can reduce σ by carefully controlling the measurement process or by restricting our attention to only part of a large population.

Increasing the sample size n also reduces the margin of error. Suppose we want to cut the margin of error in half. The square root in the formula implies that we must have four times as many observations, not just twice as many.

EXAMPLE 6.3



How confidence intervals behave

Suppose there were only 25 banks in the survey of community banks, and that \bar{x} and σ are unchanged from Example 6.2. The margin of error increases from 2.3 to

$$z^* \frac{\sigma}{\sqrt{n}} = 1.96 \frac{12.3}{\sqrt{25}} = 4.8$$

A 95% confidence interval for μ is therefore

$$\begin{aligned} \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} &= 76.7 \pm 4.8 \\ &= (71.9, 81.5) \end{aligned}$$

Figure 6.4 illustrates how the margin of error increases for the smaller sample.

Suppose next that we demand 99% confidence for the mean LTDR in Example 6.2, rather than 95%. Table D tells us that for 99% confidence, $z^* = 2.576$. The margin of error increases from 2.3 to

$$\begin{aligned} z^* \frac{\sigma}{\sqrt{n}} &= 2.576 \frac{12.3}{\sqrt{110}} \\ &= 3.0 \end{aligned}$$

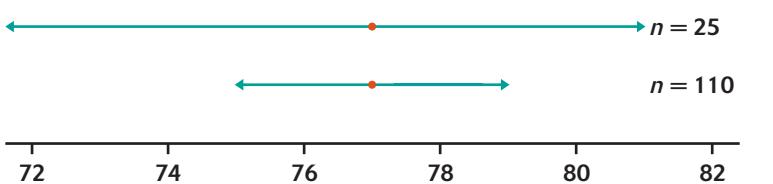


FIGURE 6.4 Confidence intervals for $n = 110$ and $n = 25$, for Examples 6.2 and 6.3.

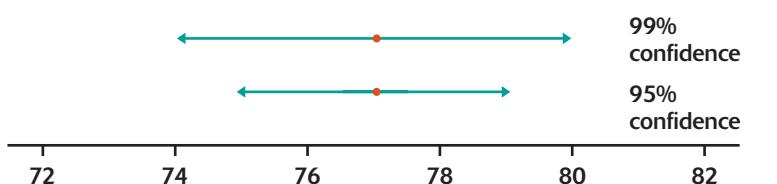


FIGURE 6.5 Confidence intervals for 95% and 99% confidence from the same data, for Example 6.3.

The 99% confidence interval is therefore

$$\begin{aligned}\bar{x} \pm \text{margin of error} &= 76.7 \pm 3.0 \\ &= (73.7, 79.7)\end{aligned}$$

Requiring 99% confidence rather than 95% confidence has increased the margin of error from 2.3 to 3.0. Figure 6.5 compares the two intervals.



Choosing the sample size

A wise user of statistics never plans data collection without at the same time planning the inference. You can arrange to have both high confidence and a small margin of error. The margin of error of the confidence interval $\bar{x} \pm z^* \sigma / \sqrt{n}$ for a Normal mean is

$$z^* \frac{\sigma}{\sqrt{n}}$$

To obtain a desired margin of error m , just set this expression equal to m , substitute the critical value z^* for your desired confidence level, and solve for the sample size n . Here is the result.

SAMPLE SIZE FOR DESIRED MARGIN OF ERROR

The confidence interval for a population mean will have a specified margin of error m when the sample size is

$$n = \left(\frac{z^* \sigma}{m} \right)^2$$

In practice, observations cost time and money. The sample size you calculate from this formula may turn out to be too expensive. Note that it is the *sample* size and not the population size that determines the margin of error. The size of the *population* (as long as it is much larger than the sample) does not influence the margin of error.

EXAMPLE 6.4**How many banks should we survey?**

In Example 6.2 we found that the margin of error was 2.3 for estimating the mean LTDR of community banks with $n = 110$ and 95% confidence. We are willing to settle for a margin of error of 3.0 when we do the survey next year. How many banks should we survey?

For 95% confidence, Table D gives $z^* = 1.960$. We will assume that the standard deviation next year will be approximately the same as the $\sigma = 12.3$ that we obtained this year. For margin of error of 3.0 we have

$$n = \left(\frac{z^* \sigma}{m} \right)^2 = \left(\frac{1.96 \times 12.3}{3.0} \right)^2 = 64.6$$

Because 64 observations will give a slightly wider interval than desired and 65 observations a slightly narrower interval, we will survey 65 banks. With this choice, our estimate of mean assets will be within 3.0 of the true value with 95% confidence.



Always round *up* to the next higher whole number when using the formula for the sample size n . In practice we often calculate the margins of error corresponding to a range of values of n . We then decide what margin of error we can afford.

APPLY YOUR KNOWLEDGE

- 6.7 Starting salaries. You are planning a survey of starting salaries for recent business major graduates from your college. From a pilot study you estimate that the standard deviation is about \$8000. What sample size do you need to have a margin of error equal to \$500 with 95% confidence?
- 6.8 Suppose that in the setting of the previous exercise you are willing to settle for a margin of error of \$1000. Will the required sample size be larger or smaller? Verify your answer by performing the calculations.

Some cautions

We have already seen that small margins of error and high confidence can require large numbers of observations. You should also be keenly aware that *any formula for inference is correct only in specific circumstances*. If the government required statistical procedures to carry warning labels like those on drugs, most inference methods would have long labels indeed. Our formula $\bar{x} \pm z^* \sigma / \sqrt{n}$ for estimating a Normal mean comes with the following list of warnings for the user:

- The data must be an SRS from the population. We are completely safe if we actually did a randomization and drew an SRS. The ABA survey that lies behind Examples 6.1 to 6.4 is based on a random sample of banks. We are not in great danger if the data can plausibly be thought of as independent observations from a population.

- The formula is not correct for probability sampling designs more complex than an SRS. Correct methods for other designs are available. We will not discuss confidence intervals based on multistage or stratified samples. If you plan such samples, be sure that you (or your statistical consultant) know how to carry out the inference you desire.
- There is no correct method for inference from data haphazardly collected with bias of unknown size. Fancy formulas cannot rescue badly produced data.
- Because \bar{x} is not resistant, outliers can have a large effect on the confidence interval. You should search for outliers and try to correct them or justify their removal before computing the interval. If the outliers cannot be removed, ask your statistical consultant about procedures that are not sensitive to outliers.
- If the sample size is small and the population is not Normal, the true confidence level will be different from the value C used in computing the interval. Examine your data carefully for skewness and other signs of non-Normality. The interval relies only on the distribution of \bar{x} , which even for quite small sample sizes is much closer to Normal than that of the individual observations. When $n \geq 15$, the confidence level is not greatly disturbed by non-Normal populations unless extreme outliers or quite strong skewness is present. We will discuss this issue in more detail in the next chapter.
- You must know the standard deviation σ of the population. This unrealistic requirement renders the interval $\bar{x} \pm z^* \sigma / \sqrt{n}$ of little use in statistical practice. We will learn in the next chapter what to do when σ is unknown. If, however, the sample is large, the sample standard deviation s will be close to the unknown σ . The formula $\bar{x} \pm z^* s / \sqrt{n}$ is then an approximate confidence interval for μ .

The most important caution concerning confidence intervals is a consequence of the first of these warnings. *The margin of error in a confidence interval covers only random sampling errors.* The margin of error is obtained from the sampling distribution and indicates how much error can be expected because of chance variation in randomized data production. Practical difficulties such as undercoverage and nonresponse in a sample survey cause additional errors. These errors can be larger than the random sampling error, particularly when the sample size is large. Remember this unpleasant fact when reading the results of an opinion poll. The practical conduct of a survey influences the trustworthiness of its results in ways that are not included in the announced margin of error.

Every inference procedure that we will meet has its own list of warnings. Because many of the warnings are similar to those above, we will not print the full warning label each time. If we use the mathematics of probability, it is easy to state conditions under which a method of inference is exactly

correct. These conditions are *never* fully met in practice. For example, no population is exactly Normal. Deciding when a statistical procedure should be used in practice often requires judgment assisted by exploratory analysis of the data. Mathematical facts are therefore only a part of statistics. The difference between statistics and mathematics can be stated thus: mathematical theorems are true; statistical methods are often effective when used with skill.

Finally, you should understand what statistical confidence does not say. We are 95% confident that the mean assets of community banks in Case 6.1 lie between \$190 million and \$250 million. This says that these numbers were calculated by a method that gives correct results in 95% of all possible samples. It does *not* say that the probability is 95% that the true mean falls between 190 and 250. No randomness remains after we draw a particular sample and get from it a particular interval. The true mean either is or is not between 190 and 250. Probability in its interpretation as a description of random behavior makes no sense in this situation. The probability calculations of standard statistical inference describe how often the *method* gives correct answers.

APPLY YOUR KNOWLEDGE

- 6.9 Internet users. A survey of users of the Internet found that males outnumbered females by nearly 2 to 1. This was a surprise, because earlier surveys had put the ratio of men to women closer to 9 to 1. Later in the article we find this information:

Detailed surveys were sent to more than 13,000 organizations on the Internet; 1,468 usable responses were received. According to Mr. Quartermann, the margin of error is 2.8 percent, with a confidence level of 95 percent.²

- What was the *response rate* for this survey? (The response rate is the percent of the planned sample that responded.)
- Do you think that the small margin of error is a good measure of the accuracy of the survey's results? Explain your answer.

BEYOND THE BASICS: THE BOOTSTRAP

Confidence intervals are based on sampling distributions. The interval $\bar{x} \pm z^* s / \sqrt{n}$ follows from the fact that the sampling distribution of \bar{x} is $N(\mu, \sigma / \sqrt{n})$ when the data are an SRS from a $N(\mu, \sigma)$ population. The central limit theorem says that this sampling distribution is also approximately right for large samples from non-Normal populations.

What if the population is clearly non-Normal and we have only a small sample? Then we do not know what the sampling distribution of \bar{x} looks like. The **bootstrap** is a procedure for approximating sampling distributions when theory cannot tell us their shape.³

The basic idea is to act as if our sample were the population. We take many samples from it. Each of these is called a **resample**. For each resample, we calculate the mean \bar{x} . We get different results from different resamples because we sample *with replacement*. That is, an observation in the original sample can appear more than once in a resample.

bootstrap

resample

The community bank assets data described in Case 6.1 are somewhat skewed toward high values. The mean \$220 million is larger than the median, \$190 million. There are two large banks with assets of \$908 and \$804 million that could be considered outliers. Because we have no reason to suspect that these banks are qualitatively different from the other community banks in the sample, we are reluctant to discard them. Rather, we will reexamine our inference with a method that is not sensitive to the Normality assumption.

If we had 1000 SRSs from the population of banks, the distribution of the 1000 values of \bar{x} would be close to the sampling distribution of \bar{x} . We have only one sample. The bootstrap idea says: take 1000 resamples of size 110 (with replacement) from our one sample of 110 banks. For each resample, compute the mean. Treat the distribution of \bar{x} 's from the 1000 resamples as if it were the sampling distribution. Because the sample is like the population, taking resamples from the sample is similar to taking samples from the population.

The bootstrap is practical only when you can use a computer to take 1000 or more resamples quickly. It is an example of how fast and cheap computing has changed the way we do statistics.

EXAMPLE 6.5



Check the interval with the bootstrap

In the discussion of Case 6.1 we found a 95% confidence interval for the mean assets of community banks using the traditional statistical procedure based on the middle 95% of a Normal distribution. The interval was (190, 250). Let's check to see if we made a serious error. The middle 95% of the 1000 \bar{x} 's from 1000 resamples runs from 193 to 249. Try it again with 1000 new resamples: we get the interval (191, 252). The two bootstrap estimates are close to each other and also close to the standard interval that assumes Normality. The standard interval is reasonably accurate for these data.



If the bootstrap method applied to this problem gave very different results, we would need to reconsider the analysis. Obtaining similar results from different methods gives us confidence that the answers we seek are coming from the data, not from the particular method that we used.

SECTION 6.1 SUMMARY

- The purpose of a **confidence interval** is to estimate an unknown parameter with an indication of how accurate the estimate is and of how confident we are that the result is correct.
- Any confidence interval has two parts: an interval computed from the data and a confidence level. The interval often has the form

$$\text{estimate} \pm \text{margin of error}$$

- The **confidence level** states the probability that the method will give a correct answer. That is, if you use 95% confidence intervals often, in the long run 95% of your intervals will contain the true parameter value. When you apply the method once, you do not know if your interval gave a correct value (this happens 95% of the time) or not (this happens 5% of the time).
- A level C confidence interval for the mean μ of a Normal population with known standard deviation σ , based on an SRS of size n , is given by

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

- Here z^* is a **critical value** obtained from the bottom row in Table D. The probability is C that a standard Normal random variable takes a value between $-z^*$ and z^* .
- The **margin of error** in the confidence interval for μ is

$$z^* \frac{\sigma}{\sqrt{n}}$$

- Other things being equal, the margin of error of a confidence interval decreases as
 - the confidence level C decreases,
 - the population standard deviation σ decreases, or
 - the sample size n increases.
- The sample size required to obtain a confidence interval of specified margin of error m for a Normal mean is

$$n = \left(\frac{z^* \sigma}{m} \right)^2$$

where z^* is the critical value for the desired level of confidence.

- A specific confidence interval formula is correct only under specific conditions. The most important conditions concern the method used to produce the data. Other factors such as the form of the population distribution may also be important.

SECTION 6.1 EXERCISES

- 6.10 Apartment rental rates. You want to rent an unfurnished one-bedroom apartment for next semester. The mean monthly rent for a random sample of 10 apartments advertised in the local newspaper is \$540. Assume that the standard deviation is \$80. Find a 95% confidence interval for the mean

monthly rent for unfurnished one-bedroom apartments available for rent in this community.

- 6.11 **Study habits.** A questionnaire about study habits was given to a random sample of students taking a large introductory statistics class. The sample of 25 students reported that they spent an average of 110 minutes per week studying statistics. Assume that the standard deviation is 40 minutes.
- Give a 95% confidence interval for the mean time spent studying statistics by students in this class.
 - Is it true that 95% of the students in the class have weekly study times that lie in the interval you found in part (a)? Explain your answer.
- 6.12 **Clothing for runners.** Your company sells exercise clothing and equipment on the Internet. To design the clothing, you collect data on the physical characteristics of your different types of customers. Here are the weights (in kilograms) for a sample of 24 male runners. Assume that these runners can be viewed as a random sample of your potential male customers. Suppose also that the standard deviation of the population is known to be $\sigma = 4.5$ kg.
- 67.8 61.9 63.0 53.1 62.3 59.7 55.4 58.9 60.9 69.2 63.7 68.3
64.7 65.6 56.0 57.8 66.0 62.9 53.6 65.0 55.8 60.4 69.3 61.7
- What is $\sigma_{\bar{x}}$, the standard deviation of \bar{x} ?
 - Give a 95% confidence interval for μ , the mean of the population from which the sample is drawn.
 - Will the interval contain the weights of approximately 95% of similar runners? Explain your answer.
- 6.13 **Pounds versus kilograms.** Suppose that the weights of the runners in Exercise 6.12 were recorded in pounds rather than kilograms. Use your answers to Exercise 6.12, and the fact that 1 kilogram equals 2.2 pounds, to answer these questions.
- What is the mean weight of these runners?
 - What is the standard deviation of the mean weight?
 - Give a 95% confidence interval for the mean weight of the population of runners that these runners represent.
- 6.14 **99% versus 95% confidence interval.** Find a 99% confidence interval for the mean weight μ of the population of male runners in Exercise 6.12. Is the 99% confidence interval wider or narrower than the 95% interval found in Exercise 6.12? Explain in plain language why this is true.
- 6.15 **Hotel managers.** A study of the career paths of hotel general managers sent questionnaires to an SRS of 160 hotels belonging to major U.S. hotel chains. There were 114 responses. The average time these 114 general managers had spent with their current company was 11.8 years. Give a 99% confidence interval for the mean number of years general managers of major-chain hotels have spent with their current company. (Take it as known that the standard deviation of time with the company for all general managers is 3.2 years.)

- 6.16 **Supermarket shoppers.** A marketing consultant observed 50 consecutive shoppers at a supermarket. One variable of interest was how much each shopper spent in the store. Here are the data (in dollars), arranged in increasing order:

3.11	8.88	9.26	10.81	12.69	13.78	15.23	15.62
17.00	17.39	18.36	18.43	19.27	19.50	19.54	20.16
20.59	22.22	23.04	24.47	24.58	25.13	26.24	26.26
27.65	28.06	28.08	28.38	32.03	34.98	36.37	38.64
39.16	41.02	42.97	44.08	44.67	45.40	46.69	48.65
50.39	52.75	54.80	59.07	61.22	70.32	82.70	85.76
86.37	93.34						

Assume that the standard deviation is \$22.00. Find a 95% confidence interval for the mean amount spent by shoppers in similar circumstances.

- 6.17 **Bank workers.** We examined the wages of a random sample of National Bank black female hourly workers in Example 1.9 on page 32. Here are the data (in dollars per year):

16015	17516	17274	16555	20788
19312	17124	18405	19090	12641
17813	18206	19338	15953	16904

Find a 95% confidence interval for the mean earnings of all black female hourly workers at National Bank. Use \$1900 for the standard deviation.

- 6.18 **Hotel managers.** How large a sample of the hotel managers in Exercise 6.15 would be needed to estimate the mean μ within ± 1 year with 99% confidence?
- 6.19 **Supermarket shoppers.** Suppose that you want to perform a study similar to the survey of supermarket shoppers described in Exercise 6.16. If you want the margin of error to be about \$4.00, how many shoppers would you need in your sample? (Use \$22.00 for the standard deviation in your calculations.)
- 6.20 **Bank employees.** In Exercise 6.17 we examined the earnings of a random sample of 15 black women who were hourly workers at National Bank. If we wanted to estimate the mean with an interval of the form $\bar{x} \pm 500$, how many observations are required?
- 6.21 **Business mergers.** A Gallup Poll asked 1004 adults about mergers between companies. One question was “Do you think the result is usually good for the economy or bad for the economy?” Forty-three percent of the sample thought mergers were good for the economy. The Gallup press release added:

For results based on this sample, one can say with 95 percent confidence that the maximum error attributable to sampling and other random effects is plus or minus 3 percentage points. In addition to sampling error, question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of public opinion polls.⁴

The Gallup Poll uses a complex multistage sample design, but the sample percent has approximately a Normal sampling distribution.

- (a) The announced poll result was $43\% \pm 3\%$. Can we be certain that the true population percent falls in this interval?
- (b) Explain to someone who knows no statistics what the announced result $43\% \pm 3\%$ means.
- (c) This confidence interval has the same form we have met earlier:

$$\text{estimate} \pm z^* \sigma_{\text{estimate}}$$

What is the standard deviation σ_{estimate} of the estimated percent?

- (d) Does the announced margin of error include errors due to practical problems such as undercoverage and nonresponse?
- 6.22 **Median household income.** When the statistic that estimates an unknown parameter has a Normal distribution, a confidence interval for the parameter has the form

$$\text{estimate} \pm z^* \sigma_{\text{estimate}}$$

In a complex sample survey design, the appropriate unbiased estimate of the population mean and the standard deviation of this estimate may require elaborate computations. But when the estimate is known to have a Normal distribution and its standard deviation is given, we can calculate a confidence interval for μ from complex sample designs without knowing the formulas that led to the numbers given.

A report based on the Current Population Survey estimates the 1999 median annual earnings of households as \$40,816 and also estimates that the standard deviation of this estimate is \$191. The Current Population Survey uses an elaborate multistage sampling design to select a sample of about 50,000 households. The sampling distribution of the estimated median income is approximately Normal. Give a 95% confidence interval for the 1999 median annual earnings of households.

- 6.23 **Household income by state.** The previous problem reports data on the median household income for the entire United States. In a detailed report based on the same sample survey, you find that the estimated median income for four-person families in Michigan is \$65,467. Is the margin of error for this estimate with 95% confidence greater or less than the margin of error for the national median? Why?
- 6.24 **More than one confidence interval.** As we prepare to take a sample and compute a 95% confidence interval, we know that the probability that the interval we compute will cover the parameter is 0.95. That's the meaning of 95% confidence. If we use several such intervals, however, our confidence that *all* give correct results is less than 95%.

Suppose we are interested in confidence intervals for the median household incomes for three states. We compute a 95% interval for each of the three, based on independent samples in the three states.

- (a) What is the probability that all three intervals cover the true median incomes? This probability (expressed as a percent) is our overall confidence level for the three simultaneous statements.
- (b) What is the probability that at least two of the three intervals cover the true median incomes?

- 6.25 **An election poll.** A newspaper headline describing a poll of registered voters taken two weeks before a recent election reads “Ringel leads with 52%.” The accompanying article says that the margin of error is 3% with 95% confidence.
- Explain in plain language to someone who knows no statistics what “95% confidence” means.
 - The poll shows Ringel leading. But the newspaper article says that the election was too close to call. Explain why.
- 6.26 **Manager trainee wages.** A newspaper ad for a manager trainee position contained the statement “Our manager trainees have a first-year earnings average of \$20,000 to \$24,000.” Do you think that the ad is describing a confidence interval? Explain your answer.
- 6.27 **Talk show poll.** A radio talk show invites listeners to enter a dispute about a proposed pay increase for city council members. “What yearly pay do you think council members should get? Call us with your number.” In all, 958 people call. The mean pay they suggest is $\bar{x} = \$9740$ per year, and the standard deviation of the responses is $s = \$1125$. For a large sample such as this, s is very close to the unknown population σ . The station calculates the 95% confidence interval for the mean pay μ that all citizens would propose for council members to be \$9669 to \$9811. Is this result trustworthy? Explain your answer.
- 6.28 **An outlier.** Exercise 6.12 gives the weights of a sample of 24 male runners. Suppose that the sample actually contained 25 runners. The extra runner claimed to weigh 92.3 kg.
- Compute the 95% confidence interval for the mean with this observation included.
 - Would you report the interval you calculated in part (a) of this question or the interval you calculated in Exercise 6.12? Explain the reasons for your choice.

6.2 Tests of Significance

Confidence intervals are one of the two most common types of formal statistical inference. They are appropriate when our goal is to estimate a population parameter. The second common type of inference is directed at a different goal: to assess the evidence provided by the data in favor of some claim about the population.

The reasoning of significance tests

A significance test is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess. The hypothesis is a statement about the parameters in a population or model. The results of a test are expressed in terms of a probability that measures how well the data

and the hypothesis agree. We use the following example to illustrate these ideas.

EXAMPLE 6.6**Banks' net income**

The community bank survey described in Case 6.1 also asked about net income and reported the percent change in net income between the first half of last year and the first half of this year. As you might expect, there is considerable bank-to-bank variation in this percent. Some banks report an increase while others report a decrease. The mean change for the 110 banks in the sample is $\bar{x} = 8.1\%$. Because the sample size is large, we are willing to use the sample standard deviation $s = 26.4\%$ as if it were the population standard deviation σ . The large sample size also makes it reasonable to assume that \bar{x} is approximately Normal.

Is the 8.1% mean increase in a *sample* good evidence that the net income for *all banks* has changed? After all, the sample result might happen just by chance even if the true mean change for all banks is $\mu = 0\%$. To answer this question, we ask another: Suppose that the truth about the population is that $\mu = 0\%$. This is our hypothesis. *What is the probability of observing a sample mean at least as far from zero as 8.1%?* The answer is 0.001. Because this probability is so small, we see that the sample mean $\bar{x} = 8.1$ is incompatible with a population mean of $\mu = 0$. We conclude that the income of community banks has changed since last year.



What are the key steps in Example 6.6? We want to know if the sample gives good evidence that the mean income for all community banks has changed in the past year. Suppose that there was no change, that is, that $\mu = 0$. Compare the sample outcome $\bar{x} = 8.1$ with the no-change population mean $\mu = 0$. The comparison takes the form of a probability: If $\mu = 0$, then a sample of size 110 will have a mean at least as far from 0 as $\bar{x} = 8.1$ only with probability 0.001.

The fact that the calculated probability is very small leads us to conclude that the average percent change in income is not in fact zero. Here's why. If the true mean is $\mu = 0$, we would see a sample mean \bar{x} as far away as 8.1% only once per 1000 samples. So there are only two possibilities:

1. $\mu = 0$ and we have observed something very unusual, or
2. μ is not zero but has some other value that makes the observed data more probable.

We calculated a probability taking the first of these choices as true. That probability guides our final choice. If the probability is very small, the data don't fit the first possibility and we conclude that the mean is not in fact zero. Here is an example in which the data lead to a different conclusion.

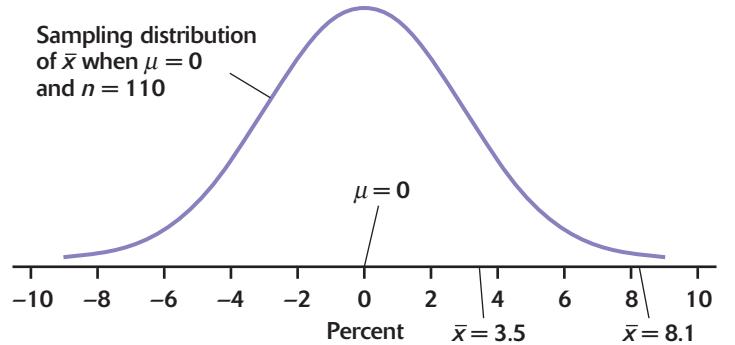


FIGURE 6.6 The mean change in net assets for a sample of 110 banks will have this sampling distribution if the mean for the population of all banks is $\mu = 0$. A sample mean $\bar{x} = 3.5\%$ could easily happen by chance. A sample mean $\bar{x} = 8.1\%$ is so far out on the curve that it would rarely happen just by chance.

EXAMPLE 6.7

Is this percent change different from zero?



Suppose that next year the percent change in net income for a sample of 110 banks is $\bar{x} = 3.5\%$. (We assume that the standard deviation is the same, $\sigma = 26.4\%$.) This sample mean is closer to the value $\mu = 0$ corresponding to no mean change in income. What is the probability that the mean of a sample of size $n = 110$ from a Normal population with mean $\mu = 0$ and standard deviation $\sigma = 26.4$ is as far away or farther away from zero as $\bar{x} = 3.5$? The answer is 0.16. A sample result this far from zero would happen just by chance in 16% of samples from a population having true mean zero. An outcome that could so easily happen just by chance is not good evidence that the population mean is different from zero.



The probabilities in Examples 6.6 and 6.7 measure the compatibility of the outcomes $\bar{x} = 8.1$ and $\bar{x} = 3.5$ with the hypothesis that $\mu = 0$. Figure 6.6 compares the two results graphically. The Normal curve centered at zero is the sampling distribution of \bar{x} when $\mu = 0$. You can see that we are not particularly surprised to observe $\bar{x} = 3.5$, but $\bar{x} = 8.1$ is an extreme outcome. That's the core reasoning of statistical tests: *A sample outcome that would be extreme if a hypothesis were true is evidence that the hypothesis is not true.* Now we proceed to the formal details of tests.

Stating hypotheses

In Example 6.6, we asked whether the bank survey data are plausible if, in fact, the true percent change in net income for all banks (μ) is zero. That is, we ask if the data provide evidence *against* the claim that the population mean is zero. The first step in a test of significance is to state a claim that we will try to find evidence against.

NULL HYPOTHESIS H_0

The statement being tested in a test of significance is called the **null hypothesis**. The test of significance is designed to assess the strength of the evidence against the null hypothesis. Usually the null hypothesis is a statement of “no effect” or “no difference.” We abbreviate “null hypothesis” as H_0 .

A null hypothesis is a statement about a population, expressed in terms of some parameter or parameters. The null hypothesis for Example 6.6 is

$$H_0: \mu = 0$$

It is convenient also to give a name to the statement we hope or suspect is true instead of H_0 . This is called the **alternative hypothesis** and is abbreviated as H_a . In Example 6.6, the alternative hypothesis states that the percent change in net income is not zero. We write this as

$$H_a: \mu \neq 0$$

Hypotheses always refer to some population or model, not to a particular outcome. For this reason, we always state H_0 and H_a in terms of population parameters.

Because H_a expresses the effect that we hope to find evidence for, we often begin with H_a and then set up H_0 as the statement that the hoped-for effect is not present. Stating H_a is not always straightforward. It is not always clear, in particular, whether H_a should be **one-sided** or **two-sided**.

The alternative $H_a: \mu \neq 0$ in the bank net income example is two-sided. We simply asked if income had changed. In any given year, income may increase or decrease, so we include both possibilities in the alternative hypothesis. Here is a setting in which a one-sided alternative is appropriate.

**one-sided and
two-sided
alternatives**

EXAMPLE 6.8**Have we reduced processing time?**

Your company hopes to reduce the mean time μ required to process customer orders. At present, this mean is 3.8 days. You study the process and eliminate some unnecessary steps. Did you succeed in decreasing the average process time? You hope to show that the mean is now less than 3.8 days, so the alternative hypothesis is one-sided, $H_a: \mu < 3.8$. The null hypothesis is as usual the “no-change” value, $H_0: \mu = 3.8$ days.

The alternative hypothesis should express the hopes or suspicions we had in mind when we decided to collect the data. It is cheating to first look at the data and then frame H_a to fit what the data show. If you do not have a specific direction firmly in mind in advance, use a two-sided alternative.

In fact, some users of statistics argue that we should *always* work with the two-sided alternative.

The choice of the hypotheses in Example 6.8 as

$$H_0: \mu = 3.8$$

$$H_a: \mu < 3.8$$

deserves a final comment. We do not expect that elimination of steps in the ordering process would actually increase the process time. However, we can allow for an increase by including this case in the null hypothesis. Then we would write

$$H_0: \mu \geq 3.8$$

$$H_a: \mu < 3.8$$

This statement is logically satisfying because the hypotheses account for all possible values of μ . However, only the parameter value in H_0 that is closest to H_a influences the form of the test in all common significance-testing situations. We will therefore take H_0 to be the simpler statement that the parameter has a specific value, in this case $H_0: \mu = 3.8$.

- APPLY YOUR KNOWLEDGE**
- 6.29 **Customer feedback.** Feedback from your customers shows that many think it takes too long to fill out the online order form for your products. You redesign the form and plan a survey of customers to determine whether or not they think that the new form is actually an improvement. Sampled customers will respond using a five-point scale: -2 if the new form takes much less time than the old form; -1 if the new form takes a little less time; 0 if the new form takes about the same time; $+1$ if the new form takes a little more time; and $+2$ if the new form takes much more time. The mean response from the sample is \bar{x} , and the mean response for all of your customers is μ . State null and alternative hypotheses that provide a framework for examining whether or not the new form is an improvement.
- 6.30 **DXA scanners.** A dual X-ray absorptiometry (DXA) scanner is used to measure bone mineral density for people who may be at risk for osteoporosis. Customers want assurance that your company's latest-model DXA scanner is accurate. You therefore supply an object called a "phantom" that has known mineral density $\mu = 1.4$ grams per square centimeter. A user scans the phantom 10 times and compares the sample mean reading \bar{x} with the theoretical mean μ using a significance test. State the null and alternative hypotheses for this test.

Test statistics

We will learn the form of significance tests in a number of common situations. Here are some principles that apply to most tests and that help in understanding the form of tests:

- The test is based on a statistic that estimates the parameter appearing in the hypotheses. Usually this is the same estimate we would use in a confidence interval for the parameter. When H_0 is true, we expect the estimate to take a value near the parameter value specified by H_0 .

- Values of the estimate far from the parameter value specified by H_0 give evidence against H_0 . The alternative hypothesis determines which directions count against H_0 .

EXAMPLE 6.9**Banks' income: the hypotheses**

For Example 6.6, the hypotheses are stated in terms of the mean change in net income for all community banks:

$$H_0: \mu = 0$$

$$H_a: \mu \neq 0$$

The estimate of μ is the sample mean \bar{x} . Because H_a is two-sided, large positive and negative values of \bar{x} (large increases and decreases of net income in the sample) count as evidence against the null hypothesis.

test statistic

A **test statistic** measures compatibility between the null hypothesis and the data. Many test statistics can be thought of as a distance between a sample estimate of a parameter and the value of the parameter specified by the null hypothesis.

EXAMPLE 6.10**Banks' income: the test statistic**

For Example 6.6, the null hypothesis is $H_0: \mu = 0$, and a sample gave $\bar{x} = 8.1$. The test statistic for this problem is the standardized version of \bar{x} :

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

This statistic is the distance between the sample mean and the hypothesized population mean in the standard scale of z -scores. In this example,

$$z = \frac{8.1 - 0}{26.4/\sqrt{110}} = 3.22$$

Even without a formal probability calculation, we know that standard score $z = 3.22$ lies far out on a Normal curve.

The income changes for individual banks do not look Normal, but the central limit theorem assures us that \bar{x} (and therefore z) is approximately Normal because the sample size is large. So if the null hypothesis $H_0: \mu = 0$ is true, the test statistic z is a random variable that has approximately the standard Normal distribution $N(0, 1)$. To move from the test statistic z to a probability, we must do Normal probability calculations. Review Sections 1.3 and 4.4 if you need to refresh your skills.

P-values

A test of significance assesses the evidence against the null hypothesis and provides a numerical summary of this evidence in terms of a probability. The idea is that “surprising” outcomes are evidence against H_0 . A surprising outcome is one that is far from what we would expect if H_0 were true. In many cases, including our bank income example, the standard scale of z -scores is one way to measure “far from what we would expect.” In Example 6.10, the standardized distance of the sample outcome \bar{x} from the hypothesized population mean $\mu = 0$ was $z = 3.22$. This suggests that the sample is not compatible with the null hypothesis $H_0: \mu = 0$. In fact, the Supreme Court of the United States has said that “two or three standard deviations” ($z = 2$ or 3) is its criterion for rejecting H_0 , and this is the criterion used in most applications involving the law.

Because not all test statistics produce z -scores, we translate the values of test statistics into a common language, the language of probability. A test of significance finds the probability of getting an outcome *as extreme or more extreme than the actually observed outcome*. “Extreme” means “far from what we would expect if H_0 were true.” The direction or directions that count as “far from what we would expect” are determined by the alternative hypothesis H_a .

P-VALUE

The probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the **P-value** of the test. The smaller the P-value, the stronger the evidence against H_0 provided by the data.

To calculate the *P*-value, we must use the sampling distribution of the test statistic. In this chapter, we need only the standard Normal distribution for the test statistic z .

EXAMPLE 6.11

Banks' income: the **P**-value



In Example 6.6 the observations are an SRS of size $n = 110$ from a population of community banks with $\sigma = 26.4$. The observed average percent change in net income is $\bar{x} = 8.1$. In Example 6.10, we found that the test statistic for testing $H_0: \mu = 0$ versus $H_a: \mu \neq 0$ is

$$z = \frac{8.1 - 0}{26.4/\sqrt{110}} = 3.22$$

If the null hypothesis is true, we expect z to take a value not too far from 0. Because the alternative is two-sided, values of z far from 0 *in either direction* count as evidence against H_0 and in favor of H_a . So the *P*-value is

$$P(z \geq 3.22) + P(z \leq -3.22)$$

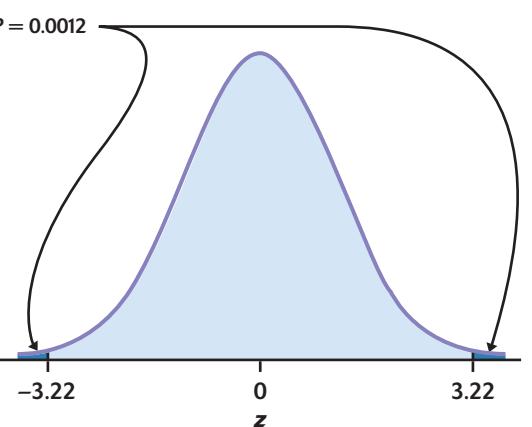


FIGURE 6.7 The P -value for Example 6.11. The two-sided P -value is the probability (when H_0 is true) that \bar{x} takes a value at least as far from 0 as the actually observed value.

If H_0 is true, then z has approximately the standard Normal $N(0, 1)$ distribution. Figure 6.7 shows the P -value as a very small area under the standard Normal curve.

From Table A, we find

$$P(Z \geq 3.22) = 1 - 0.9994 = 0.0006$$

Similarly, from Table A or using the symmetry of the Normal distribution,

$$P(Z \leq -3.22) = 0.0006$$

That is,

$$P = 2P(Z \geq 3.22) = 0.0012$$

In Example 6.6 we rounded this value to 0.001.



APPLY YOUR KNOWLEDGE

- 6.31 **Spending on housing.** The Census Bureau reports that households spend an average of 31% of their total spending on housing. A homebuilders association in Cleveland wonders if the national finding applies in their area. They interview a sample of 40 households in the Cleveland metropolitan area to learn what percent of their spending goes toward housing. Take μ to be the mean percent of spending devoted to housing among all Cleveland households. We want to test the hypotheses

$$H_0: \mu = 31\%$$

$$H_a: \mu \neq 31\%$$

The population standard deviation is $\sigma = 9.6\%$.

- (a) The study finds $\bar{x} = 28.6\%$ for the 40 households in the sample. What is the value of the test statistic z ? Sketch a standard Normal curve and mark z on the axis. Shade the area under the curve that represents the P -value.

- (b) Calculate the *P*-value. Are you convinced that Cleveland differs from the national average?
- 6.32 In the setting of the previous exercise, suppose that the Cleveland home-builders were convinced, before interviewing their sample, that residents of Cleveland spend less than the national average on housing. Do the interviews support their conviction? State null and alternative hypotheses. Find the *P*-value, using the interview results given in the previous problem. Why do the same data give different *P*-values in these two problems?
- 6.33 The homebuilders wonder if the national finding applies in the Cleveland area. They have no idea whether Cleveland residents spend more or less than the national average. Because their interviews find that $\bar{x} = 28.6\%$, less than the national 31%, their analyst tests

$$H_0: \mu = 31\%$$
$$H_a: \mu < 31\%$$

Explain why this is incorrect.

Statistical significance

Statistical software automates the task of calculating the test statistic and its *P*-value. You must still decide which test is appropriate and whether to use a one-sided or two-sided test. You must also decide what conclusion the computer's numbers support. We know that smaller *P*-values indicate stronger evidence against the null hypothesis. But how strong is strong enough?

One approach is to announce in advance how much evidence against H_0 we will require to reject H_0 . We compare the *P*-value with a level that says "this evidence is strong enough." The decisive level is called the **significance level**. It is denoted by α , the Greek letter alpha. If we choose $\alpha = 0.05$, we are requiring that the data give evidence against H_0 so strong that it would happen no more than 5% of the time (1 time in 20) when H_0 is true. If we choose $\alpha = 0.01$, we are insisting on stronger evidence against H_0 , evidence so strong that it would appear only 1% of the time (1 time in 100) if H_0 is in fact true.

STATISTICAL SIGNIFICANCE

If the *P*-value is as small or smaller than α , we say that the data are statistically significant at level α .

"Significant" in the statistical sense does not mean "important." The original meaning of the word is "signifying something." In statistics, the term is used to indicate only that the evidence against the null hypothesis reached the standard set by α . Significance at level 0.01 is often expressed by

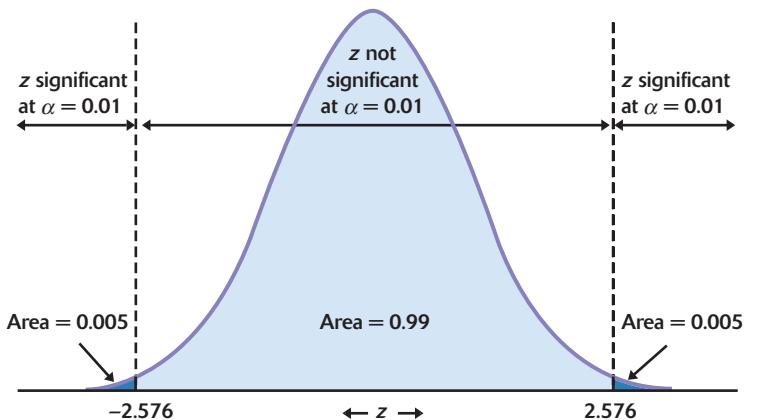


FIGURE 6.8 Values of z in the extreme area 0.01 under a standard Normal curve are significant at $\alpha = 0.01$.

the statement “The results were significant ($P < 0.01$).” Here P stands for the P -value. The P -value is more informative than a statement of significance because we can then assess significance at any level we choose. For example, a result with $P = 0.03$ is significant at the $\alpha = 0.05$ level but is not significant at the $\alpha = 0.01$ level.

You need not actually find the P -value to assess significance at a fixed level α . You need only compare the observed statistic z with a *critical value* that marks off area α in one or both tails of the standard Normal curve.

EXAMPLE 6.12



Significant?

In Example 6.11, the test statistic took the value $z = 3.22$. Is this significant at the $\alpha = 0.01$ level? Values in the extreme 0.01 area of the standard Normal curve are significant at the 0.01 level. Because the alternative is two-sided, this area is divided equally between the two tails of the curve. So z is significant if it lies in the extreme area 0.005 at either end. Look at the z^* row at the bottom of Table D. The extreme 0.005 in the right tail starts at $z^* = 2.576$. That is, the z -values that are significant at the $\alpha = 0.01$ level are $z > 2.576$ and $z < -2.576$. Figure 6.8 shows why. The bank survey outcome $z = 3.22$ is statistically significant ($P < 0.01$).

APPLY YOUR KNOWLEDGE

- 6.34 How to show that you are rich. Every society has its own marks of wealth and prestige. In ancient China, it appears that owning pigs was such a mark. Evidence comes from examining burial sites. If the skulls of sacrificed pigs tend to appear along with expensive ornaments, that suggests that the pigs, like the ornaments, signal the wealth and prestige of the person buried. A study of burials from around 3500 B.C. concluded that “there are striking differences in grave goods between burials with pig skulls and

burials without them. ... A test indicates that the two samples of total artifacts are significantly different at the 0.01 level.⁵ Explain clearly why “significantly different at the 0.01 level” gives good reason to think that there really is a systematic difference between burials that contain pig skulls and those that lack them.

- 6.35 **Significance.** You are testing $H_0: \mu = 0$ against $H_a: \mu \neq 0$ based on an SRS of 20 observations from a Normal population. What values of the z statistic are statistically significant at the $\alpha = 0.005$ level?
- 6.36 **Significance.** You are testing $H_0: \mu = 0$ against $H_a: \mu > 0$ based on an SRS of 20 observations from a Normal population. What values of the z statistic are statistically significant at the $\alpha = 0.005$ level?
- 6.37 **The Supreme Court speaks.** Court cases in such areas as employment discrimination often involve statistical evidence. The Supreme Court has said that z -scores beyond $z^* = 2$ or 3 are generally convincing statistical evidence. For a two-sided test, what significance level corresponds to $z^* = 2$? To $z^* = 3$?

Tests for a population mean

There are four steps in carrying out a significance test:

1. State the hypotheses.
2. Calculate the test statistic.
3. Find the P -value.
4. State your conclusion in the context of your specific setting.

Once you have stated your hypotheses and identified the proper test, you or your software can do Steps 2 and 3 by following a recipe. We now present the recipe for the test we have used in our examples.

We have an SRS of size n drawn from a Normal population with unknown mean μ . We want to test the hypothesis that μ has a specified value. Call the specified value μ_0 . The null hypothesis is

$$H_0: \mu = \mu_0$$

The test is based on the sample mean \bar{x} . Because Normal calculations require standardized variables, we will use as our test statistic the standardized sample mean

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

one-sample z statistic

This **one-sample z statistic** has the standard Normal distribution when H_0 is true. The P -value of the test is the probability that z takes a value at least as extreme as the value for our sample. What counts as extreme is determined by the alternative hypothesis H_a . Here is a summary.

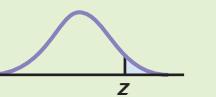
z TEST FOR A POPULATION MEAN

To test the hypothesis $H_0: \mu = \mu_0$ based on an SRS of size n from a population with unknown mean μ and known standard deviation σ , compute the one-sample z statistic

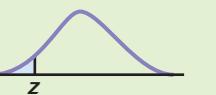
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

In terms of a variable Z having the standard Normal distribution, the P -value for a test of H_0 against

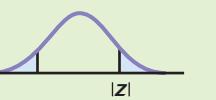
$$H_a: \mu > \mu_0 \quad \text{is} \quad P(Z \geq z)$$



$$H_a: \mu < \mu_0 \quad \text{is} \quad P(Z \leq z)$$



$$H_a: \mu \neq \mu_0 \quad \text{is} \quad 2P(|Z| \geq |z|)$$



These P -values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

EXAMPLE 6.13**Blood pressures of executives**

The medical director of a large company is concerned about the effects of stress on the company's younger executives. According to the National Center for Health Statistics, the mean systolic blood pressure for males 35 to 44 years of age is 128 and the standard deviation in this population is 15. The medical director examines the records of 72 executives in this age group and finds that their mean systolic blood pressure is $\bar{x} = 129.93$. Is this evidence that the mean blood pressure for all the company's young male executives is higher than the national average? As usual in this chapter, we make the unrealistic assumption that the population standard deviation is known, in this case that executives have the same $\sigma = 15$ as the general population.

Step 1: Hypotheses. The hypotheses about the unknown mean μ of the executive population are

$$H_0: \mu = 128$$

$$H_a: \mu > 128$$

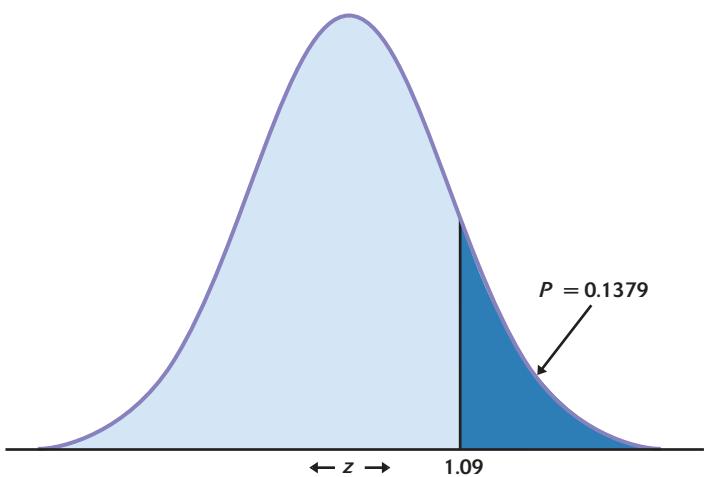


FIGURE 6.9 The P -value for the one-sided test in Example 6.13.

Step 2: Test statistic. The z test requires that the 72 executives in the sample are an SRS from the population of the company's young male executives. We must ask how the data were produced. If records are available only for executives with recent medical problems, for example, the data are of little value for our purpose. It turns out that all executives are given a free annual medical exam and that the medical director selected 72 exam results at random. The one-sample z statistic is

$$\begin{aligned} z &= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{129.93 - 128}{15/\sqrt{72}} \\ &= 1.09 \end{aligned}$$

Step 3: P -value. Draw a picture to help find the P -value. Figure 6.9 shows that the P -value is the probability that a standard Normal variable Z takes a value of 1.09 or greater. From Table A we find that this probability is

$$P = P(Z \geq 1.09) = 1 - 0.8621 = 0.1379$$

Step 4: Conclusion. About 14% of the time, an SRS of size 72 from the general male population would have a mean blood pressure as high as that of the executive sample. The observed $\bar{x} = 129.93$ is not significantly higher than the national average at any of the usual significance levels.



The data in Example 6.13 do *not* establish that the mean blood pressure μ for this company's middle-aged male executives is 128. We sought evidence that μ was higher than 128 and failed to find convincing evidence. That is all we can say. No doubt the mean blood pressure of the entire executive population is not exactly equal to 128. A large enough sample would give evidence of the difference, even if it is very small. Tests of significance assess

the evidence *against* H_0 . If the evidence is strong, we can confidently reject H_0 in favor of the alternative. Failing to find evidence against H_0 means only that the data are consistent with H_0 , not that we have clear evidence that H_0 is true.

EXAMPLE 6.14**A company-wide health promotion campaign**

The company medical director institutes a health promotion campaign to encourage employees to exercise more and eat a healthier diet. One measure of the effectiveness of such a program is a drop in blood pressure. Choose a random sample of 50 employees, and compare their blood pressures from annual physical examinations given before the campaign and again a year later. The mean change in blood pressure for these $n = 50$ employees is $\bar{x} = -6$. We take the population standard deviation to be $\sigma = 20$.

Step 1: Hypotheses. We want to know if the health campaign reduced average blood pressure. Taking μ to be the mean change in blood pressure for all employees,

$$\begin{aligned} H_0: \mu &= 0 \\ H_a: \mu &< 0 \end{aligned}$$

Step 2: Test statistic. The one-sample z test is appropriate:

$$\begin{aligned} z &= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{-6 - 0}{20/\sqrt{50}} \\ &= -2.12 \end{aligned}$$

Step 3: P-value. Because H_a is one-sided on the low side, large negative values of z count against H_0 . See Figure 6.10. From Table A, we find that the P -value is

$$P = P(Z \leq -2.12) = 0.0170$$

Step 4: Conclusion. A mean change in blood pressure of -6 or better would occur only 17 times in 1000 samples if the campaign had no effect on the blood pressures of the employees. This is convincing evidence that the mean blood pressure in the population of all employees has decreased.



Our conclusion in Example 6.14 is cautious. We would like to conclude that the health campaign *caused* the drop in mean blood pressure. But the data are not protected from confounding. If, for example, the local television station runs a series on the risk of heart attacks and the value of better diet and exercise, many employees may be moved to reform their health habits even without encouragement from the company. Only a randomized comparative experiment protects against such lurking variables. The medical director preferred to launch a company-wide campaign that

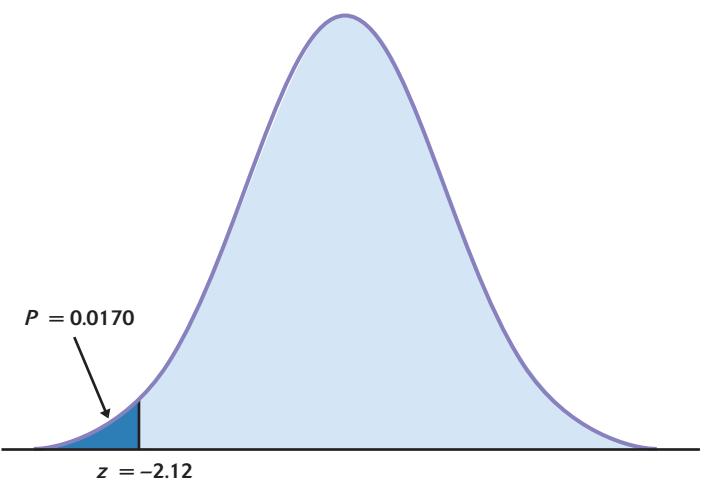


FIGURE 6.10 The P -value for the one-sided test in Example 6.14.

appealed to all employees. This is no doubt a sound medical decision, but the absence of a control group weakens the statistical conclusion.

APPLY YOUR KNOWLEDGE

- 6.38 **Testing a random number generator.** Statistical software has a “random number generator” that is supposed to produce numbers uniformly distributed between 0 to 1. If this is true, the numbers generated come from a population with $\mu = 0.5$. A command to generate 100 random numbers gives outcomes with mean $\bar{x} = 0.532$ and $s = 0.316$. Because the sample is reasonably large, take the population standard deviation also to be $\sigma = 0.316$. Do we have evidence that the mean of all numbers produced by this software is not 0.5?
- 6.39 A test of the null hypothesis $H_0: \mu = \mu_0$ gives test statistic $z = 1.8$.
- What is the P -value if the alternative is $H_a: \mu > \mu_0$?
 - What is the P -value if the alternative is $H_a: \mu < \mu_0$?
 - What is the P -value if the alternative is $H_a: \mu \neq \mu_0$?
- 6.40 **A new supplier.** A new supplier offers a good price on a catalyst used in your production process. You compare the purity of this catalyst with that from your current supplier. The P -value for a test of “no difference” is 0.15. Can you be confident that the purity of the new product is the same as the purity of the product that you have been using? Discuss.

Two-sided significance tests and confidence intervals

A 95% confidence interval captures the true value of μ in 95% of all samples. If we are 95% confident that the true μ lies in our interval, we are also confident that values of μ that fall outside our interval are incompatible with the data. That sounds like the conclusion of a test of significance. In fact, there is an intimate connection between 95% confidence intervals and significance at the 5% level. The same connection holds between 99% confidence intervals and significance at the 1% level, and so on.

CONFIDENCE INTERVALS AND TWO-SIDED TESTS

A level α two-sided significance test rejects a hypothesis $H_0: \mu = \mu_0$ exactly when the value μ_0 falls outside a level $1 - \alpha$ confidence interval for μ .

EXAMPLE 6.15

Testing pharmaceutical products

The Deely Laboratory analyzes pharmaceutical products to verify the concentration of active ingredients. Such chemical analyses are not perfectly precise. Repeated measurements on the same specimen will give slightly different results. The results of repeated measurements follow a Normal distribution quite closely. The analysis procedure has no bias, so that the mean μ of the population of all measurements is the true concentration in the specimen. The standard deviation of this distribution is a property of the analytical procedure and is known to be $\sigma = 0.0068$ grams per liter. The laboratory analyzes each specimen three times and reports the mean result.

A client sends a specimen for which the concentration of active ingredient is supposed to be 0.86%. Deely's three analyses give concentrations

0.8403 0.8363 0.8447

Is there significant evidence at the 1% level that the true concentration is not 0.86%? This calls for a test of the hypotheses

$$H_0: \mu = 0.86$$

$$H_a: \mu \neq 0.86$$

We will carry out the test twice, with the usual significance test and then from a 99% confidence interval.



First, the test. The mean of the three analyses is $\bar{x} = 0.8404$. The one-sample z test statistic is therefore

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{0.8404 - 0.86}{0.0068/\sqrt{3}} = -4.99$$

We need not find the exact P -value to assess significance at the $\alpha = 0.01$ level. Look in Table D under tail area 0.005 because the alternative is two-sided. The z -values that are significant at the 1% level are $z > 2.576$ and $z < -2.576$. Our observed $z = -4.99$ is significant ($P < 0.01$).

To compute a 99% confidence interval for the mean concentration, find in Table D the critical value for 99% confidence. It is $z^* = 2.576$, the same critical value that marked off significant z 's in our test. The confidence interval is

$$\begin{aligned} \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} &= 0.8404 \pm 2.576 \frac{0.0068}{\sqrt{3}} \\ &= 0.8404 \pm 0.0101 \\ &= (0.8303, 0.8505) \end{aligned}$$

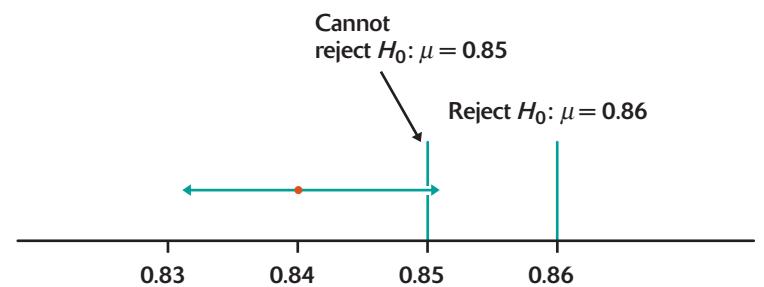


FIGURE 6.11 Values of μ falling outside a 99% confidence interval can be rejected at the 1% significance level. Values falling inside the interval cannot be rejected.

The hypothesized value $\mu_0 = 0.86$ falls outside this confidence interval. We are therefore 99% confident that μ is *not* equal to 0.86, so we can reject

$$H_0: \mu = 0.86$$

at the 1% significance level. On the other hand, we cannot reject

$$H_0: \mu = 0.85$$

at the 1% level in favor of the two-sided alternative $H_a: \mu \neq 0.85$, because 0.85 lies inside the 99% confidence interval for μ . Figure 6.11 illustrates both cases.

- 6.41 The P -value for a two-sided test of the null hypothesis $H_0: \mu = 10$ is 0.06.
 - (a) Does the 95% confidence include the value 10? Why?
 - (b) Does the 90% confidence include the value 10? Why?
- 6.42 A 95% confidence interval for a population mean is (28, 35).
 - (a) Can you reject the null hypothesis that $\mu = 34$ at the 5% significance level? Why?
 - (b) Can you reject the null hypothesis that $\mu = 36$ at the 5% significance level? Why?

P-values versus fixed α

In Example 6.15, we concluded that the test statistic $z = -4.99$ is significant at the 1% level. This isn't the whole story. The observed z is far beyond the critical value for $\alpha = 0.01$, and the evidence against H_0 is far stronger than 1% significance suggests. The P -value $P = 0.0000006$ (from software) gives a better sense of how strong the evidence is. *The P-value is the smallest level α at which the data are significant.* Knowing the P -value allows us to assess significance at any level.

EXAMPLE 6.16

How significant?

In Example 6.14, we tested the hypotheses

$$H_0: \mu = 0$$

$$H_a: \mu < 0$$

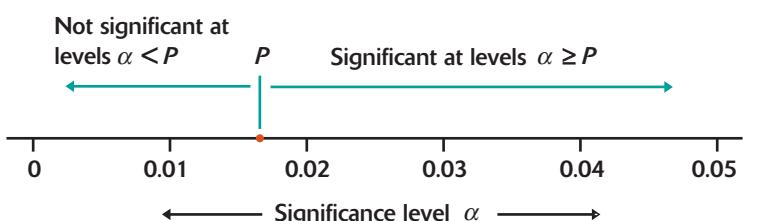


FIGURE 6.12 An outcome with P -value P is significant at all levels α at or above P and is not significant at smaller levels α .

to evaluate the effectiveness of a health promotion program on blood pressure. The test had the P -value $P = 0.0170$. This result is not significant at the $\alpha = 0.01$ level, because $0.0170 \geq 0.01$. However, it is significant at the $\alpha = 0.05$ level, because the P -value is less than 0.05. See Figure 6.12.

A P -value is more informative than a reject-or-not finding at a fixed significance level. But assessing significance at a fixed level α is easier, because no probability calculation is required. You need only look up a critical value in a table. Because the practice of statistics almost always employs software that calculates P -values automatically, tables of critical values are becoming outdated. We include the usual tables of critical values (such as Table D) at the end of the book for learning purposes and to rescue students without good computing facilities. The tables can be used directly to carry out fixed α tests. They also allow us to approximate P -values quickly without a probability calculation. The following example illustrates the use of Table D to find an approximate P -value.

EXAMPLE 6.17 Fill the bottles

Bottles of a popular cola drink are supposed to contain 300 milliliters (ml) of cola. There is some variation from bottle to bottle because the filling machinery is not perfectly precise. The distribution of the contents is Normal with standard deviation $\sigma = 3$ ml. An inspector who suspects that the bottler is underfilling measures the contents of six bottles. The results are

299.4 297.7 301.0 298.9 300.2 297.0

Is this convincing evidence that the mean contents of cola bottles is less than the advertised 300 ml? The hypotheses are

$$H_0: \mu = 300$$

$$H_a: \mu < 300$$

The sample mean contents of the six bottles measured is $\bar{x} = 299.03$ ml. The z test statistic is therefore

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{299.03 - 300}{3/\sqrt{6}} = -0.792$$

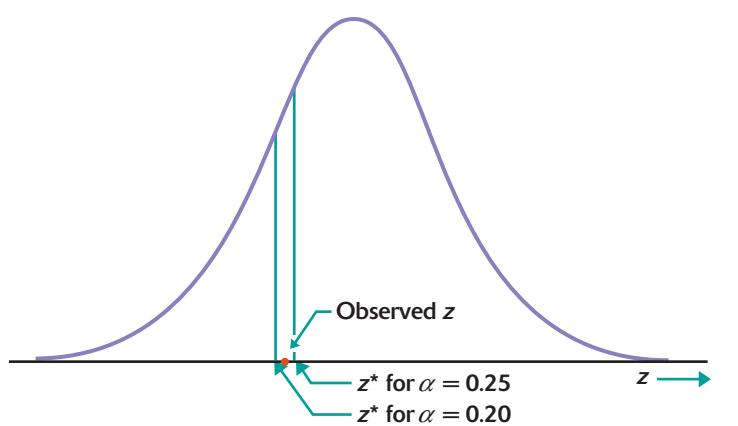


FIGURE 6.13 The P -value of a z statistic can be approximated by noting which levels from Table D it falls between. Here, P lies between 0.20 and 0.25.

Small values of z count against H_0 because H_a is one-sided on the low side. The P -value is

$$P(Z \leq -0.792)$$

Rather than compute this probability, compare $z = -0.792$ with the critical values in Table D. According to this table, we would

reject H_0 at level $\alpha = 0.25$ if $z \leq -0.674$

reject H_0 at level $\alpha = 0.20$ if $z \leq -0.841$

As Figure 6.13 illustrates, the observed $z = -0.792$ lies between these two critical values. We can reject at $\alpha = 0.25$ but not at $\alpha = 0.20$. That is the same as saying that the P -value lies between 0.25 and 0.20. A sample mean at least as small as that observed will occur in between 20% and 25% of all samples if the population mean is $\mu = 300$ ml. There is no convincing evidence that the mean is below 300, and there is no need to calculate the P -value more exactly.

-
- APPLY YOUR KNOWLEDGE**
- 6.43 The P -value for a significance test is 0.078.
 - (a) Do you reject the null hypothesis at level $\alpha = 0.05$?
 - (b) Do you reject the null hypothesis at level $\alpha = 0.01$?
 - (c) Explain your answers.

 - 6.44 The P -value for a significance test is 0.03.
 - (a) Do you reject the null hypothesis at level $\alpha = 0.05$?
 - (b) Do you reject the null hypothesis at level $\alpha = 0.01$?
 - (c) Explain your answers.

SECTION 6.2 SUMMARY

- A test of significance is intended to assess the evidence provided by data against a **null hypothesis** H_0 and in favor of an **alternative hypothesis** H_a .

It provides a method for ruling out chance as an explanation for data that deviate from what we expect under H_0 .

- The hypotheses are stated in terms of population parameters. Usually H_0 is a statement that no effect is present, and H_a says that a parameter differs from its null value, in a specific direction (**one-sided alternative**) or in either direction (**two-sided alternative**).
- The test is based on a **test statistic**. The **P-value** is the probability, computed assuming that H_0 is true, that the test statistic will take a value at least as extreme as that actually observed. Small P-values indicate strong evidence against H_0 . Calculating P-values requires knowledge of the sampling distribution of the test statistic when H_0 is true.
- If the P-value is as small or smaller than a specified value α , the data are **statistically significant** at significance level α .
- Significance tests for the hypothesis $H_0: \mu = \mu_0$ concerning the unknown mean μ of a population are based on the ***z* statistic**:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- The *z* test assumes an SRS of size n , known population standard deviation σ , and either a Normal population or a large sample. P-values are computed from the Normal distribution (Table A). Fixed α tests use the table of **standard Normal critical values** (z^* row in Table D).

SECTION 6.2 EXERCISES

- 6.45 Hypotheses.** Each of the following situations requires a significance test about a population mean μ . State the appropriate null hypothesis H_0 and alternative hypothesis H_a in each case.
- The mean area of the several thousand apartments in a new development is advertised to be 1250 square feet. A tenant group thinks that the apartments are smaller than advertised. They hire an engineer to measure a sample of apartments to test their suspicion.
 - Larry's car averages 32 miles per gallon on the highway. He now switches to a new motor oil that is advertised as increasing gas mileage. After driving 3000 highway miles with the new oil, he wants to determine if his gas mileage actually has increased.
 - The diameter of a spindle in a small motor is supposed to be 5 millimeters. If the spindle is either too small or too large, the motor will not perform properly. The manufacturer measures the diameter in a sample of motors to determine whether the mean diameter has moved away from the target.
- 6.46 Hypotheses.** In each of the following situations, a significance test for a population mean μ is called for. State the null hypothesis H_0 and the alternative hypothesis H_a in each case.
- Experiments on learning in animals sometimes measure how long it takes a mouse to find its way through a maze. The mean time is

- 18 seconds for one particular maze. A researcher thinks that a loud noise will cause the mice to complete the maze faster. She measures how long each of 10 mice takes with a noise as stimulus.
- (b) The examinations in a large accounting class are scaled after grading so that the mean score is 50. A self-confident teaching assistant thinks that his students have a higher mean score than the class as a whole. His students this semester can be considered a sample from the population of all students he might teach, so he compares their mean score with 50.
- (c) A university gives credit in French language courses to students who pass a placement test. The language department wants to know if students who get credit in this way differ in their understanding of spoken French from students who actually take the French courses. Experience has shown that the mean score of students in the courses on a standard listening test is 24. The language department gives the same listening test to a sample of 40 students who passed the credit examination to see if their performance is different.
- 6.47 Hypotheses.** In each of the following situations, state an appropriate null hypothesis H_0 and alternative hypothesis H_a . Be sure to identify the parameters that you use to state the hypotheses. (We have not yet learned how to test these hypotheses.)
- (a) An economist believes that among employed young adults there is a positive correlation between income and the percent of disposable income that is saved. To test this, she gathers income and savings data from a sample of employed persons in her city aged 25 to 34.
- (b) A sociologist asks a large sample of high school students which academic subject they like best. She suspects that a higher percent of males than of females will name economics as their favorite subject.
- (c) An education researcher randomly divides sixth-grade students into two groups for physical education class. He teaches both groups basketball skills, using the same methods of instruction in both classes. He encourages Group A with compliments and other positive behavior but acts cool and neutral toward Group B. He hopes to show that positive teacher attitudes result in a higher mean score on a test of basketball skills than do neutral attitudes.
- 6.48 Hypotheses.** Translate each of the following research questions into appropriate H_0 and H_a .
- (a) Census Bureau data show that the mean household income in the area served by a shopping mall is \$62,500 per year. A market research firm questions shoppers at the mall to find out whether the mean household income of mall shoppers is higher than that of the general population.
- (b) Last year, your company's service technicians took an average of 2.6 hours to respond to trouble calls from business customers who had purchased service contracts. Do this year's data show a different average response time?
- 6.49 Blood pressure and calcium.** A randomized comparative experiment examined whether a calcium supplement in the diet reduces the blood pressure of healthy men.⁶ The subjects received either a calcium supplement or a placebo for 12 weeks. The statistical analysis was quite complex, but one conclusion

was that “the calcium group had lower seated systolic blood pressure ($P = 0.008$) compared with the placebo group.” Explain this conclusion, especially the P -value, as if you were speaking to a doctor who knows no statistics.

- 6.50 **California brushfires.** We often see televised reports of brushfires threatening homes in California. Some people argue that the modern practice of quickly putting out small fires allows fuel to accumulate and so increases the damage done by large fires. A detailed study of historical data suggests that this is wrong—the damage has risen simply because there are more houses in risky areas.⁷ As usual, the study report gives statistical information tersely. Here is the summary of a regression of number of fires per decade (9 data points, for the 1910s to the 1990s):

Collectively, since 1910, there has been a highly significant increase ($r^2 = 0.61$, $P < 0.01$) in the number of fires per decade.

How would you explain this statement to someone who knows no statistics? Include an explanation of both the description given by r^2 and the statistical significance.

- 6.51 **Exercise and statistics exams.** A study examined whether exercise affects how students perform on their final exam in statistics. The P -value was given as 0.87.
- State null and alternative hypotheses that could be used for this study. (Note that there is more than one correct answer.)
 - Do you reject the null hypothesis? State your conclusion in plain language.
 - What other facts about the study would you like to know for a proper interpretation of the results?
- 6.52 **Financial aid.** The financial aid office of a university asks a sample of students about their employment and earnings. The report says that “for academic year earnings, a significant difference ($P = 0.038$) was found between the sexes, with men earning more on the average. No difference ($P = 0.476$) was found between the earnings of black and white students.”⁸ Explain both of these conclusions, for the effects of sex and of race on mean earnings, in language understandable to someone who knows no statistics.
- 6.53 **Who is the author?** Statistics can help decide the authorship of literary works. Sonnets by a certain Elizabethan poet are known to contain an average of $\mu = 6.9$ new words (words not used in the poet’s other works). The standard deviation of the number of new words is $\sigma = 2.7$. Now a manuscript with 5 new sonnets has come to light, and scholars are debating whether it is the poet’s work. The new sonnets contain an average of $\bar{x} = 11.2$ words not used in the poet’s known works. We expect poems by another author to contain more new words, so to see if we have evidence that the new sonnets are not by our poet we test

$$H_0: \mu = 6.9$$

$$H_a: \mu > 6.9$$

Give the z test statistic and its P -value. What do you conclude about the authorship of the new poems?

- 6.54 **Study habits.** The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures the motivation, attitude toward school, and study habits of students. Scores range from 0 to 200. The mean score for U.S. college students is about 115, and the standard deviation is about 30. A teacher who suspects that older students have better attitudes toward school gives the SSHA to 20 students who are at least 30 years of age. Their mean score is $\bar{x} = 135.2$.
- (a) Assuming that $\sigma = 30$ for the population of older students, carry out a test of

$$H_0: \mu = 115$$

$$H_a: \mu > 115$$

Report the P -value of your test, and state your conclusion clearly.

- (b) Your test in (a) required two important assumptions in addition to the assumption that the value of σ is known. What are they? Which of these assumptions is most important to the validity of your conclusion in (a)?

- 6.55 **Corn yield.** The mean yield of corn in the United States is about 135 bushels per acre. A survey of 40 farmers this year gives a sample mean yield of $\bar{x} = 138.8$ bushels per acre. We want to know whether this is good evidence that the national mean this year is not 135 bushels per acre. Assume that the farmers surveyed are an SRS from the population of all commercial corn growers and that the standard deviation of the yield in this population is $\sigma = 10$ bushels per acre. Give the P -value for the test of

$$H_0: \mu = 135$$

$$H_a: \mu \neq 135$$

Are you convinced that the population mean is not 135 bushels per acre? Is your conclusion correct if the distribution of corn yields is somewhat non-Normal? Why?

- 6.56 **Cigarettes.** According to data from the Tobacco Institute Testing Laboratory, Camel Lights King Size cigarettes contain an average of 1.4 milligrams of nicotine. An advocacy group commissions an independent test to see if the mean nicotine content is higher than the industry laboratory claims.

- (a) What are H_0 and H_a ?
(b) Suppose that the test statistic is $z = 2.42$. Is this result significant at the 5% level?
(c) Is the result significant at the 1% level?

- 6.57 **Academic probation and TV watching.** There are other z statistics that we have not yet met. We can use Table D to assess the significance of any z statistic. A study compares the habits of students who are on academic probation with students whose grades are satisfactory. One variable measured is the hours spent watching television last week. The null hypothesis is “no difference” between the means for the two populations. The alternative hypothesis is two-sided. The value of the test statistic is $z = -1.37$.

- (a) Is the result significant at the 5% level?
(b) Is the result significant at the 1% level?
6.58 Explain in plain language why a significance test that is significant at the 1% level must always be significant at the 5% level.

- 6.59 You have performed a two-sided test of significance and obtained a value of $z = 3.3$. Use Table D to find the approximate P -value for this test.
- 6.60 You have performed a one-sided test of significance and obtained a value of $z = 0.215$. Use Table D to find the approximate P -value for this test.
- 6.61 You will perform a significance test of $H_0: \mu = 0$ versus $H_a: \mu > 0$.
- What values of z would lead you to reject H_0 at the 5% level?
 - If the alternative hypothesis was $H_a: \mu \neq 0$, what values of z would lead you to reject H_0 at the 5% level?
 - Explain why your answers to parts (a) and (b) are different.
- 6.62 Between what critical values from Table D does the P -value for the outcome $z = -1.37$ in Exercise 6.57 lie? Calculate the P -value using Table A, and verify that it lies between the values you found from Table D.
- 6.63 **Radon.** Radon is a colorless, odorless gas that is naturally released by rocks and soils and may concentrate in tightly closed houses. Because radon is slightly radioactive, there is some concern that it may be a health hazard. Radon detectors are sold to homeowners worried about this risk, but the detectors may be inaccurate. University researchers placed 12 detectors in a chamber where they were exposed to 105 picocuries per liter (pCi/l) of radon over 3 days.⁹ Here are the readings given by the detectors:

91.9 97.8 111.4 122.3 105.4 95.0
103.8 99.6 96.6 119.3 104.8 101.7

- Assume (unrealistically) that you know that the standard deviation of readings for all detectors of this type is $\sigma = 9$.
- Give a 95% confidence interval for the mean reading μ for this type of detector.
 - Is there significant evidence at the 5% level that the mean reading differs from the true value 105? State hypotheses and base a test on your confidence interval from (a).
- 6.64 **Clothing for runners.** Your company sells exercise clothing and equipment on the Internet. To design the clothing, you collect data on the physical characteristics of your different types of customers. Here are the weights for a sample of 24 male runners. Assume that these runners can be viewed as a random sample of your potential customers. The weights are expressed in kilograms.

67.8 61.9 63.0 53.1 62.3 59.7 55.4 58.9 60.9 69.2 63.7 68.3
64.7 65.6 56.0 57.8 66.0 62.9 53.6 65.0 55.8 60.4 69.3 61.7

- Exercise 6.12 (page 377) asks you to find a 95% confidence interval for the mean weight of the population of all such runners, assuming that the population standard deviation is $\sigma = 4.5$ kg.

- Give the confidence interval from that exercise, or calculate the interval if you did not do the exercise.
- Based on this confidence interval, does a test of

$$H_0: \mu = 61.3 \text{ kg}$$

$$H_a: \mu \neq 61.3 \text{ kg}$$

reject H_0 at the 5% significance level?

- (c) Would $H_0: \mu = 63$ be rejected at the 5% level if tested against a two-sided alternative?
- 6.65 **Cockroaches.** Your company is developing a better means to eliminate cockroaches from buildings. In the process, your R&D team studies the absorption of sugar by these insects.¹⁰ They feed cockroaches a diet containing measured amounts of a particular sugar. After 10 hours, the cockroaches are killed and the concentration of the sugar in various body parts is determined by a chemical analysis. The paper that reports the research states that a 95% confidence interval for the mean amount (in milligrams) of the sugar in the hindguts of the cockroaches is 4.2 ± 2.3 .
- Does this paper give evidence that the mean amount of sugar in the hindguts under these conditions is not equal to 7 mg? State H_0 and H_a and base a test on the confidence interval.
 - Would the hypothesis that $\mu = 5$ mg be rejected at the 5% level in favor of a two-sided alternative?
- 6.66 **Market pioneers.** Market pioneers, companies that are among the first to develop a new product or service, tend to have higher market shares than latecomers to the market. What accounts for this advantage? Here is an excerpt from the conclusions of a study of a sample of 1209 manufacturers of industrial goods:
- Can patent protection explain pioneer share advantages? Only 21% of the pioneers claim a significant benefit from either a product patent or a trade secret. Though their average share is two points higher than that of pioneers without this benefit, the increase is not statistically significant ($z = 1.13$). Thus, at least in mature industrial markets, product patents and trade secrets have little connection to pioneer share advantages.¹¹*

Find the P -value for the given z . Then explain to someone who knows no statistics what “not statistically significant” in the study’s conclusion means. Why does the author conclude that patents and trade secrets don’t help, even though they contributed 2 percentage points to average market share?

6.3 Using Significance Tests

Significance tests are widely used in reporting the results of research in many fields of applied science and in industry. New pharmaceutical products require significant evidence of effectiveness and safety. Courts inquire about statistical significance in hearing class-action discrimination cases. Marketers want to know whether a new ad campaign significantly outperforms the old one, and medical researchers want to know whether a new therapy performs significantly better. In all these uses, statistical significance is valued because it points to an effect that is unlikely to occur simply by chance.

Carrying out a test of significance is often quite simple, especially if you get a P -value effortlessly from a calculator or computer. Using tests wisely is not so simple. Here are some points to keep in mind when using or interpreting significance tests.

How small a P is convincing?

The purpose of a test of significance is to describe the degree of evidence provided by the sample against the null hypothesis. The P -value does this. But how small a P -value is convincing evidence against the null hypothesis? This depends mainly on two circumstances:

- *How plausible is H_0 ?* If H_0 represents an assumption that the people you must convince have believed for years, strong evidence (small P) will be needed to persuade them.
- *What are the consequences of rejecting H_0 ?* If rejecting H_0 in favor of H_a means making an expensive changeover from one type of product packaging to another, you need strong evidence that the new packaging will boost sales.

These criteria are a bit subjective. Different people will often insist on different levels of significance. Giving the P -value allows each of us to decide individually if the evidence is sufficiently strong.

Users of statistics have often emphasized standard levels of significance such as 10%, 5%, and 1%. This emphasis reflects the time when tables of critical values rather than computer software dominated statistical practice. The 5% level ($\alpha = 0.05$) is particularly common. There is no sharp border between “significant” and “insignificant,” only increasingly strong evidence as the P -value decreases. There is no practical distinction between the P -values 0.049 and 0.051. It makes no sense to treat $P \leq 0.05$ as a universal rule for what is significant.

APPLY YOUR KNOWLEDGE

- 6.67 Is it significant? More than 200,000 people worldwide take the GMAT examination each year as they apply for MBA programs. Their scores vary Normally with mean about $\mu = 525$ and standard deviation about $\sigma = 100$. One hundred students go through a rigorous training program designed to raise their GMAT scores. Test the hypotheses

$$H_0 : \mu = 525$$

$$H_a : \mu > 525$$

in each of the following situations:

- (a) The students' average score is $\bar{x} = 541.4$. Is this result significant at the 5% level?
- (b) The average score is $\bar{x} = 541.5$. Is this result significant at the 5% level?

The difference between the two outcomes in (a) and (b) is of no importance. Beware attempts to treat $\alpha = 0.05$ as sacred.

Statistical significance and practical significance

When a null hypothesis (“no effect” or “no difference”) can be rejected at the usual levels, $\alpha = 0.05$ or $\alpha = 0.01$, there is good evidence that an effect is present. But that effect may be extremely small. When large

samples are available, even tiny deviations from the null hypothesis will be significant.

EXAMPLE 6.18**It's significant. So what?**

We are testing the hypothesis of no correlation between two variables. With 1000 observations, an observed correlation of only $r = 0.08$ is significant evidence at the $\alpha = 0.01$ level that the correlation in the population is not zero but positive. The low significance level does not mean there is a strong association, only that there is strong evidence of some association. The true population correlation is probably quite close to the observed sample value, $r = 0.08$. We might well conclude that for practical purposes we can ignore the association between these variables, even though we are confident (at the 1% level) that the correlation is positive.

Remember the wise saying: *statistical significance is not the same as practical significance*. On the other hand, if we fail to reject the null hypothesis, it may be because H_0 is true or because our sample size is insufficient to detect the alternative. Exercise 6.73 demonstrates in detail the effect on P of increasing the sample size.

The remedy for attaching too much importance to statistical significance is to pay attention to the actual experimental results as well as to the P -value. Plot your data and examine them carefully. Are there outliers or other deviations from a consistent pattern? A few outlying observations can produce highly significant results if you blindly apply common tests of significance. Outliers can also destroy the significance of otherwise convincing data. The foolish user of statistics who feeds the data to a computer without exploratory analysis will often be embarrassed. Is the effect that you are seeking visible in your plots? If not, ask yourself if the effect is large enough to be of practical importance. It is usually wise to give a confidence interval for the parameter in which you are interested. A confidence interval actually estimates the size of an effect rather than simply asking if it is too large to reasonably occur by chance alone. Confidence intervals are not used as often as they should be, while tests of significance are perhaps overused.

6.68 How far do rich parents take us? How much education children get is strongly associated with the wealth and social status of their parents. In social science jargon, this is “socioeconomic status,” or SES. But the SES of parents has little influence on whether children who have graduated from college go on to yet more education. One study looked at whether college graduates took the graduate admissions tests for business, law, and other graduate programs. The effects of the parents’ SES on taking the LSAT test for law school were “both statistically insignificant and small.”

- (a) What does “statistically insignificant” mean?
- (b) Why is it important that the effects were small in size as well as insignificant?

APPLY YOUR KNOWLEDGE

Statistical inference is not valid for all sets of data

We know that badly designed surveys or experiments often produce useless results. Formal statistical inference cannot correct basic flaws in the design. A statistical test is valid only in certain circumstances, with properly produced data being particularly important. The z test, for example, should bear the same warning label that we attached on page 372 to the z confidence interval. Similar warnings accompany the other tests that we will learn.

Tests of significance and confidence intervals are based on the laws of probability. Randomization in sampling or experimentation ensures that these laws apply. But we must often analyze data that do not arise from randomized samples or experiments. To apply statistical inference to such data, we must have confidence in a probability model for the data. The diameters of successive holes bored in auto engine blocks during production, for example, may behave like independent observations on a Normal distribution. We can check this probability model by examining the data. If the model appears correct, we can apply the recipes of this chapter to do inference about the process mean diameter μ . Do ask how the data were produced, and don't be too impressed by P -values on a printout until you are confident that the data deserve a formal analysis.

6.69 Give an example of a set of data for which statistical inference is not valid.

APPLY YOUR KNOWLEDGE**Beware of searching for significance**

Statistical significance ought to mean that you have found an effect that you were looking for. The reasoning behind statistical significance works well if you decide what effect you are seeking, design a study to search for it, and use a test of significance to weigh the evidence you get. In other settings, significance may have little meaning.

EXAMPLE 6.19**Cell phones and brain cancer**

Might the radiation from cell phones be harmful to users? Many studies have found little or no connection between using cell phones and various illnesses. Here is part of a news account of one study:

A hospital study that compared brain cancer patients and a similar group without brain cancer found no statistically significant association between cell phone use and a group of brain cancers known as gliomas. But when 20 types of glioma were considered separately, an association was found between phone use and one rare form. Puzzlingly, however, this risk appeared to decrease rather than increase with greater mobile phone use.¹²

Think for a moment: Suppose that the 20 null hypotheses for these 20 significance tests are all true. Then each test has a 5% chance of being significant at the 5% level. That's what $\alpha = 0.05$ means—results this extreme occur only 5% of the time just by chance when the null hypothesis is true. Because 5% is 1/20, we expect about 1 of 20 tests to give a significant result just by chance. Running 1 test

and reaching the $\alpha = 0.05$ level is reasonably good evidence that you have found something; running 20 tests and reaching that level only once is not.



The peril of multiple analyses is increased now that a few simple commands will set software to work performing all manner of complicated tests and operations on your data. We will state it as a law that any large set of data—even several pages of a table of random digits—contains some unusual pattern. Sufficient computer time will discover that pattern, and when you test specifically for the pattern that turned up, the result will be significant. That's much like testing for 20 kinds of cancer and finding one test significant. Significance levels assume that you chose one hypothesis before searching your data and then tested that one hypothesis.

Searching data for suggestive patterns is certainly legitimate. Exploratory data analysis is an important part of statistics. But the reasoning of formal inference does not apply when your search for a striking effect in the data is successful. The remedy is clear. Once you have a hypothesis, design a study to search specifically for the effect you now think is there. If the result of this study is statistically significant, you have real evidence.

6.70 Predicting success of trainees. What distinguishes managerial trainees who eventually become executives from those who, after expensive training, don't succeed and leave the company? We have abundant data on past trainees—data on their personalities and goals, their college preparation and performance, even their family backgrounds and their hobbies. Statistical software makes it easy to perform dozens of significance tests on these dozens of variables to see which ones best predict later success. We find that future executives are significantly more likely than washouts to have an urban or suburban upbringing and an undergraduate degree in a technical field.

Explain clearly why using these “significant” variables to select future trainees is not wise. Then suggest a follow-up study using this year's trainees as subjects that should clarify the importance of the variables identified by the first study.

APPLY YOUR KNOWLEDGE

SECTION 6.3 SUMMARY

- *P*-values are more informative than the reject-or-not result of a fixed level α test. Beware of placing too much weight on traditional values of α , such as $\alpha = 0.05$.
- Very small effects can be highly significant (small *P*), especially when a test is based on a large sample. A statistically significant effect need not be practically important. Plot the data to display the effect you are seeking, and use confidence intervals to estimate the actual value of parameters.
- On the other hand, lack of significance does not imply that H_0 is true, especially when the test is based on a small sample.

■ Significance tests are not always valid. Faulty data collection, outliers in the data, and testing a hypothesis on the same data that suggested the hypothesis can invalidate a test. Many tests run at once will probably produce some significant results by chance alone, even if all the null hypotheses are true.

SECTION 6.3 EXERCISES

- 6.71 **Significance tests.** Which of the following questions does a test of significance answer?
- Is the sample or experiment properly designed?
 - Is the observed effect due to chance?
 - Is the observed effect important?
- 6.72 **Vitamin C and colds.** In a study of the suggestion that taking vitamin C will prevent colds, 400 subjects are assigned at random to one of two groups. The experimental group takes a vitamin C tablet daily, while the control group takes a placebo. At the end of the experiment, the researchers calculate the difference between the percents of subjects in the two groups who were free of colds. This difference is statistically significant ($P = 0.03$) in favor of the vitamin C group. Can we conclude that vitamin C has a strong effect in preventing colds? Explain your answer.
- 6.73 **Coaching for the SAT.** Every user of statistics should understand the distinction between statistical significance and practical importance. A sufficiently large sample will declare very small effects statistically significant. Let us suppose that SAT mathematics (SATM) scores in the absence of coaching vary Normally with mean $\mu = 515$ and $\sigma = 100$. Suppose further that coaching may change μ but does not change σ . An increase in the SATM score from 515 to 518 is of no importance in seeking admission to college, but this unimportant change can be statistically very significant. To see this, calculate the P -value for the test of

$$H_0: \mu = 515$$

$$H_a: \mu > 515$$

in each of the following situations:

- A coaching service coaches 100 students; their SATM scores average $\bar{x} = 518$.
 - By the next year, the service has coached 1000 students; their SATM scores average $\bar{x} = 518$.
 - An advertising campaign brings the number of students coached to 10,000; their average score is still $\bar{x} = 518$.
- 6.74 **Coaching for the SAT.** Give a 99% confidence interval for the mean SATM score μ after coaching in each part of the previous exercise. For large samples, the confidence interval says, “Yes, the mean score is higher after coaching, but only by a small amount.”
- 6.75 **Student loan poll.** A local television station announces a question for a call-in opinion poll on the six o’clock news and then gives the response on the eleven o’clock news. Today’s question concerns a proposed increase in funds

for student loans. Of the 2372 calls received, 1921 oppose the increase. The station, following standard statistical practice, makes a confidence statement: “81% of the Channel 13 Pulse Poll sample oppose the increase. We can be 95% confident that the proportion of all viewers who oppose the increase is within 1.6% of the sample result.” Is the station’s conclusion justified? Explain your answer.

- 6.76 **Extrasensory perception.** A researcher looking for evidence of extrasensory perception (ESP) tests 500 subjects. Four of these subjects do significantly better ($P < 0.01$) than random guessing.

- Is it proper to conclude that these four people have ESP? Explain your answer.
- What should the researcher now do to test whether any of these four subjects have ESP?

- 6.77 **Bonferroni procedure** 6.77 **More than one test.** A P -value based on a single test is misleading if you perform several tests. The **Bonferroni procedure** gives a significance level for several tests together. Level α then means that if *all* the null hypotheses are true, the probability is α that *any* of the tests rejects its null hypothesis.

If you perform 2 tests and want to use the $\alpha = 5\%$ significance level, Bonferroni says to require a P -value of $0.05/2 = 0.025$ to declare either one of the tests significant. In general, if you perform k tests and want protection at level α , use α/k as your cutoff for statistical significance for each test.

You perform 6 tests and obtain individual P -values of 0.476, 0.032, 0.241, 0.008, 0.010, and 0.001. Which of these are statistically significant using the Bonferroni procedure with $\alpha = 0.05$?

- 6.78 **More than one test.** Refer to the previous problem. A researcher has performed 12 tests of significance and wants to apply the Bonferroni procedure with $\alpha = 0.05$. The calculated P -values are 0.041, 0.569, 0.050, 0.416, 0.001, 0.004, 0.256, 0.041, 0.888, 0.010, 0.002, and 0.433. Which of these tests reject their null hypotheses with this procedure?

- 6.79 **Many tests.** Long ago, a group of psychologists carried out 77 separate significance tests and found that 2 were significant at the 5% level. Suppose that these tests are independent of each other. (In fact, they were not independent, because all involved the same subjects.) If all of the null hypotheses are true, each test has probability 0.05 of being significant at the 5% level.

- What is the distribution of the number X of tests that are significant?
- Find the probability that 2 or more of the tests are significant.

6.4 Power and Inference as a Decision*

Although we prefer to use P -values rather than the reject-or-not view of the fixed α significance test, the latter view is important for planning statistical

*Although the topics in this section are important in planning and interpreting significance tests, they can be omitted without loss of continuity.

studies. We will first explain why, then discuss different views of statistical tests.

The power of a statistical test

In examining the usefulness of a confidence interval, we are concerned with both the level of confidence and the margin of error. The confidence level tells us how reliable the method is in repeated use. The margin of error tells us how sensitive the method is, that is, how closely the interval pins down the parameter being estimated. Fixed level α significance tests are closely related to confidence intervals—in fact, we saw that a two-sided test can be carried out directly from a confidence interval. The significance level, like the confidence level, says how reliable the method is in repeated use. If we use 5% significance tests repeatedly when H_0 is in fact true, we will be wrong (the test will reject H_0) 5% of the time and right (the test will fail to reject H_0) 95% of the time.

High confidence is of little value if the interval is so wide that few values of the parameter are excluded. Similarly, a test with a small level of α is of little value if it almost never rejects H_0 even when the true parameter value is far from the hypothesized value. We must be concerned with the ability of a test to detect that H_0 is false, just as we are concerned with the margin of error of a confidence interval. This ability is measured by the probability that the test will reject H_0 when an alternative is true. The higher this probability is, the more sensitive the test is.

POWER

The probability that a fixed level α significance test will reject H_0 when a particular alternative value of the parameter is true is called the **power** of the test.

EXAMPLE 6.20

Does exercise make strong bones?

Can a 6-month exercise program increase the total body bone mineral content (TBBMC) of young women? A team of researchers is planning a study to examine this question. Based on the results of a previous study, they are willing to assume that $\sigma = 2$ for the percent change in TBBMC over the 6-month period. A change in TBBMC of 1% would be considered important, and the researchers would like to have a reasonable chance of detecting a change this large or larger. Are 25 subjects a large enough sample for this project?



We will answer this question by calculating the power of the significance test that will be used to evaluate the data to be collected. The calculation consists of three steps:

1. State H_0 , H_a , the particular alternative we want to detect, and the significance level α .

2. Find the values of \bar{x} that will lead us to reject H_0 .
3. Calculate the probability of observing these values of \bar{x} when the alternative is true.

Step 1 The null hypothesis is that the exercise program has no effect on TBBMC. In other words, the mean percent change is zero. The alternative is that exercise is beneficial; that is, the mean change is positive. Formally, we have

$$H_0: \mu = 0$$

$$H_a: \mu > 0$$

The alternative of interest is $\mu = 1\%$ increase in TBBMC. A 5% test of significance will be used.

Step 2 The z test rejects H_0 at the $\alpha = 0.05$ level whenever

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 0}{2/\sqrt{25}} \geq 1.645$$

Be sure you understand why we use 1.645. Rewrite this in terms of \bar{x} :

$$\text{reject } H_0 \text{ when } \bar{x} \geq 1.645 \frac{2}{\sqrt{25}}$$

$$\text{reject } H_0 \text{ when } \bar{x} \geq 0.658$$

Because the significance level is $\alpha = 0.05$, this event has probability 0.05 of occurring *when the population mean μ is 0*.

Step 3 The power for the alternative $\mu = 1\%$ is the probability that H_0 will be rejected *when in fact $\mu = 1$* . We calculate this probability by standardizing \bar{x} , using the value $\mu = 1$, the population standard deviation $\sigma = 2$, and the sample size $n = 25$. The power is

$$\begin{aligned} P(\bar{x} \geq 0.658 \text{ when } \mu = 1) &= P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{0.658 - 1}{2/\sqrt{25}}\right) \\ &= P(Z \geq -0.855) = 0.80 \end{aligned}$$

Figure 6.14 illustrates the power with the sampling distribution of \bar{x} when $\mu = 1$. This significance test rejects the null hypothesis that exercise has no effect on TBBMC 80% of the time if the true effect of exercise is a 1% increase in TBBMC. If the true effect of exercise is greater than a 1% increase, the test will have greater power; it will reject with a higher probability.

High power is desirable. Along with 95% confidence intervals and 5% significance tests, 80% power is becoming a standard. Many U.S. government agencies that provide research funds require that the sample size for the funded studies be sufficient to detect important results 80% of the time using a 5% test of significance.

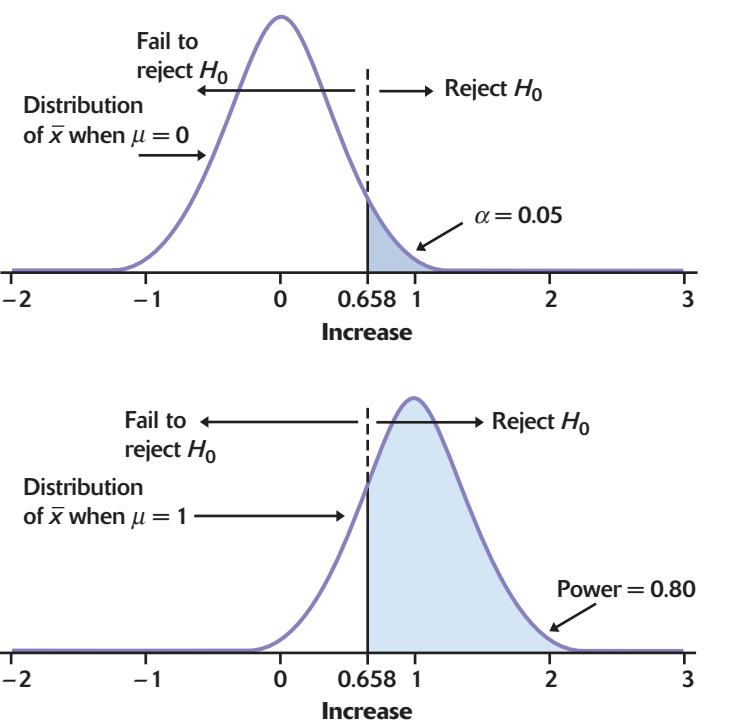


FIGURE 6.14 The sampling distributions of \bar{x} when $\mu = 0$ and when $\mu = 1$ with α and the power. Power is the probability that the test rejects H_0 when the alternative is true.

Increasing the power

Suppose you have performed a power calculation and found that the power is too small. What can you do to increase it? Here are four ways:

- Increase α . A 5% test of significance will have a greater chance of rejecting the alternative than a 1% test because the strength of evidence required for rejection is less.
- Consider a particular alternative that is farther away from μ_0 . Values of μ that are in H_a but lie close to the hypothesized value μ_0 are harder to detect (lower power) than values of μ that are far from μ_0 .
- Increase the sample size. More data will provide more information about \bar{x} so we have a better chance of distinguishing values of μ .
- Decrease σ . This has the same effect as increasing the sample size: more information about μ . Improving the measurement process and restricting attention to a subpopulation are two common ways to decrease σ .

Power calculations are important in planning studies. Using a significance test with low power makes it unlikely that you will find a significant effect even if the truth is far from the null hypothesis. A null hypothesis that is in fact false can become widely believed if repeated attempts to find evidence against it fail because of low power. The following example illustrates this point.

EXAMPLE 6.21**Are stock markets efficient?**

The “efficient market hypothesis” for stock prices says that future stock prices show only random variation. No information available now will help us predict stock prices in the future, because the efficient working of the market has already incorporated all available information in the present price. Many studies tested the claim that one or another kind of information is helpful. In these studies, the efficient market hypothesis is H_0 , and the claim that prediction is possible is H_a . Almost all studies failed to find good evidence against H_0 . As a result, the efficient market theory became quite popular.

An examination of the significance tests employed found that the power was generally low. Failure to reject H_0 when using tests of low power is not evidence that H_0 is true. As one expert said, “The widespread impression that there is strong evidence for market efficiency may be due just to a lack of appreciation of the low power of many statistical tests.”¹³ More careful studies later showed that the size of a company and measures of value such as the ratio of stock price to earnings do help predict future stock price movements.



Here is another example of a power calculation, this time for a two-sided z test.

EXAMPLE 6.22**Power for a two-sided test**

Example 6.15 (page 395) presented a test of

$$\begin{aligned} H_0: \mu &= 0.86 \\ H_a: \mu &\neq 0.86 \end{aligned}$$

at the 1% level of significance. What is the power of this test against the specific alternative $\mu = 0.845$?

The test rejects H_0 when $|z| \geq 2.576$. The test statistic is

$$z = \frac{\bar{x} - 0.86}{0.0068/\sqrt{3}}$$

Some arithmetic shows that the test rejects when either of the following is true:

$$\begin{aligned} z &\geq 2.576 \quad (\text{in other words, } \bar{x} \geq 0.870) \\ z &\leq -2.576 \quad (\text{in other words, } \bar{x} \leq 0.850) \end{aligned}$$

These are disjoint events, so the power is the sum of their probabilities, *computed assuming that the alternative $\mu = 0.845$ is true*. We find that

$$P(\bar{x} \geq 0.87) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{0.87 - 0.845}{0.0068/\sqrt{3}}\right)$$

$$= P(Z \geq 6.37) \doteq 0$$

$$P(\bar{x} \leq 0.85) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq \frac{0.85 - 0.845}{0.0068/\sqrt{3}}\right)$$

$$= P(Z \leq 1.27) = 0.8980$$

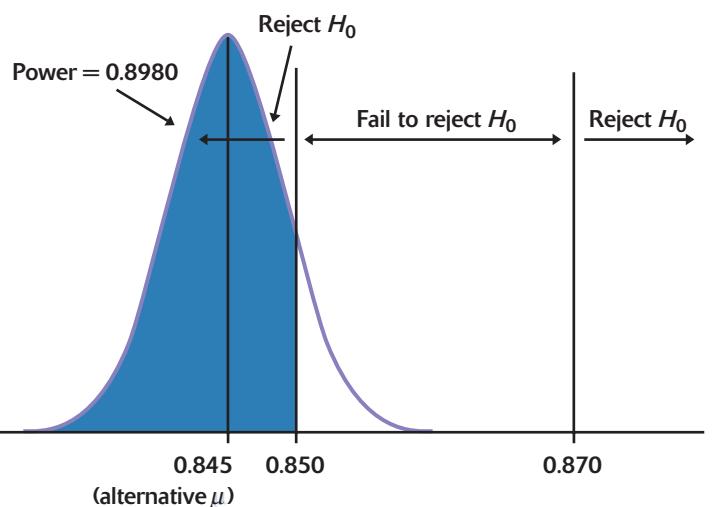


FIGURE 6.15 The power for Example 6.22.

Figure 6.15 illustrates this calculation. Because the power is about 0.9, we are quite confident that the test will reject H_0 when this alternative is true.



Inference as decision*

We have presented tests of significance as methods for assessing the strength of evidence against the null hypothesis. This assessment is made by the P -value, which is a probability computed under the assumption that H_0 is true. The alternative hypothesis (the statement we seek evidence for) enters the test only to show us what outcomes count against the null hypothesis. Most users of statistics think of tests in this way.

But signs of another way of thinking were present in our discussion of significance tests with fixed level α . A level of significance α chosen in advance points to the outcome of the test as a *decision*. If the P -value is less than α , we reject H_0 in favor of H_a . Otherwise we fail to reject H_0 . There is a big distinction between measuring the strength of evidence and making a decision. Many statisticians feel that making decisions is too grand a goal for statistics alone. Decision makers should take account of the results of statistical studies, but their decisions are based on many additional factors that are hard to reduce to numbers.

Yet there are circumstances in which a statistical test leads directly to a decision. **Acceptance sampling** is one such circumstance. A producer of bearings and the consumer of the bearings agree that each carload lot shall meet certain quality standards. When a carload arrives, the consumer inspects a random sample of bearings. On the basis of the sample outcome, the consumer either accepts or rejects the carload. Some statisticians argue

*acceptance
sampling*

*The purpose of this section is to clarify the reasoning of significance tests by contrast with a related type of reasoning. It can be omitted without loss of continuity.

that if “decision” is given a broad meaning, almost all problems of statistical inference can be posed as problems of making decisions in the presence of uncertainty. We will not venture further into the arguments over how we ought to think about inference. We do want to show how a different concept—*inference as decision*—changes the reasoning used in tests of significance.

Two types of error

Tests of significance concentrate on H_0 , the null hypothesis. If a decision is called for, however, there is no reason to single out H_0 . There are simply two hypotheses, and we must accept one and reject the other. It is convenient to call the two hypotheses H_0 and H_a , but H_0 no longer has the special status (the statement we try to find evidence against) that it had in tests of significance. In the acceptance sampling problem, we must decide between

H_0 : the lot of bearings meets standards

H_a : the lot does not meet standards

on the basis of a sample of bearings.

We hope that our decision will be correct, but sometimes it will be wrong. There are two types of incorrect decisions. We can accept a bad lot of bearings, or we can reject a good lot. Accepting a bad lot injures the consumer, while rejecting a good lot hurts the producer. To help distinguish these two types of error, we give them specific names.

TYPE I AND TYPE II ERRORS

If we reject H_0 (accept H_a) when in fact H_0 is true, this is a Type I error.

If we accept H_0 (reject H_a) when in fact H_a is true, this is a Type II error.

The possibilities are summed up in Figure 6.16. If H_0 is true, our decision either is correct (if we accept H_0) or is a Type I error. If H_a is true, our

		Truth about the population	
		H_0 true	H_a true
Decision based on sample	Reject H_0	Type I error	Correct decision
	Accept H_0	Correct decision	Type II error

FIGURE 6.16 The two types of error in testing hypotheses.

		Truth about the lot	
		Does meet standards	Does not meet standards
Decision based on sample	Reject the lot	Type I error	Correct decision
	Accept the lot	Correct decision	Type II error

FIGURE 6.17 The two types of error in the acceptance sampling setting.

decision either is correct or is a Type II error. Only one error is possible at one time. Figure 6.17 applies these ideas to the acceptance sampling example.

Error probabilities

We can assess any rule for making decisions in terms of the probabilities of the two types of error. This is in keeping with the idea that statistical inference is based on probability. We cannot (short of inspecting the whole lot) guarantee that good lots of bearings will never be rejected and bad lots will never be accepted. But by random sampling and the laws of probability, we can say what the probabilities of both kinds of error are.

Significance tests with fixed level α give a rule for making decisions, because the test either rejects H_0 or fails to reject it. If we adopt the decision-making way of thought, failing to reject H_0 means deciding that H_0 is true. We can then describe the performance of a test by the probabilities of Type I and Type II errors.

EXAMPLE 6.23

Are the bearings acceptable?

The mean diameter of a type of bearing is supposed to be 2.000 centimeters (cm). The bearing diameters vary Normally with standard deviation $\sigma = 0.010$ cm. When a lot of the bearings arrives, the consumer takes an SRS of 5 bearings from the lot and measures their diameters. The consumer rejects the bearings if the sample mean diameter is significantly different from 2 at the 5% significance level.

This is a test of the hypotheses

$$H_0: \mu = 2$$

$$H_a: \mu \neq 2$$

To carry out the test, the consumer computes the z statistic:

$$z = \frac{\bar{x} - 2}{0.01/\sqrt{5}}$$

and rejects H_0 if

$$z < -1.96 \text{ or } z > 1.96$$

A Type I error is to reject H_0 when in fact $\mu = 2$.

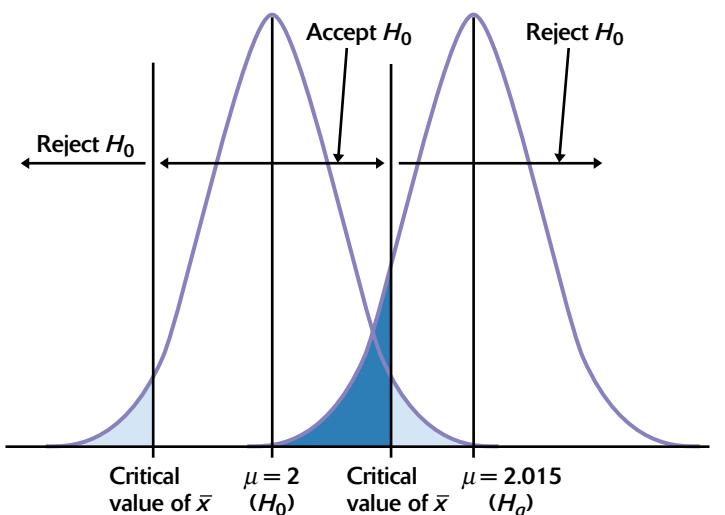


FIGURE 6.18 The two error probabilities for Example 6.23. The probability of a Type I error (light shaded area) is the probability of rejecting $H_0: \mu = 2$ when in fact $\mu = 2$. The probability of a Type II error (dark shaded area) is the probability of accepting H_0 when in fact $\mu = 2.015$.

What about Type II errors? Because there are many values of μ in H_a , we will concentrate on one value. The producer and the consumer agree that a lot of bearings with mean 0.015 cm away from the desired mean 2.000 should be rejected. So a particular Type II error is to accept H_0 when in fact $\mu = 2.015$.

Figure 6.18 shows how the two probabilities of error are obtained from the two sampling distributions of \bar{x} , for $\mu = 2$ and for $\mu = 2.015$. When $\mu = 2$, H_0 is true and to reject H_0 is a Type I error. When $\mu = 2.015$, accepting H_0 is a Type II error. We will now calculate these error probabilities.



The probability of a Type I error is the probability of rejecting H_0 when it is really true. In Example 6.23, this is the probability that $|z| \geq 1.96$ when $\mu = 2$. But this is exactly the significance level of the test. The critical value 1.96 was chosen to make this probability 0.05, so we do not have to compute it again. The definition of “significant at level 0.05” is that sample outcomes this extreme will occur with probability 0.05 when H_0 is true.

SIGNIFICANCE AND TYPE I ERROR

The significance level α of any fixed level test is the probability of a Type I error. That is, α is the probability that the test will reject the null hypothesis H_0 when H_0 is in fact true.

The probability of a Type II error for the particular alternative $\mu = 2.015$ in Example 6.23 is the probability that the test will fail to reject H_0 when μ has this alternative value. The *power* of the test against the alternative

$\mu = 2.015$ is just the probability that the test *does* reject H_0 . By following the method of Example 6.22, we can calculate that the power is about 0.92. The probability of a Type II error is therefore $1 - 0.92$, or 0.08.

POWER AND TYPE II ERROR

The power of a fixed level test against a particular alternative is 1 minus the probability of a Type II error for that alternative.

The two types of error and their probabilities give another interpretation of the significance level and power of a test. The distinction between tests of significance and tests as rules for deciding between two hypotheses does not lie in the calculations but in the reasoning that motivates the calculations. In a test of significance we focus on a single hypothesis (H_0) and a single probability (the P -value). The goal is to measure the strength of the sample evidence against H_0 . Calculations of power are done to check the sensitivity of the test. If we cannot reject H_0 , we conclude only that there is not sufficient evidence against H_0 , not that H_0 is actually true. If the same inference problem is thought of as a decision problem, we focus on two hypotheses and give a rule for deciding between them based on the sample evidence. We therefore must focus equally on two probabilities, the probabilities of the two types of error. We must choose one or the other hypothesis and cannot abstain on grounds of insufficient evidence.

The common practice of testing hypotheses

Such a clear distinction between the two ways of thinking is helpful for understanding. In practice, the two approaches often merge. We continued to call one of the hypotheses in a decision problem H_0 . The common practice of *testing hypotheses* mixes the reasoning of significance tests and decision rules as follows:

1. State H_0 and H_a just as in a test of significance.
2. Think of the problem as a decision problem, so that the probabilities of Type I and Type II errors are relevant.
3. Because of step 1, Type I errors are more serious. So choose an α (significance level) and consider only tests with probability of Type I error no greater than α .
4. Among these tests, select one that makes the probability of a Type II error as small as possible (that is, power as large as possible). If this probability is too large, you will have to take a larger sample to reduce the chance of an error.

Testing hypotheses may seem to be a hybrid approach. It was, historically, the effective beginning of decision-oriented ideas in statistics. An impressive mathematical theory of hypothesis testing was developed between 1928 and

1938 by Jerzy Neyman and Egon Pearson. The decision-making approach came later (1940s). Because decision theory in its pure form leaves you with two error probabilities and no simple rule on how to balance them, it has been used less often than either tests of significance or tests of hypotheses.

SECTION 6.4 SUMMARY

- The power of a significance test measures its ability to detect an alternative hypothesis. Power against a specific alternative is calculated as the probability that the test will reject H_0 when that alternative is true. This calculation requires knowledge of the sampling distribution of the test statistic under the alternative hypothesis. Increasing the size of the sample increases the power when the significance level remains fixed.
- In the case of testing H_0 versus H_a , decision analysis chooses a decision rule on the basis of the probabilities of two types of error. A Type I error occurs if H_0 is rejected when it is in fact true. A Type II error occurs if H_0 is accepted when in fact H_a is true.
- In a fixed level α significance test, the significance level α is the probability of a Type I error, and the power against a specific alternative is 1 minus the probability of a Type II error for that alternative.

SECTION 6.4 EXERCISES

- 6.80 Net income of banks. In Example 6.11, we ask if the net income of community banks has changed in the last year. If μ is the mean percent change for all such banks, the hypotheses are $H_0: \mu = 0$ and $H_a: \mu \neq 0$. The data come from an SRS of 110 banks and we assume that the population standard deviation is $\sigma = 26.4$. The test rejects H_0 at the 1% level of significance when $z \geq 2.576$ or $z < -2.576$, where

$$z = \frac{\bar{x} - 0}{26.4/\sqrt{110}}$$

Is this test sufficiently sensitive to usually detect a percent change in net income of 5%? Answer this question by calculating the power of the test against the alternative $\mu = 5$.

- 6.81 Fill the bottles. Example 6.17 discusses a test of $H_0: \mu = 300$ against $H_a: \mu < 300$, where μ is the mean content of cola bottles, in milliliters (ml). The sample size is $n = 6$, and the population is assumed to have a Normal distribution with $\sigma = 3$. The test statistic is therefore

$$z = \frac{\bar{x} - 300}{3/\sqrt{6}}$$

The test rejects H_0 at the 5% level of significance if $z \leq -1.645$. Power calculations help us see how large a shortfall in bottle content the test can be expected to detect.

- Find the power of this test against the alternative $\mu = 298$ ml.
- Is the power against $\mu = 295$ ml higher or lower than the value you found in (a)? Why?

- 6.82 **Mail-order catalog sales.** You want to see if a redesign of the cover of a mail-order catalog will increase sales. A very large number of customers will receive the original catalog, and a random sample of customers will receive the one with the new cover. For planning purposes, you are willing to assume that the mean sales from the new catalog will be approximately Normal with $\sigma = 50$ dollars and that the mean for the original catalog will be $\mu = 25$ dollars. You decide to use a sample size of $n = 900$. You wish to test

$$H_0: \mu = 25$$

$$H_a: \mu > 25$$

You decide to reject H_0 if $\bar{x} > 26$ and to accept H_0 otherwise.

- (a) Find the probability of a Type I error, that is, the probability that your test rejects H_0 when in fact $\mu = 25$ dollars.
 - (b) Find the probability of a Type II error when $\mu = 28$ dollars. This is the probability that your test accepts H_0 when in fact $\mu = 28$.
 - (c) Find the probability of a Type II error when $\mu = 30$.
 - (d) The distribution of sales is not Normal because many customers buy nothing. Why is it nonetheless reasonable in this circumstance to assume that the mean will be approximately Normal?
- 6.83 **Fill the bottles.** Increasing the sample size increases the power of a test when the level α is unchanged. Suppose that in Exercise 6.81 a sample of 20 bottles rather than 6 bottles had been measured. The significance test still rejects H_0 when $z \leq -1.645$, but the z statistic is now

$$z = \frac{\bar{x} - 300}{3/\sqrt{20}}$$

Find the power of this test against the alternative $\mu = 298$. Compare your result with the power from Exercise 6.81.

- 6.84 **Net income of banks.** Use the result of Exercise 6.80 to give the probabilities of Type I and Type II errors for the test discussed there.
- 6.85 **Fill the bottles.** Use the result of Exercise 6.81(a) to give the probability of a Type I error and the probability of a Type II error for the test in that exercise.
- 6.86 **Decide.** You must decide which of two discrete distributions a random variable X has. We will call the distributions p_0 and p_1 . Here are the probabilities that they assign to the values x of X :

x	0	1	2	3	4	5	6
p_0	0.1	0.1	0.1	0.1	0.2	0.1	0.3
p_1	0.2	0.1	0.1	0.2	0.2	0.1	0.1

You have a single observation on X and wish to test

$$H_0: p_0 \text{ is correct}$$

$$H_a: p_1 \text{ is correct}$$

One possible decision procedure is to accept H_0 if $X = 4$ or $X = 6$ and reject H_0 otherwise.

- (a) Find the probability of a Type I error, that is, the probability that you reject H_0 when p_0 is the correct distribution.
- (b) Find the probability of a Type II error.
- 6.87 **A Web-based business.** You are in charge of marketing for a Web site that offers automated medical diagnoses. The program will scan the results of routine medical tests (pulse rate, blood pressure, urinalysis, etc.) and either clear the patient or refer the case to a doctor. You are marketing the program for use as part of a preventive-medicine system to screen many thousands of persons who do not have specific medical complaints. The program makes a decision about each patient.
- (a) What are the two hypotheses and the two types of error that the program can make? Describe the two types of error in terms of “false-positive” and “false-negative” test results.
- (b) The program can be adjusted to decrease one error probability at the cost of an increase in the other error probability. Which error probability would you choose to make smaller, and why? (This is a matter of judgment. There is no single correct answer.)
- 6.88 **(Optional) Acceptance sampling.** The acceptance sampling test in Example 6.23 has probability 0.05 of rejecting a good lot of bearings and probability 0.08 of accepting a bad lot. The consumer of the bearings may imagine that acceptance sampling guarantees that most accepted lots are good. Alas, it is not so. Suppose that 90% of all lots shipped by the producer are bad.
- (a) Draw a tree diagram for shipping a lot (the branches are “bad” and “good”) and then inspecting it (the branches at this stage are “accept” and “reject”).
- (b) Write the appropriate probabilities on the branches, and find the probability that a lot shipped is accepted.
- (c) Use the definition of conditional probability or Bayes’s formula (page 349) to find the probability that a lot is bad, given that the lot is accepted. This is the proportion of bad lots among the lots that the sampling plan accepts.

STATISTICS IN SUMMARY

Statistical inference draws conclusions about a population on the basis of sample data and uses probability to indicate how reliable the conclusions are. A confidence interval estimates an unknown parameter. A significance test shows how strong the evidence is for some claim about a parameter.

The probabilities in both confidence intervals and tests tell us what would happen if we used the recipe for the interval or test very many times. A confidence level is the probability that the recipe for a confidence interval actually produces an interval that contains the unknown parameter. A 95% confidence interval gives a correct result 95% of the time when we use it repeatedly. A *P*-value is the probability that the sample would produce a result at least as extreme as the observed result if the null hypothesis really were true. That is, a *P*-value tells us how surprising the observed outcome is. Very surprising outcomes (small *P*-values) are good evidence that the null hypothesis is not true.

These ideas are the foundation for the rest of this book. We will have much to say about many statistical methods and their use in practice. In every case, the basic reasoning of confidence intervals and significance tests remains the same. Here are the most important things you should be able to do after studying this chapter.

A. CONFIDENCE INTERVALS

1. State in nontechnical language what is meant by “95% confidence” or other statements of confidence in statistical reports.
2. Calculate a confidence interval for the mean μ of a Normal population with known standard deviation σ , using the recipe $\bar{x} \pm z^* \sigma / \sqrt{n}$.
3. Recognize when you can safely use this confidence interval recipe and when the data collection design or a small sample from a skewed population makes it inaccurate. Understand that the margin of error does not include the effects of undercoverage, nonresponse, or other practical difficulties.
4. Understand how the margin of error of a confidence interval changes with the sample size and the level of confidence C .
5. Find the sample size required to obtain a confidence interval of specified margin of error m when the confidence level and other information are given.

B. SIGNIFICANCE TESTS

1. State the null and alternative hypotheses in a testing situation when the parameter in question is a population mean μ .
2. Explain in nontechnical language the meaning of the P -value when you are given the numerical value of P for a test.
3. Calculate the one-sample z statistic and the P -value for both one-sided and two-sided tests about the mean μ of a Normal population.
4. Assess statistical significance at standard levels α , either by comparing P to α or by comparing z to standard Normal critical values z^* .
5. Recognize that significance testing does not measure the size or importance of an effect.
6. Recognize when you can use the z test and when the data collection design or a small sample from a skewed population makes it inappropriate.

CHAPTER 6 REVIEW EXERCISES

- 6.89 Company cash flow and investment. How much a company invests in its business depends on how much cash flow it has. What factors influence the relationship between cash flow and investment? Here’s a clever suggestion: If an industry has an active market in used equipment, investment will be less sensitive to cash flow (“lower elasticity” in economic jargon). Companies in these industries can borrow easily because lenders know they can sell the company’s equipment if it defaults. A study of 270 manufacturing

industries measured SHRUSED, the proportion of secondhand equipment in the industry's total investment. The study found that "industries with SHRUSED values above the median have smaller cash flow elasticities than those with lower SHRUSED values; the difference is significant at the 5% level in the full sample."¹⁴ Explain to someone who knows no statistics why this study gives good reason to think that an active used-equipment market really does change the relationship between cash flow and investment.

- 6.90 **Foreign investment and exchange rates.** We might suspect that foreign direct investment (FDI), in which U.S. companies buy or build facilities overseas, depends on the rate at which the dollar can be exchanged with foreign currencies. A study of 3036 FDI transactions found that "there is no statistically significant relationship between the level of the exchange rate and foreign investment relative to domestic investment."¹⁵

- (a) Explain this conclusion to someone who knows no statistics.
- (b) We are reasonably confident that if there were a relationship between FDI and exchange rate that was large enough to be of interest, this study would have found it. Why?

- 6.91 **Wine.** Many food products contain small quantities of substances that would give an undesirable taste or smell if they were present in large amounts. An example is the "off-odors" caused by sulfur compounds in wine. Oenologists (wine experts) have determined the odor threshold, the lowest concentration of a compound that the human nose can detect. For example, the odor threshold for dimethyl sulfide (DMS) is given in the oenology literature as 25 micrograms per liter of wine ($\mu\text{g/l}$). Untrained noses may be less sensitive, however. Here are the DMS odor thresholds for 10 beginning students of oenology:

31 31 43 36 23 34 32 30 20 24

Assume (this is not realistic) that the standard deviation of the odor threshold for untrained noses is known to be $\sigma = 7 \mu\text{g/l}$.

- (a) Make a stemplot to verify that the distribution is roughly symmetric with no outliers. (A Normal quantile plot confirms that there are no systematic departures from Normality.)
 - (b) Give a 95% confidence interval for the mean DMS odor threshold among all beginning oenology students.
 - (c) Are you convinced that the mean odor threshold for beginning students is higher than the published threshold, $25 \mu\text{g/l}$? Carry out a significance test to justify your answer.
- 6.92 **Annual household income.** A government report gives a 90% confidence interval for the 1999 median annual household income as $\$40,816 \pm \314 . This result was calculated by advanced methods from the Current Population Survey, a multistage random sample of about 50,000 households.
- (a) Would a 95% confidence interval be wider or narrower? Explain your answer.
 - (b) Would the null hypothesis that the 1999 median household income was $\$40,000$ be rejected at the 10% significance level in favor of the two-sided alternative?

- 6.93 **Annual household income.** Refer to the previous problem. Give a 90% confidence interval for the 1999 median *weekly* household income. Use 52.14 as the number of weeks in a year.
- 6.94 **Too much cellulose to be profitable?** Excess cellulose in alfalfa reduces the “relative feed value” of the product that will be fed to dairy cows. If the cellulose content is too high, the price will be lower and the producer will have less profit. An agronomist examines the cellulose content of one type of alfalfa hay. Suppose that the cellulose content in the population has standard deviation $\sigma = 8$ milligrams per gram (mg/g). A sample of 15 cuttings has mean cellulose content $\bar{x} = 145$ mg/g.
- Give a 90% confidence interval for the mean cellulose content in the population.
 - A previous study claimed that the mean cellulose content was $\mu = 140$ mg/g, but the agronomist believes that the mean is higher than that figure. State H_0 and H_a and carry out a significance test to see if the new data support this belief.
 - The statistical procedures used in (a) and (b) are valid when several assumptions are met. What are these assumptions?
- 6.95 **Where do you buy?** Consumers can purchase nonprescription medications at food stores, mass merchandise stores such as Kmart and Wal-Mart, or pharmacies. About 45% of consumers make such purchases at pharmacies. What accounts for the popularity of pharmacies, which often charge higher prices?
- A study examined consumers’ perceptions of overall performance of the three types of stores, using a long questionnaire that asked about such things as “neat and attractive store,” “knowledgeable staff,” and “assistance in choosing among various types of nonprescription medication.” A performance score was based on 27 such questions. The subjects were 201 people chosen at random from the Indianapolis telephone directory. Here are the means and standard deviations of the performance scores for the sample:¹⁶

Store type	\bar{x}	s
Food stores	18.67	24.95
Mass merchandisers	32.38	33.37
Pharmacies	48.60	35.62

We do not know the population standard deviations, but a sample standard deviation s from so large a sample is usually close to σ . Use s in place of the unknown σ in this exercise.

- What population do you think that the authors of the study want to draw conclusions about? What population are you certain that they can draw conclusions about?
- Give 95% confidence intervals for the mean performance for each type of store.
- Based on these confidence intervals, are you convinced that consumers think that pharmacies offer higher performance than the other types of stores? Note that in Chapter 12 we will study a statistical method for comparing means of several groups.

- 6.96 **CEO pay.** A study of the pay of corporate chief executive officers (CEOs) examined the increase in cash compensation of the CEOs of 104 companies, adjusted for inflation, in a recent year. The mean increase in real compensation was $\bar{x} = 6.9\%$, and the standard deviation of the increases was $s = 55\%$. Is this good evidence that the mean real compensation μ of all CEOs increased that year? The hypotheses are

$$H_0: \mu = 0 \quad (\text{no increase})$$

$$H_a: \mu > 0 \quad (\text{an increase})$$

Because the sample size is large, the sample s is close to the population σ , so take $\sigma = 55\%$.

- (a) Sketch the Normal curve for the sampling distribution of \bar{x} when H_0 is true. Shade the area that represents the P -value for the observed outcome $\bar{x} = 6.9\%$.
- (b) Calculate the P -value.
- (c) Is the result significant at the $\alpha = 0.05$ level? Do you think the study gives strong evidence that the mean compensation of all CEOs went up?
- 6.97 **Large samples.** Statisticians prefer large samples. Describe briefly the effect of increasing the size of a sample (or the number of subjects in an experiment) on each of the following:
- (a) The width of a level C confidence interval.
- (b) The P -value of a test, when H_0 is false and all facts about the population remain unchanged as n increases.
- (c) The power of a fixed level α test, when α , the alternative hypothesis, and all facts about the population remain unchanged.
- 6.98 **Roulette.** A roulette wheel has 18 red slots among its 38 slots. You observe many spins and record the number of times that red occurs. Now you want to use these data to test whether the probability of a red has the value that is correct for a fair roulette wheel. State the hypotheses H_0 and H_a that you will test. (We will describe the test for this situation in Chapter 8.)
- 6.99 **Significant.** When asked to explain the meaning of “statistically significant at the $\alpha = 0.05$ level,” a student says, “This means there is only probability 0.05 that the null hypothesis is true.” Is this a correct explanation of statistical significance? Explain your answer.
- 6.100 **Significant.** Another student, when asked why statistical significance appears so often in research reports, says, “Because saying that results are significant tells us that they cannot easily be explained by chance variation alone.” Do you think that this statement is correct? Explain your answer.
- 6.101 **Welfare reform.** A study compares two groups of mothers with young children who were on welfare two years ago. One group attended a voluntary training program offered free of charge at a local vocational school and advertised in the local news media. The other group did not choose to attend the training program. The study finds a significant difference ($P < 0.01$) between the proportions of the mothers in the two groups who are still on welfare. The difference is not only significant but quite large. The report says that with 95% confidence the percent of the nonattending group still

- on welfare is $21\% \pm 4\%$ higher than that of the group who attended the program. You are on the staff of a member of Congress who is interested in the plight of welfare mothers and who asks you about the report.
- (a) Explain briefly and in nontechnical language what “a significant difference ($P < 0.01$)” means.
 - (b) Explain clearly and briefly what “95% confidence” means.
 - (c) Is this study good evidence that requiring job training of all welfare mothers would greatly reduce the percent who remain on welfare for several years?
- 6.102 **Simulation.** Use a computer to generate $n = 5$ observations from the Normal distribution $N(20, 5)$ with mean 20 and standard deviation 5. Find the 95% confidence interval for μ . Repeat this process 100 times and then count the number of times that the confidence interval includes the value $\mu = 20$. Explain your results.
- 6.103 **Simulation.** Use a computer to generate $n = 5$ observations from the Normal distribution $N(20, 5)$ with mean 20 and standard deviation 5. Test $H_0: \mu = 20$ versus $H_a: \mu \neq 20$ at the $\alpha = 0.05$ significance level. Repeat this process 100 times and then count the number of times that you reject H_0 . Explain your results.
- 6.104 **Simulation.** Use the same procedure for generating data as in the previous exercise. Now test the null hypothesis that $\mu = 22.5$. Explain your results.
- 6.105 **Simulation.** Figure 6.2 (page 366) demonstrates the behavior of a confidence interval in repeated sampling by showing the results of 25 samples from the same population. Now you will do a similar demonstration, though in an artificial setting. Suppose that the net assets of a large population of community banks follow the Uniform distribution between -20 and 500 (in millions of dollars). Then the mean assets of these banks are $\mu = 240$ and the standard deviation is $\sigma = 150$.
- (a) Simulate the drawing of 25 SRSs of size $n = 100$ from this population.
 - (b) For calculating the confidence intervals, you are willing to assume that the sample means are approximately Normal. Explain why this is a reasonable assumption.
 - (c) The 50% confidence interval for the population mean μ has the form $\bar{x} \pm m$. What is the margin of error m ? (Use 150 for the standard deviation.)
 - (d) Use your software to calculate the 50% confidence interval for μ for each of your 25 samples. Verify the computer’s calculations by checking the interval given for the first sample against your result in (b). Use the \bar{x} reported by the software.
 - (e) How many of the 25 confidence intervals contain the true mean $\mu = 240$? If you repeated the simulation, would you expect exactly the same number of intervals to contain μ ? In a very large number of samples, what percent of the confidence intervals would contain μ ?
- 6.106 **Simulation.** In the previous exercise you simulated the assets of 25 SRSs of 100 community banks. Now use these samples to demonstrate the behavior of a significance test. We know that the population standard deviation is $\sigma = 150$ and we are willing to assume that the sample means are approximately Normal.

- (a) Use your software to carry out a test of

$$H_0: \mu = 240$$

$$H_a: \mu \neq 240$$

for each of the 25 samples.

- (b) Verify the computer's calculations by using Table A to find the *P*-value of the test for the first of your samples. Use the \bar{x} reported by your software.
- (c) How many of your 25 tests reject the null hypothesis at the $\alpha = 0.05$ significance level? (That is, how many have *P*-value 0.05 or smaller?)
- (d) Because the simulation was done with $\mu = 240$, samples that lead to rejecting H_0 produce the wrong conclusion. In a very large number of samples, what percent would falsely reject the hypothesis?



CHAPTER 6 CASE STUDY EXERCISES

CASE STUDY 6.1: Older customers in restaurants. Persons aged 55 and over represented 21.3% of the U.S. population in the year 2000. This group is expected to increase to 30.5% by 2025. In terms of actual numbers of people, the increase is from 58.6 million to 101.4 million. Restaurateurs have found this market to be important and would like to make their businesses attractive to older customers. One study used a questionnaire to collect data from people aged 50 and over.¹⁷ For one part of the analysis, individuals were classified into two age groups: 50–64 and 65–79. There were 267 people in the first group and 263 in the second. One set of items concerned ambience, menu design, and service. A series of questions was rated on a 1 to 5 scale with 1 representing "strongly disagree" and 5 representing "strongly agree." In some cases the wording of questions has been shortened in the table below. Here are the means:

Question	50–64	65–79
Ambience		
Most restaurants are too dark	2.75	2.93
Most restaurants are too noisy	3.33	3.43
Background music is often too loud	3.27	3.55
Restaurants are too smoky	3.17	3.12
Tables are too small	3.00	3.19
Tables are too close together	3.79	3.81
Menu design		
Print size is not large enough	3.68	3.77
Glare makes menus difficult to read	2.81	3.01
Colors of menus make them difficult to read	2.53	2.72
Service		
It is difficult to hear the service staff	2.65	3.00
I would rather be served than serve myself	4.23	4.14
I would rather pay the server than a cashier	3.88	3.48
Service is too slow	3.13	3.10

First examine the means of the people who are 50 to 64. Order the statements according to the means and describe the results. Then do the same for the older group. For each question compute the z statistic and the associated P -value for the comparison between the two groups. For these calculations you can assume that the denominator in the test statistic is 0.08, so z is simply the difference in the means divided by 0.08. Note that you are performing 13 significance tests in this exercise. Keep this in mind when you interpret your results. Write a report summarizing your work.

CASE STUDY 6.2: Accessibility concerns. Refer to the previous question. The questionnaire also asked about accessibility both inside the restaurant and outside. Analyze these data and write a report.

Question	50–64	65–79
Accessibility and comfort inside		
Most chairs are too small	2.49	2.56
Bench seats are usually too narrow	3.03	3.25
Salad bars and buffets are difficult to reach	3.04	3.09
Serving myself from salad bars and buffets is difficult	2.58	2.75
Floors around salad bars and buffets are often slippery	2.84	3.01
Aisles are too narrow	3.04	3.20
Bathroom stalls are too narrow	2.82	3.10
Outside accessibility		
Doors are too heavy	2.51	3.01
Curbs near entrance are difficult	2.54	3.07
Parking spaces are too narrow	2.83	3.16
Distance from parking lot is too far	2.33	2.64
Parking lots are too dark at night	2.84	3.26

Prelude

Diversify your investments

A basic principle of investing is “don’t put all your eggs in one basket.” In previous chapters we learned that \bar{x} has variance σ^2/n when the data are an SRS from a population. The variance goes down as the number of observations goes up. More general versions of the same theory tell us that averages of many random quantities are less variable than averages of only a few. This applies to investing: diversification reduces the variability of returns, and so reduces the investor’s risk. If you put all of your retirement money in the stock of one company, the company may be very successful and you will be rich. On the other hand, the company may go bankrupt and you will be very poor. If you invest in several stocks, you are unlikely to be very rich or very poor.

Reputable investment advisors accept the idea that diversification should guide the investment strategy of individuals who are investing for retirement. An investor with a portfolio of several hundred thousand dollars sued his broker and brokerage firm because they did not diversify his investments. The conflict was settled by an arbitration panel that gave “substantial damages” to the investor.¹ Statistical methods were used to show that this investor’s return was substantially worse than the Standard & Poor’s (S&P) average return.

Goal Output

CHAPTER 7

Inference for Distributions

- 7.1 Inference for the Mean
of a Population**
- 7.2 Comparing Two Means**
- 7.3 Optional Topics
in Comparing Distributions**

Introduction

We began our study of data analysis in Chapter 1 by learning graphical and numerical tools for describing the distribution of a single variable and for comparing several distributions. Our study of the practice of statistical inference begins in the same way, with inference about a single distribution and comparison of two distributions. Comparing more than two distributions requires more elaborate methods, which are presented in Chapters 15 and 16.

Two important aspects of any distribution are its center and spread. If the distribution is Normal, we describe its center by the mean μ and its spread by the standard deviation σ . In this chapter, we will meet confidence intervals and significance tests for inference about a population mean μ and for comparing the means of two populations. The previous chapter emphasized the reasoning of tests and confidence intervals. Now we emphasize statistical practice, so we no longer assume that population standard deviations are known. The t procedures for inference about means are among the most common statistical methods.

7.1

Inference for the Mean of a Population

Both confidence intervals and tests of significance for the mean μ of a Normal population are based on the sample mean \bar{x} , which estimates the unknown μ . The sampling distribution of \bar{x} depends on σ . This fact causes no difficulty when σ is known. When σ is unknown, however, we must estimate σ even though we are primarily interested in μ . The sample standard deviation s is used to estimate the population standard deviation σ .

The t distributions

Suppose that we have a simple random sample (SRS) of size n from a Normally distributed population with mean μ and standard deviation σ . The sample mean \bar{x} then has the Normal distribution with mean μ and standard deviation σ/\sqrt{n} . When σ is not known, we estimate it with the sample standard deviation s , and then we estimate the standard deviation of \bar{x} by s/\sqrt{n} . This quantity is called the *standard error* of the sample mean \bar{x} and we denote it by $SE_{\bar{x}}$.

STANDARD ERROR

When the standard deviation of a statistic is estimated from the data, the result is called the **standard error** of the statistic. The standard error of the sample mean is

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The term “standard error” is sometimes used for the actual standard deviation of a statistic, σ/\sqrt{n} in the case of \bar{x} . The estimated value s/\sqrt{n} is then called the “estimated standard error.” In this book we will use the term “standard error” only when the standard deviation of a statistic is estimated from the data. The term has this meaning in the output of many statistical computer packages and in reports of research in many fields that apply statistical methods.

The standardized sample mean, or one-sample z statistic,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

is the basis of the z procedures for inference about μ when σ is known. This statistic has the standard Normal distribution $N(0, 1)$. When we substitute the standard error s/\sqrt{n} for the standard deviation σ/\sqrt{n} of \bar{x} , the statistic does *not* have a Normal distribution. It has a distribution that is new to us, called a *t distribution*.

THE t DISTRIBUTIONS

Suppose that an SRS of size n is drawn from an $N(\mu, \sigma)$ population. Then the **one-sample t statistic**

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has the *t distribution* with $n - 1$ degrees of freedom.

degrees of freedom

There is a different t distribution for each sample size. A particular t distribution is specified by giving the **degrees of freedom**. The degrees of freedom for this t statistic come from the sample standard deviation s in the denominator of t . We saw in Chapter 1 (page 42) that s has $n - 1$ degrees of freedom. There are other t statistics with different degrees of freedom, some of which we will meet later in this chapter.

We use $t(k)$ to stand for the t distribution with k degrees of freedom.* The density curves of the $t(k)$ distributions are similar in shape to the standard Normal curve. That is, they are symmetric about 0 and are bell-shaped. The spread of the t distributions is a bit greater than that of the standard Normal distribution. This is due to the extra variability caused by substituting the random variable s for the fixed parameter σ . As the degrees of freedom k increase, the $t(k)$ density curve approaches the $N(0, 1)$ curve ever more closely. This reflects the fact that s approaches σ as the sample size increases.

*The t distributions were discovered in 1908 by William S. Gosset. Gosset was a statistician employed by the Guinness brewing company, which did not permit him to publish his discoveries under his own name. He therefore wrote under the pen name “Student.” The t distribution is often called “Student’s t ” in his honor.

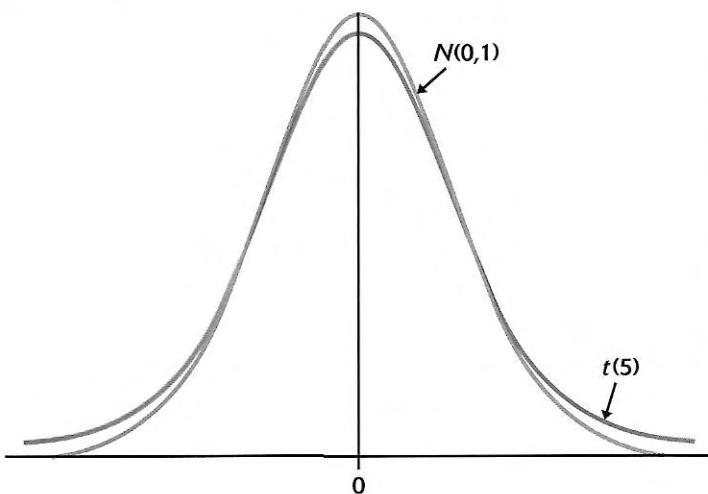


FIGURE 7.1 Density curves for the standard Normal and $t(5)$ distributions. Both are symmetric with center 0. The t distributions have more probability in their tails than does the standard Normal distribution.

Figure 7.1 compares the density curves of the standard Normal distribution and the t distribution with 5 degrees of freedom. The similarity in shape is apparent, as is the fact that the t distribution has more probability in the tails and less in the center than does the standard Normal distribution.

Table D in the back of the book gives critical values for the t distributions. For convenience, we have labeled the table entries by both the critical value p needed for significance tests and by the confidence level C (in percent) required for confidence intervals. The standard Normal critical values in the bottom row of entries are labeled z^* . As in the case of the Normal table (Table A), computer software often makes Table D unnecessary.

APPLY YOUR KNOWLEDGE

- 7.1 **Apartment rents.** Your local newspaper contains a large number of advertisements for unfurnished one-bedroom apartments. You choose 10 at random and calculate that their mean monthly rent is \$540 and that the standard deviation of their rents is \$80.
 - (a) What is the standard error of the mean?
 - (b) What are the degrees of freedom for a one-sample t statistic?
- 7.2 **Critical values.** What critical value would you use for a 95% confidence interval based on the $t(20)$ distribution? For a 99% confidence interval based on the $t(60)$ distribution?

The one-sample t confidence interval

With the t distributions to help us, we can analyze samples from Normal populations with unknown σ by replacing the standard deviation σ/\sqrt{n} of \bar{x} by its standard error s/\sqrt{n} in the z procedures of Chapter 6. The z statistic then becomes the one-sample t statistic. We must now employ P -values or critical values from t in place of the corresponding Normal values (z). Here are the details for the confidence interval.

THE ONE-SAMPLE t CONFIDENCE INTERVAL

Suppose that an SRS of size n is drawn from a population having unknown mean μ . A level C confidence interval for μ is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the value for the $t(n - 1)$ density curve with area C between $-t^*$ and t^* . The margin of error is

$$t^* \frac{s}{\sqrt{n}}$$

This interval is exact when the population distribution is Normal and is approximately correct for large n in other cases.

margin of error

The one-sample t confidence interval is similar in both reasoning and computational detail to the z confidence interval of Chapter 6. There, the margin of error for the population mean was $z^* \sigma / \sqrt{n}$. Here, we replace σ by its estimate s and z^* by t^* . So the margin of error for the population mean when we use the data to estimate σ is $t^* s / \sqrt{n}$.



CASE 7.1

PRODUCING A FORTIFIED FOOD PRODUCT

Many food products are fortified by adding nutrients, especially vitamins. In a recent year, the U.S. Agency for International Development purchased 238,300 metric tons of corn soy blend (CSB) for development programs and emergency relief in countries throughout the world. CSB is a highly nutritious, low-cost fortified food that is partially precooked and can be incorporated into different food preparations by the recipients. The addition of the nutrients is a difficult and important part of the production process. If too little is present, the product will be ineffective. If too much is present, consumption of the product could have some adverse effects. Vitamin C is an important nutrient used to fortify CSB. A study of the quality of the production process for CSB therefore measured vitamin C in specimens taken at the factory.²

EXAMPLE 7.1

CASE 7.1

Estimating the level of vitamin C

The following data are the amounts of vitamin C, measured in milligrams per 100 grams (mg/100 g) of blend (dry basis), for a random sample of size 8 from a production run:

26 31 23 22 11 22 14 31

We want to find a 95% confidence interval for μ , the mean vitamin C content of the CSB produced during this run.

Check that the sample mean is $\bar{x} = 22.50$ and the standard deviation is $s = 7.19$ with degrees of freedom $n - 1 = 7$. The standard error of \bar{x} is

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{7.19}{\sqrt{8}} = 2.54$$

From Table D we find $t^* = 2.365$. The 95% confidence interval is

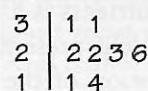
df = 7		
t^*	1.895	2.365
C	90%	95%

$$\begin{aligned}\bar{x} \pm t^* \frac{s}{\sqrt{n}} &= 22.50 \pm 2.365 \frac{7.19}{\sqrt{8}} \\ &= 22.5 \pm (2.365)(2.54) \\ &= 22.5 \pm 6.0 \\ &= (16.5, 28.5)\end{aligned}$$

We are 95% confident that the mean vitamin C content of the CSB for this run is between 16.5 and 28.5 mg/100 g.

In this example we have given the actual interval (16.5, 28.5) as our answer. Sometimes, we prefer to report the mean and margin of error: the mean vitamin C content is 22.5 mg/100 g with a margin of error of 6.0 mg/100 g.

The use of the t confidence interval in Example 7.1 rests on conditions that cannot easily be checked but are reasonable in this case. The samples from the production run were taken at regular time intervals, and experience has shown that this procedure gives data that can be treated as random samples from a Normal population. The condition that the population distribution is Normal cannot be effectively checked with only 8 observations. On the other hand, when we make a stemplot, we can see clearly that there are no outliers:



- APPLY YOUR KNOWLEDGE**
- 7.3 **Apartment rents.** You want to rent an unfurnished one-bedroom apartment for next semester. You take a random sample of 10 apartments advertised in the local newspaper and record the rental rates. Here are the rents (in dollars per month):

500, 650, 600, 505, 450, 550, 515, 495, 650, 395

Find a 95% confidence interval for the mean monthly rent for unfurnished one-bedroom apartments available for rent in this community.

- 7.4 If you chose 99% rather than 95% confidence, would your margin of error in the previous exercise be larger or smaller? Explain your answer and verify it by doing the calculations.

The one-sample t test

In tests as in confidence intervals, we allow for unknown σ by using the standard error and replacing z by t . Here are the details.

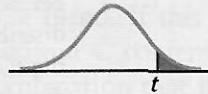
THE ONE-SAMPLE t TEST

Suppose that an SRS of size n is drawn from a population having unknown mean μ . To test the hypothesis $H_0: \mu = \mu_0$ based on an SRS of size n , compute the one-sample t statistic.

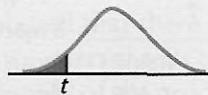
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

In terms of a random variable T having the $t(n - 1)$ distribution, the P -value for a test of H_0 against

$$H_a: \mu > \mu_0 \text{ is } P(T \geq t)$$



$$H_a: \mu < \mu_0 \text{ is } P(T \leq t)$$



$$H_a: \mu \neq \mu_0 \text{ is } 2P(T \geq |t|)$$



These P -values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

EXAMPLE 7.2

Is the vitamin C level correct?



The specifications for the CSB described in Case 7.1 state that the mixture should contain 2 pounds of vitamin premix for every 2000 pounds of product. These specifications are designed to produce a mean (μ) vitamin C content in the final product of 40 mg/100 g. We can test a null hypothesis that the mean vitamin C content of the production run in Example 7.1 conforms to these specifications. Specifically, we test

$$H_0: \mu = 40$$

$$H_a: \mu \neq 40$$

Recall that $n = 8$, $\bar{x} = 22.50$, and $s = 7.19$. The t test statistic is

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{22.5 - 40}{7.19/\sqrt{8}} \\ &= -6.88 \end{aligned}$$

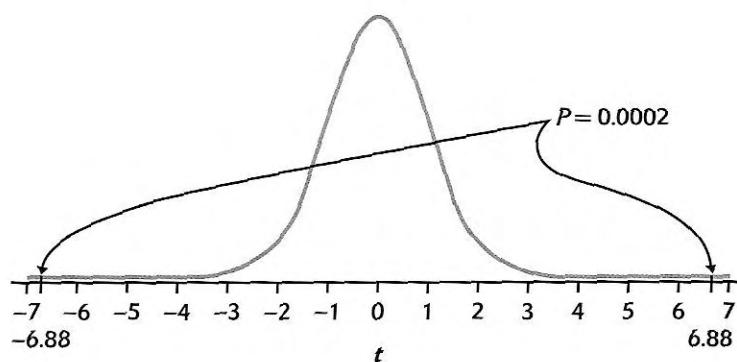


FIGURE 7.2 The P -value for Example 7.2.

Because the degrees of freedom are $n - 1 = 7$, this t statistic has the $t(7)$ distribution. Figure 7.2 shows that the P -value is $2P(T \geq 6.88)$, where T has the $t(7)$ distribution. From the largest entry in the $df = 7$ line of Table D we see that $P(T \geq 5.408) = 0.0005$.

We conclude that the P -value is less than 2×0.0005 , or $P < 0.001$. Software gives a more exact value as $P = 0.0002$. Clearly, these data are incompatible with a process mean of $\mu = 40$ mg/100 g. We reject H_0 and conclude that the vitamin C content for this run is below the specifications. Changes are needed to improve the process.

$df = 7$		
p	0.001	0.0005
t^*	4.785	5.408

In Example 7.2 we tested the specifications $\mu = 40$ mg/100 g against the two-sided alternative $\mu \neq 40$ mg/100 g because we had no prior suspicion that the production run would produce CSB with too much or too little vitamin C. If we knew that the proper amount of vitamin C was added to the mixture but suspected that some of the vitamin might be lost or destroyed by the production process, we could use a one-sided test. It is *wrong*, however, to examine the data first and then decide to do a one-sided test in the direction indicated by the data. If in doubt, use a two-sided test.

EXAMPLE 7.3



Has vitamin C been lost in production?

To test whether the vitamin C content is low, perhaps because vitamin C is lost or destroyed in production, our hypotheses are

$$H_0: \mu = 40$$

$$H_a: \mu < 40$$

The t test statistic does not change: $t = -6.88$. As Figure 7.3 illustrates, however, the P -value is now $P(T \leq -6.88)$, half of the value in the previous exercise. From Table D we can determine that $P < 0.0005$; software gives the exact value as $P = 0.0001$. We conclude that the production process has lost or destroyed some of the vitamin C.

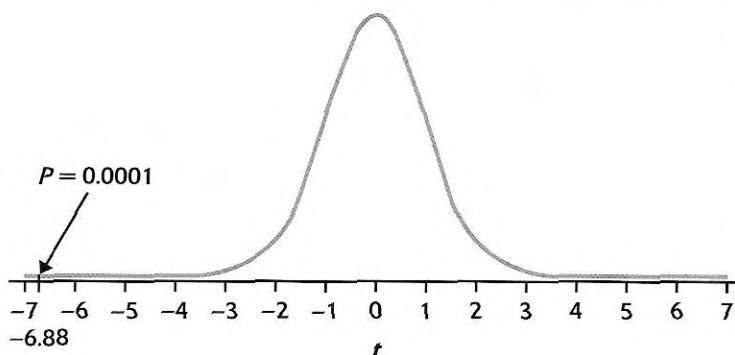


FIGURE 7.3 The P -value for Example 7.3.

The conclusion drawn in this example rests on the condition that the proper amount of vitamin C was added to the blend. If this condition is not correct, we still have evidence that the vitamin C content of the CSB produced is below the specification, but the explanation that this is caused by loss or destruction of the vitamin C in the production process is not necessarily true. Our hypothesis test has established that there is a problem with the production process, and our confidence interval tells us something about the extent of the problem. More detailed study of the process is needed to determine how to fix it.

APPLY YOUR KNOWLEDGE

- 7.5 **Apartment rents.** A random sample of 10 one-bedroom apartments from your local newspaper has these monthly rents (dollars):

500, 650, 600, 505, 450, 550, 515, 495, 650, 395

Do these data give good reason to believe that the mean rent of all advertised apartments is greater than \$500 per month? State hypotheses, find the t statistic and its P -value, and state your conclusion.

- 7.6 **Significant?** A test of a null hypothesis versus a two-sided alternative gives $t = 2.55$.

- (a) The sample size is 25. Is the test result significant at the 5% level? Explain how you obtained your answer.
- (b) The sample size is 5. Is the test result now significant at the 5% level? What general fact about statistical tests do your two answers illustrate?

- 7.7 **Have sales changed?** You will have complete sales information for last month in a week, but right now you have data from a random sample of 50 stores. The mean change in sales in the sample is +4.8% and the standard deviation of the changes is 15%. Are average sales for all stores different from last month?

- (a) State appropriate null and alternative hypotheses. Explain how you decided between the one- and two-sided alternatives.
- (b) Find the t statistic and its P -value. State your conclusion.
- (c) If the test gives strong evidence against the null hypothesis, would you conclude that sales are up in every one of your stores? Explain your answer.

Using software

For small data sets such as the one in Example 7.1 it is easy to perform the computations for confidence intervals and significance tests with an ordinary calculator. For larger data sets, however, software or a statistical calculator eases our work.

EXAMPLE 7.4

Diversify or be sued

An investor with a stock portfolio worth several hundred thousand dollars sued his broker and brokerage firm because lack of diversification in his portfolio led to poor performance. Table 7.1 gives the rates of return for the 39 months that the account was managed by the broker.³ The arbitration panel compared these returns with the average of the S&P 500 for the same period. Consider the 39 monthly returns as a random sample from the population of monthly returns the brokerage would generate if it managed the account forever. Are these returns compatible with a population mean of $\mu = 0.95\%$, the S&P 500 average? Our hypotheses are

$$H_0: \mu = 0.95$$

$$H_a: \mu \neq 0.95$$

Minitab and SPSS outputs appear in Figure 7.4. Output from other software will look similar.

TABLE 7.1 Monthly rates of return on a portfolio (percent)

-8.36	1.63	-2.27	-2.93	-2.70	-2.93	-9.14	-2.64
6.82	-2.35	-3.58	6.13	7.00	-15.25	-8.66	-1.03
-9.16	-1.25	-1.22	-10.27	-5.11	-0.80	-1.44	1.28
-0.65	4.34	12.22	-7.21	-0.09	7.34	5.04	-7.24
-2.14	-1.01	-1.41	12.03	-2.56	4.33	2.35	

Minitab

Test of $\mu = 0.950$ vs $\mu \neq 0.950$

Variable	N	Mean	StDev	SE MEAN	T	P
RETURN	39	-1.100	5.991	0.959	-2.14	0.039

SPSS

One-Sample Test

Test Value = 0.95	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
RETURN	-2.137	38	.039	-2.0497	-3.9918	-.1077

FIGURE 7.4 Minitab and SPSS output for Example 7.4.

Here is one way to report the conclusion: the mean monthly return on investment for this client's account was $\bar{x} = -1.1\%$. This differs significantly from the performance of the S&P 500 for the same period ($t = -2.14$, $df = 38$, $P < 0.039$).

The hypothesis test in Example 7.4 leads us to conclude that the mean return on the client's account differs from that of the stock index. Now let's assess the return on the client's account with a confidence interval.

EXAMPLE 7.5 Estimating mean monthly return

The mean monthly return on the client's portfolio was $\bar{x} = -1.1\%$ and the standard deviation was $s = 5.99\%$. Figure 7.5 gives the Minitab, SPSS, and Excel output for a 95% confidence interval for the population mean μ . Note that Excel gives the margin of error next to the label "Confidence Level (95.0%)" rather than the actual confidence interval. We see that the 95% confidence interval is $(-3.04, 0.84)$, or (from Excel) -1.0997 ± 1.9420 .

Because the S&P 500 return, 0.95% , falls outside this interval, we know that μ differs significantly from 0.95% at the $\alpha = 0.05$ level. Example 7.4 gave the actual P -value as $P < 0.039$.

The confidence interval suggests that the broker's management of this account had a long-term mean somewhere between a loss of 3.04% and a gain of 0.84% per month. We are interested not in the actual mean but in the difference between the broker's process and the diversified S&P 500 index.

Minitab

Variable	N	Mean	StDev	SE MEAN	95.0%	CI
RETURN	39	-1.100	5.991	0.959	(-3.042,	0.842)

SPSS

Descriptives

RETURN	Mean	Statistic	Std. Error
		-1.0997	0.9593
95% Confidence Interval for Mean		-3.0418	
		0.8423	
Lower Bound			
Upper Bound			

FIGURE 7.5 Minitab and SPSS output for Example 7.5 (continued).

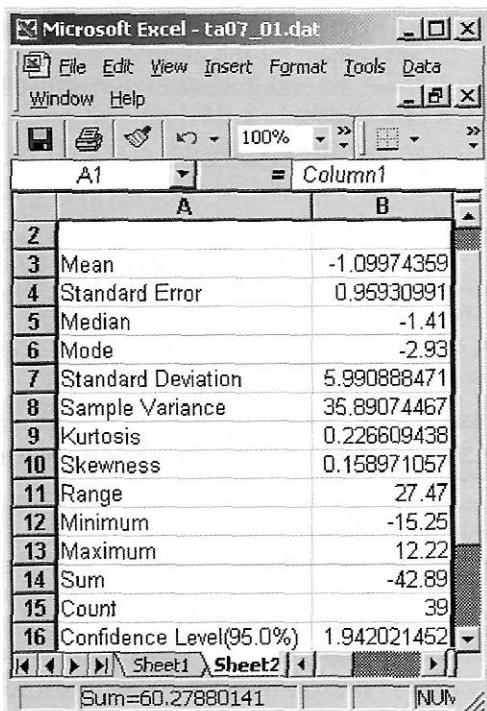


FIGURE 7.5 (continued) Excel output for Example 7.5.

EXAMPLE 7.6**Estimating difference from a standard**

Following the analysis accepted by the arbitration panel, we are considering the S&P 500 monthly average return as a constant standard. (It is easy to envision scenarios where we would want to treat this type of quantity as random.) The difference between the mean of the investor's account and the S&P 500 is $\bar{x} - \mu = -1.10 - 0.95 = -2.05\%$. In Example 7.5 we found that the 95% confidence interval for the investor's account was $(-3.04, 0.84)$. To obtain the corresponding interval for the difference, subtract 0.95 from each of the endpoints. The resulting interval is $(-3.04 - 0.95, 0.84 - 0.95)$, or $(-3.99, -0.11)$. We conclude with 95% confidence that the underperformance was between -3.99% and -0.11% . This estimate helps to set the compensation owed the investor.

APPLY YOUR KNOWLEDGE

- 7.8 Estimating mean personal income. The data set *individuals.dat*, described in the Data Appendix, contains the personal incomes of a random sample of 55,899 people between the ages of 25 and 65 who have worked outside agriculture. Use software to give a 99% confidence interval for the mean income in the entire population of such people.
- 7.9 You are using software to calculate *t* procedures for a large sample, n greater than 5000.

- (a) For a two-sided test, how large must the t statistic be to reach significance at the 1% level? (Note that either a large positive number or a large negative number will be significant.)
- (b) In which row of Table D did you find your result in part (a)? What general fact about the t distributions does this illustrate?

matched pairs

Matched pairs t procedures

The CSB production problem of Case 7.1 concerns only a single population. We know that comparative studies are usually preferred to single-sample investigations because of the protection they offer against confounding. For that reason, inference about a parameter of a single distribution is less common than comparative inference. One common comparative design, however, makes use of single-sample procedures. In a **matched pairs** study, subjects are matched in pairs and the outcomes are compared within each matched pair. The experimenter can toss a coin to assign two treatments to the two subjects in each pair. Matched pairs are also common when randomization is not possible. One situation calling for matched pairs is before-and-after observations on the same subjects, as illustrated in the next example.

EXAMPLE 7.7

The effects of language instruction

A company contracts with a language institute to provide individualized instruction in foreign languages for its executives who will be posted overseas. Is the instruction effective?

Last year, 20 executives studied French. All had some knowledge of French, so they were given the Modern Language Association's listening test of understanding of spoken French before the instruction began. After several weeks of immersion in French, the executives took the listening test again. (The actual French spoken in the two tests was different, so that simply taking the first test should not improve the score on the second test.) Table 7.2 gives the pretest and posttest scores. The

TABLE 7.2

French listening scores for executives

Executive	Pretest	Posttest	Gain	Executive	Pretest	Posttest	Gain
1	32	34	2	11	30	36	6
2	31	31	0	12	20	26	6
3	29	35	6	13	24	27	3
4	10	16	6	14	24	24	0
5	30	33	3	15	31	32	1
6	33	36	3	16	30	31	1
7	22	24	2	17	15	15	0
8	25	28	3	18	32	34	2
9	32	26	-6	19	23	26	3
10	20	26	6	20	23	26	3

maximum possible score on the test is 36.⁴ To analyze these data, subtract the pretest score from the posttest score to obtain the improvement for each executive. These 20 differences form a single sample. They appear in the “Gain” columns in Table 7.2. The first executive, for example, improved from 32 to 34, so the gain is $34 - 32 = 2$.

To assess whether the institute significantly improved the executives’ comprehension of spoken French, we test

$$H_0: \mu = 0$$

$$H_a: \mu > 0$$

Here μ is the mean improvement that would be achieved if the entire population of executives received similar instruction. The null hypothesis says that no improvement occurs, and H_a says that posttest scores are higher on the average.

The 20 differences have

$$\bar{x} = 2.5 \text{ and } s = 2.893$$

The one-sample t statistic is therefore

$$t = \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{2.5}{2.893/\sqrt{20}} \\ = 3.86$$

The P -value is found from the $t(19)$ distribution. Remember that the degrees of freedom are 1 less than the sample size. Table D shows that 3.86 lies between the upper 0.001 and 0.0005 critical values of the $t(19)$ distribution. The P -value therefore lies between these values. Software gives the value $P = 0.00053$. The improvement in listening scores is very unlikely to be due to chance alone. We have strong evidence that the posttest scores are systematically higher. In scholarly writing it is usual to omit the details of routine statistical procedures; our test would be reported in the form: “The improvement in scores was significant ($t = 3.86$, $df = 19$, $P = 0.00053$).”

df = 19		
p	0.001	0.0005
t^*	3.579	3.883

The significance test shows that the mean score has improved. A critic could argue that the test does not show that the instruction *caused* the improvement—for example, the executives might simply be less nervous when taking the test a second time. A more elaborate study with a control group who receive no instruction would be more convincing, but the company is unlikely to assign executives to “no instruction.” Also note that the subjects all knew that they would receive overseas assignments. Our conclusions apply only to the population of such executives, not to the larger population of all the company’s managers.

A statistically significant but very small improvement in language ability would not justify the expense of the individualized instruction. A confidence interval allows us to estimate the *amount* of the improvement.

EXAMPLE 7.8

A 90% confidence interval for the mean improvement in the entire population requires the critical value $t^* = 1.729$ from Table D. The confidence interval is

$$\begin{aligned}\bar{x} \pm t^* \frac{s}{\sqrt{n}} &= 2.5 \pm 1.729 \frac{2.893}{\sqrt{20}} \\ &= 2.5 \pm 1.12 \\ &= (1.38, 3.62)\end{aligned}$$

The estimated average improvement is 2.5 points, with margin of error 1.12 for 90% confidence. Though statistically significant, the effect of the instruction was rather small.

A look at the data discloses one reason for the small average improvement. Several of the executives had pretest scores close to the maximum of 36. They could not improve their scores very much even if their mastery of French increased substantially. This is a weakness in the listening test that is the measuring instrument in this study.

Here are some key points to remember concerning matched pairs:

- A matched pairs analysis is needed when there are two measurements or observations on each individual and we want to examine the change from the first to the second. Frequently, the observations are “before” and “after” measures in some sense.
- For each individual, subtract the “before” measure from the “after” measure.
- Analyze the difference using the one-sample confidence interval and significance-testing procedures that we learned in this section.

7.10 Does the product lose value when cooked? The researchers studying vitamin C in CSB in Case 7.1 were also interested in a similar commodity called wheat soy blend (WSB). Both CSB and WSB are mixed with other ingredients and cooked. Cooking can cause loss of vitamin C. In Haiti, gruel (“bouillie” in Creole) is made from WSB, sugar, milk, banana, and seasonings. The researchers collected specimens of gruel prepared in Haitian households and measured the vitamin C content before and after cooking. Here are the results (milligrams per 100 grams of blend, dry basis):



Household	1	2	3	4	5
Before	73	79	86	88	78
After	20	27	29	36	17

It is not possible for cooking to increase the amount of vitamin C. State appropriate hypotheses and carry out a matched pairs t test for these data.



- 7.11 How much loss? Exercise 7.10 demonstrates that cooking reduces the vitamin C content of food made with wheat soy blend. This fact is neither new nor surprising. The real question is how much vitamin C is lost. Use the data in Exercise 7.10 to give a 95% confidence interval for the amount of vitamin C lost in preparing and cooking gruel in Haiti.

Robustness of the t procedures

The matched pairs t procedures use one-sample t confidence intervals and significance tests for differences between pairs. They are therefore valid under the condition that the population of differences has a Normal distribution. The differences in Table 7.2 show departures from Normality. One executive actually lost 6 points between the pretest and the posttest. This one subject lowered the sample mean from 2.95 for the other 19 subjects to 2.5 for all 20. A Normal quantile plot (Figure 7.6) displays this outlier as well as granularity due to the fact that only whole-number scores are possible. The overall pattern of the plot is otherwise roughly straight. Does this non-Normality forbid use of the t test? The behavior of the t procedures when the population does not have a Normal distribution is one of their most important properties.

The results of one-sample t procedures are exactly correct only when the population is Normal. Real populations are never exactly Normal. All inference procedures are based on some conditions, such as Normality: procedures that are not strongly affected by violations of a condition are called *robust*.

ROBUST PROCEDURES

A statistical inference procedure is called **robust** if the probability calculations required are insensitive to violations of the conditions that usually justify the procedure.

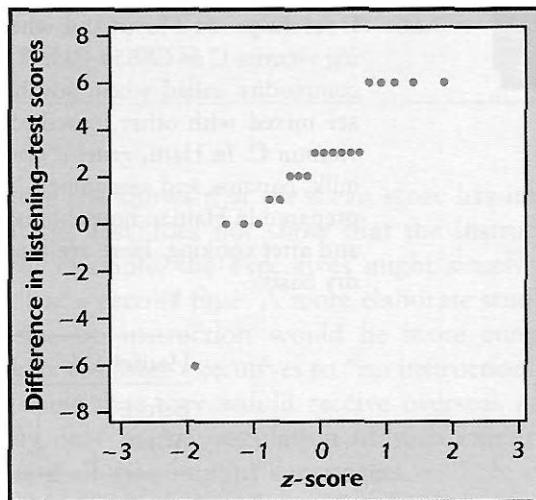


FIGURE 7.6 Normal quantile plot for the change in French listening score.

The condition that the population be Normal rules out outliers, so the presence of outliers shows that this condition is not fulfilled. The *t* procedures are not robust against outliers, because \bar{x} and s are not resistant to outliers.

Fortunately, the *t* procedures are quite robust against non-Normality of the population except when outliers or strong skewness are present. Larger samples improve the accuracy of *P*-values and critical values from the *t* distributions when the population is not Normal. The *t* procedures rely only on the Normality of the sample mean \bar{x} . This condition is satisfied when the population is Normal, but the central limit theorem tells us that a mean \bar{x} from a large sample follows a Normal distribution closely even when individual observations are not Normally distributed.

Before using the *t* procedures for small samples, make a Normal quantile plot, stemplot, or boxplot to check for skewness and outliers. For most purposes, the one-sample *t* procedures can be safely used when $n \geq 15$ unless an outlier or clearly marked skewness is present. Except in the case of small samples, the condition that the data are an SRS from the population of interest is more crucial than the condition that the population distribution is Normal. Here are practical guidelines for inference on a single mean:⁵

- *Sample size less than 15:* Use *t* procedures if the data are close to Normal. If the data are clearly non-Normal or if outliers are present, do not use *t*.
- *Sample size at least 15:* The *t* procedures can be used except in the presence of outliers or strong skewness.
- *Large samples:* The *t* procedures can be used even for clearly skewed distributions when the sample is large, roughly $n \geq 40$.

The French instruction data are a borderline case: there are 20 observations, with one clear outlier. The *t* procedures give approximately correct results, but we would be happier if a good reason, such as illness on the posttest date, justified removing the outlier.

APPLY YOUR KNOWLEDGE

- 7.12 The Platinum Gasaver.** National Fuelsaver Corporation manufactures the Platinum Gasaver, a device they claim “may increase gas mileage by 22%.” Here are the percent changes in gas mileage for 15 identical vehicles, as presented in one of the company’s advertisements:

48.3	46.9	46.8	44.6	40.2	38.5	34.6	33.7
28.7	28.7	24.8	10.8	10.4	6.9	-12.4	

Would you recommend use of a *t* confidence interval to estimate the mean fuel savings in the population of all such vehicles? Explain your answer.

- 7.13 *t* is robust.** A manufacturer of small appliances employs a market research firm to estimate retail sales of its products. Here are last month’s sales of electric can openers from an SRS of 50 stores in the Midwest sales region:

19	19	16	19	25	26	24	63	22	16
13	26	34	10	48	16	20	14	13	24
34	14	25	16	26	25	25	26	11	79
17	25	18	15	13	35	17	15	21	12
19	20	32	19	24	19	17	41	24	27

- Make a stemplot of the data. The distribution is skewed to the right and has several high outliers. The *bootstrap* (page 374) is a modern computer-intensive tool for getting accurate confidence intervals without the Normality condition. Three bootstrap simulations, each with 10,000 repetitions, give these 95% confidence intervals for mean sales in the entire region: (20.42, 27.26), (20.40, 27.18), and (20.48, 27.28).
- Find the 95% t confidence interval for the mean. It is essentially the same as the bootstrap intervals. The lesson is that for sample sizes as large as $n = 50$, t procedures are very robust.

The power of the t test*

The power of a statistical test measures its ability to detect deviations from the null hypothesis. In practice, we carry out the test in the hope of showing that the null hypothesis is false, so high power is important. The power of the one-sample t test against a specific alternative value of the population mean μ is the probability that the test will reject the null hypothesis when the alternative is really true. To calculate the power, we assume a fixed level of significance, often $\alpha = 0.05$.

Calculation of the exact power of the t test takes into account the estimation of σ by s and requires special software. Fortunately, an approximate calculation that acts as if σ were known is almost always adequate for planning a study. This calculation is very much like that for the z test, presented in Section 6.4: write the event that the test rejects H_0 in terms of \bar{x} , and then find the probability of this event when the population mean has the alternative value.

EXAMPLE 7.9

The company in Example 7.7 is planning to evaluate the next language instruction session. We decide that an improvement of 2 points or more is large enough to be useful in practice. Can we rely on a sample of 20 executives to detect an improvement of this size?

We wish to compute the power of the t test for

$$H_0: \mu = 0$$

$$H_a: \mu > 0$$

against the alternative $\mu = 2$ when $n = 20$. We must have a rough guess of the size of σ in order to compute the power. In planning a large study, a pilot study is often run for this and other purposes. In this case, listening-score improvements in past language instruction programs for similar executives have had sample standard deviations of about 3. We therefore take both $\sigma = 3$ and $s = 3$ in our approximate calculation.

The t test with 20 observations rejects H_0 at the 5% significance level if the t statistic

$$t = \frac{\bar{x} - 0}{s/\sqrt{20}}$$

*This section can be omitted without loss of continuity.

exceeds the upper 5% point of $t(19)$, which is 1.729. Taking $s = 3$, the event that the test rejects H_0 is therefore

$$\begin{aligned} t &= \frac{\bar{x}}{3/\sqrt{20}} \geq 1.729 \\ \bar{x} &\geq 1.729 \cdot \frac{3}{\sqrt{20}} \\ \bar{x} &\geq 1.160 \end{aligned}$$

The power is the probability that $\bar{x} \geq 1.160$ when $\mu = 2$. Taking $\sigma = 3$, we find this probability by standardizing \bar{x} :

$$\begin{aligned} P(\bar{x} \geq 1.160 \text{ when } \mu = 2) &= P\left(\frac{\bar{x} - 2}{3/\sqrt{20}} \geq \frac{1.160 - 2}{3/\sqrt{20}}\right) \\ &= P(Z \geq -1.252) \\ &= 1 - 0.1056 = 0.8944 \end{aligned}$$

A true difference of 2 points in the population mean scores will produce significance at the 5% level in 89% of all possible samples. We can be reasonably confident of detecting a difference this large.

APPLY YOUR KNOWLEDGE

- 7.14 If you repeat the calculation in Example 7.9 for other values of μ that are larger than 2, would you expect the power to be higher or lower than 0.8944? Why?
- 7.15 Verify your answer to the previous exercise by doing the calculation for the alternative $\mu = 2.3$.

Inference for non-Normal populations*

We have not discussed how to do inference about the mean of a clearly non-Normal distribution based on a small sample. If you face this problem, you should consult an expert. Three general strategies are available:

- In some cases a distribution other than a Normal distribution describes the data well. There are many non-Normal models for data, and inference procedures for these models are available.
- Because skewness is the chief barrier to the use of t procedures on data without outliers, you can attempt to transform skewed data so that the distribution is symmetric and as close to Normal as possible. Confidence levels and P -values from the t procedures applied to the transformed data will be quite accurate for even moderate sample sizes.
- The third strategy is to use a **distribution-free** inference procedure. Such procedures do not assume that the population distribution has any specific form, such as Normal. Distribution-free procedures are often called **nonparametric procedures**. The *bootstrap* (page 374 and Chapter 14)

*distribution-free
procedures*

*nonparametric
procedures*

*This section can be omitted without loss of continuity. It requires knowledge of the Binomial distributions from the optional Chapter 5.

is a modern computer-intensive nonparametric procedure that is especially useful for confidence intervals. Chapter 18 discusses traditional nonparametric procedures, especially significance tests.

Each of these strategies quickly carries us beyond the basic practice of statistics. We emphasize procedures based on Normal distributions because they are the most common in practice, because their robustness makes them widely useful, and (most important) because we are first of all concerned with understanding the principles of inference. We will be content with illustrating by example a simple and useful distribution-free procedure.

The Sign Test

Distribution-free significance tests do not require that the data follow any specific type of distribution such as Normal. The gain in generality isn't free: if the data really are close to Normal, distribution-free tests have less power than *t* tests. They also don't answer quite the same question. The *t* tests concern the population *mean*. Distribution-free tests ask about the population *median*, as is natural for distributions that may be skewed. The simplest distribution-free test, and one of the most useful, is the *sign test*. The following example illustrates this test.

EXAMPLE 7.10

Example 7.7 (page 443) gives data on the improvement in French listening scores after individualized instruction. In that example we used the matched pairs *t* test on these data, despite granularity and an outlier that make the *P*-value only roughly correct. The sign test is based on the following simple observation: of the 17 executives whose scores changed, 16 improved and only 1 did more poorly. This is evidence that the instruction improved French listening skills.

To perform a significance test based on the count of executives whose scores improved, let p be the probability that a randomly chosen executive would improve if she attended the institute. The null hypothesis of "no effect" says that the posttest is just a repeat of the pretest with no change in ability, so an executive is equally likely to do better on either test. We therefore want to test

$$H_0: p = 1/2$$

$$H_a: p > 1/2$$

The 17 executives whose scores changed are 17 independent trials, so the number who improve has the Binomial distribution $B(17, 1/2)$ if H_0 is true. The *P*-value for the observed count 16 is therefore $P(X \geq 16)$, where X has the $B(17, 1/2)$ distribution. You can compute this probability with software or from the Binomial probability formula (page 323):

$$\begin{aligned} P(X \geq 16) &= P(X = 16) + P(X = 17) \\ &= \binom{17}{16} \left(\frac{1}{2}\right)^{16} \left(\frac{1}{2}\right)^1 + \binom{17}{17} \left(\frac{1}{2}\right)^{17} \left(\frac{1}{2}\right)^0 \\ &= (17) \left(\frac{1}{2}\right)^{17} + \left(\frac{1}{2}\right)^{17} \\ &= 0.00014 \end{aligned}$$

As in Example 7.7, there is very strong evidence that participation in the institute has improved performance on the listening test.

There are several varieties of the sign test, all based on counts and the Binomial distribution. The sign test for matched pairs (Example 7.10) is the most useful. The null hypothesis of “no effect” is then always $H_0: p = 1/2$. The alternative can be one-sided in either direction or two-sided, depending on the type of change we are looking for. The test gets its name from the fact that we look only at the signs of the differences, not their actual values.

THE SIGN TEST FOR MATCHED PAIRS

Ignore pairs with difference 0; the number of trials n is the count of the remaining pairs. The test statistic is the count X of pairs with a positive difference. P -values for X are based on the Binomial $B(n, 1/2)$ distribution.

The matched pairs t test in Example 7.7 tested the hypothesis that the mean of the distribution of differences (score after the instruction minus score before) is 0. The sign test in Example 7.10 is in fact testing the hypothesis that the *median* of the differences is 0. If p is the probability that a difference is positive, then $p = 1/2$, when the median is 0. This is true because the median of the distribution is the point with probability 1/2 lying to its right. As Figure 7.7 illustrates, $p > 1/2$, when the median is greater than 0, again because the probability to the right of the median is always 1/2. The sign test of $H_0: p = 1/2$ against $H_a: p > 1/2$ is a test of

$$H_0: \text{population median} = 0$$

$$H_a: \text{population median} > 0$$

The sign test in Example 7.10 makes no use of the actual scores—it just counts how many executives improved. The executives whose scores did not change were ignored altogether. Because the sign test uses so little of the available information, it is much less powerful than the t test when the population is close to Normal. Chapter 18 describes other distribution-free tests that are more powerful than the sign test.

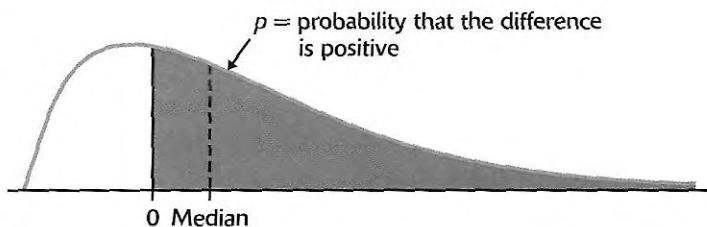


FIGURE 7.7 Why the sign test tests the median difference: when the median is greater than 0, the probability p of a positive difference is greater than 1/2, and vice versa.

APPLY YOUR KNOWLEDGE

- 7.16 Does the product lose value when cooked? Exercise 7.10 (page 445) gives data on the amount of vitamin C in gruel made from wheat soy blend in 5 Haitian households before and after cooking. Is there evidence that the median amount of vitamin C is less after cooking? State hypotheses, carry out a sign test, and report your conclusion.

CASE 7.1
**SECTION 7.1 SUMMARY**

- Significance tests and confidence intervals for the mean μ of a Normal population are based on the sample mean \bar{x} of an SRS. Because of the central limit theorem, the resulting procedures are approximately correct for other population distributions when the sample is large.
- The standardized sample mean, or one-sample z statistic,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

has the $N(0, 1)$ distribution. If the standard deviation σ/\sqrt{n} of \bar{x} is replaced by the standard error s/\sqrt{n} , the one-sample t statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has the **t distribution** with $n - 1$ degrees of freedom.

- There is a t distribution for every positive degrees of freedom k . All are symmetric distributions similar in shape to Normal distributions. The $t(k)$ distribution approaches the $N(0, 1)$ distribution as k increases.
- A level C confidence interval for the mean μ of a Normal population is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the value for the $t(n - 1)$ density curve with area C between $-t^*$ and t^* . The quantity

$$t^* \frac{s}{\sqrt{n}}$$

is the **margin of error**.

- Significance tests for $H_0: \mu = \mu_0$ are based on the t statistic. P -values or fixed significance levels are computed from the $t(n - 1)$ distribution.
- These one-sample procedures are used to analyze **matched pairs** data by first taking the differences within the matched pairs to produce a single sample.
- The t procedures are relatively **robust** against lack of Normality, especially for larger sample sizes. The t procedures are useful for non-Normal data when $n \geq 15$ unless the data show outliers or strong skewness.

- The power of the t test is calculated like that of the z test, using an approximate value for both σ and s .
- The sign test is a distribution-free test because it uses probability calculations that are correct for a wide range of population distributions.
- The sign test for “no treatment effect” in matched pairs counts the number of positive differences. The P -value is computed from the $B(n, 1/2)$ distribution, where n is the number of non-0 differences. The sign test is less powerful than the t test in cases where use of the t test is justified.

SECTION 7.1 EXERCISES

- 7.17 What critical value t^* from Table D should be used for a confidence interval for the mean of the population in each of the following situations?
- A 90% confidence interval based on $n = 12$ observations.
 - A 95% confidence interval from an SRS of 30 observations.
 - An 80% confidence interval from a sample of size 18.
- 7.18 Use software to find the critical values t^* that you would use for each of the following confidence intervals for the mean.
- A 99% confidence interval based on $n = 55$ observations.
 - A 90% confidence interval from an SRS of 35 observations.
 - A 95% confidence interval from a sample of size 90.
- 7.19 The one-sample t statistic for testing

$$H_0: \mu = 0$$

$$H_a: \mu > 0$$

from a sample of $n = 15$ observations has the value $t = 1.97$.

- What are the degrees of freedom for this statistic?
 - Give the two critical values t^* from Table D that bracket t .
 - What are the right-tail probabilities p for these two entries?
 - Between what two values does the P -value of the test fall?
 - Is the value $t = 1.97$ significant at the 5% level? Is it significant at the 1% level?
 - If you have software available, find the exact P -value.
- 7.20 The one-sample t statistic from a sample of $n = 30$ observations for the two-sided test of

$$H_0: \mu = 64$$

$$H_a: \mu \neq 64$$

has the value $t = 1.12$.

- What are the degrees of freedom for t ?
- Locate the two critical values t^* from Table D that bracket t . What are the right-tail probabilities p for these two values?
- How would you report the P -value for this test?

- (d) Is the value $t = 1.12$ statistically significant at the 10% level? At the 5% level?
 (e) If you have software available, find the exact P -value.

7.21 The one-sample t statistic for a test of

$$\begin{aligned} H_0: \mu &= 20 \\ H_a: \mu &< 20 \end{aligned}$$

based on $n = 12$ observations has the value $t = -2.45$.

- (a) What are the degrees of freedom for this statistic?
 (b) How would you report the P -value based on Table D?
 (c) If you have software available, find the exact P -value.

7.22 LSAT scores. The scores of four roommates on the Law School Admission Test are

$$628, 593, 455, 503$$

Find the mean, the standard deviation, and the standard error of the mean. Is it appropriate to calculate a confidence interval based on these data? Explain why or why not.

7.23 Earnings of bank workers. Case 1.2 of Chapter 1 (page 30) concerns the annual earnings of hourly-paid workers at National Bank. Table 1.8 (page 31) gives the earnings of a random sample of 20 white female workers. Their annual earnings (dollars) are

$$\begin{array}{cccccccccc} 25,249 & 19,029 & 17,233 & 26,606 & 28,346 & 31,176 & 18,863 \\ 15,904 & 22,477 & 19,102 & 18,002 & 21,596 & 26,885 & 24,780 \\ 14,698 & 19,308 & 17,576 & 24,497 & 20,612 & 17,757 & \end{array}$$

- (a) Display the data in a stemplot and, if your software permits, a normal quantile plot. The distribution shows no strong departures from Normality.
 (b) Give a 95% confidence interval for the mean annual earnings of all white female hourly workers at this bank.

7.24 The cost of Internet access. How much do users pay for Internet service? Here are the monthly fees (in dollars) paid by a random sample of 50 users of commercial Internet service providers in August 2000:⁶

$$\begin{array}{cccccccccc} 20 & 40 & 22 & 22 & 21 & 21 & 20 & 10 & 20 & 20 \\ 20 & 13 & 18 & 50 & 20 & 18 & 15 & 8 & 22 & 25 \\ 22 & 10 & 20 & 22 & 22 & 21 & 15 & 23 & 30 & 12 \\ 9 & 20 & 40 & 22 & 29 & 19 & 15 & 20 & 20 & 20 \\ 20 & 15 & 19 & 21 & 14 & 22 & 21 & 35 & 20 & 22 \end{array}$$

- (a) Make a stemplot of the data. Also make a Normal quantile plot if your software permits. The data are not Normal: there are stacks of observations taking the same values, and the distribution is more spread out in both directions and somewhat skewed to the right. The t procedures are nonetheless approximately correct because $n = 50$ and there are no extreme outliers.
 (b) Give a 95% confidence interval for the mean monthly cost of Internet access in August 2000.

- 7.25 **Earnings of bank workers.** Does the sample in Exercise 7.23 give good evidence that the mean annual earnings of all white female hourly workers at National Bank are greater than \$20,000?
- 7.26 **The cost of Internet access.** The data in Exercise 7.24 show that many people paid \$20 per month for Internet access, presumably because major providers such as AOL charged this amount. Do the data give good reason to think that the mean cost for all Internet users differs from \$20 per month?
- 7.27 **Supermarket shoppers.** A marketing consultant observed 50 consecutive shoppers at a supermarket. One variable of interest was how much each shopper spent in the store. Here are the data (in dollars), arranged in increasing order:

3.11	8.88	9.26	10.81	12.69	13.78	15.23	15.62	17.00
17.39	18.36	18.43	19.27	19.50	19.54	20.16	20.59	22.22
23.04	24.47	24.58	25.13	26.24	26.26	27.65	28.06	28.08
28.38	32.03	34.98	36.37	38.64	39.16	41.02	42.97	44.08
44.67	45.40	46.69	48.65	50.39	52.75	54.80	59.07	61.22
70.32	82.70	85.76	86.37	93.34				

- (a) Display the data using a stemplot. Make a Normal quantile plot if your software allows. The data are clearly non-Normal. In what way? Because $n = 50$, the t procedures remain quite accurate.
- (b) Calculate the mean, the standard deviation, and the standard error of the mean.
- (c) Find a 95% t confidence interval for the mean spending for all shoppers at this store.
- 7.28 **The influence of big shoppers.** Eliminate the four largest observations and redo parts (a), (b), and (c) of the previous exercise. Do these observations have a large influence on the results?
- 7.29 **Economic impact of the Internet.** Exercise 7.24 gives the fees paid for Internet access by a national random sample of clients of Internet service providers in 2000. The Census Bureau estimates that 44 million households had Internet access in 2000. Use your confidence interval from Exercise 7.24 to give a 95% confidence interval for the total amount these households paid in Internet access fees. This is one aspect of the national economic impact of the Internet.
- 7.30 **Loss of vitamin C, reconsidered.** Exercise 7.10 gives these data on the amount of vitamin C in gruel made from wheat soy blend in 5 Haitian households before and after cooking:



Household	1	2	3	4	5
Before	73	79	86	88	78
After	20	27	29	36	17

The units are milligrams per 100 grams of blend, dry basis. In Exercise 7.11, you were asked to give a confidence interval for the amount of vitamin C

lost. The result is hard to interpret without some background information. A loss of 50 mg/100 g would be of little concern if we started with 5000 mg/100 g but would be serious if we started with 75 mg/100 g. In fact, the specifications call for the blend to contain 98 mg/100 g of vitamin C. The “before” values differ from the specification due to variation in manufacturing and in handling the product before it was used to prepare gruel in Haiti. Express the “after” data as percent of specification and give a 95% confidence interval for the mean percent.

- 7.31 **Corn prices.** The U.S. Department of Agriculture (USDA) uses sample surveys to obtain important economic estimates. One USDA pilot study estimated the price received by farmers for corn sold in October from a sample of 22 farms. The mean price was reported as \$2.08 per bushel with a standard error of \$0.176 per bushel. Give a 95% confidence interval for the mean price received by farmers for corn sold in October.⁷
- 7.32 **Health care costs.** The cost of health care is the subject of many studies that use statistical methods. One such study estimated that the average length of service for home health care among people over the age of 65 who use this type of service is 96.0 days with a standard error of 5.1 days. Assuming a large sample, calculate a 90% confidence interval for the mean length of service for all users of home health care.⁸
- 7.33 **A customer satisfaction survey.** Many organizations do surveys to determine the satisfaction of their customers. Attitudes toward various aspects of campus life were the subject of one such study conducted at Purdue University. Each question was rated on a 1 to 5 scale, with 5 being the highest rating. The mean response of 1406 first-year students to “Feeling welcomed at Purdue” was 3.9 and the standard deviation of the responses was 0.98. Assuming that the respondents are an SRS, give a 99% confidence interval for the mean of all first-year students.
- 7.34 **Piano lessons.** Do piano lessons improve the spatial-temporal reasoning of preschool children? Neurobiological arguments suggest that this may be true. A study designed to test this hypothesis measured the spatial-temporal reasoning of 34 preschool children before and after six months of piano lessons.⁹ (The study also included children who took computer lessons and a control group, but we are not concerned with those here.) The changes in the reasoning scores are

2	5	7	-2	2	7	4	1	0	7	3	4	3	4	9	4	5
2	9	6	0	3	6	-1	3	4	6	7	-2	7	-3	3	4	4

- (a) Display the data and summarize the distribution.
- (b) Find the mean, the standard deviation, and the standard error of the mean.
- (c) Give a 95% confidence interval for the mean improvement in reasoning scores.
- 7.35 **Are the results statistically significant?** Using the data of the previous exercise, test the null hypothesis that there is no improvement versus the alternative suggested by the neurobiological arguments. State the hypotheses, and give the test statistic with degrees of freedom and the *P*-value. What do you conclude? From your answer to part (c) of the previous exercise, what can be concluded from this significance test?

7.36 Credit card fees. A bank wonders whether omitting the annual credit card fee for customers who charge at least \$2400 in a year would increase the amount charged on its credit card. The bank makes this offer to an SRS of 250 of its existing credit card customers. It then compares how much these customers charge this year with the amount that they charged last year. The mean increase is \$342, and the standard deviation is \$108.

- Is there significant evidence at the 1% level that the mean amount charged increases under the no-fee offer? State H_0 and H_a and carry out a t test.
- Give a 95% confidence interval for the mean amount of the increase.
- The distribution of the amount charged is skewed to the right, but outliers are prevented by the credit limit that the bank enforces on each card. Use of the t procedures is justified in this case even though the population distribution is not Normal. Explain why.
- A critic points out that the customers would probably have charged more this year than last even without the new offer because the economy is more prosperous and interest rates are lower. Briefly describe the design of an experiment to study the effect of the no-fee offer that would avoid this criticism.

7.37 Cockroaches. Your company is developing a better means to eliminate cockroaches from buildings. In the process, your R&D team studies the absorption of sugar by these insects. They feed cockroaches a diet containing measured amounts of D-glucose. The cockroaches are killed at different times after eating, and the concentration of this sugar in various body parts is determined by a chemical analysis. Five cockroaches dissected after 10 hours had the following amounts (in micrograms) of D-glucose in their hindguts:¹⁰

55.95 68.24 52.73 21.50 23.78

Find a 95% confidence interval for the mean amount of D-glucose in cockroach hindguts under these conditions.

7.38 Accuracy of radon detectors. How accurate are radon detectors of a type sold to homeowners? To answer this question, university researchers placed 12 detectors in a chamber that exposed them to 105 picocuries per liter (pCi/l) of radon. The detector readings were as follows:¹¹

91.9	97.8	111.4	122.3	105.4	95.0
103.8	99.6	96.6	119.3	104.8	101.7

- Make a stemplot of the data. The distribution is somewhat skewed to the right, but not strongly enough to forbid use of the t procedures.
- Is there convincing evidence that the mean reading of all detectors of this type differs from the true value 105? Carry out a test in detail and write a brief conclusion.

7.39 Effect of storage and shipment on a product. The researchers in the project described in Case 7.1 are also interested in wheat soy blend (WSB). One question is whether some of the vitamin C content is destroyed as a result of storage and shipment of the product. The researchers marked a sample of bags at the factory and tested each to determine the vitamin C content. Five months later in Haiti they found the



marked bags and again tested their contents. The data consist of two vitamin C measures for each bag, one at the time of production in the factory and the other five months later in Haiti. The units are milligrams of vitamin C per 100 grams of WSB. Here are the data:

Factory	Haiti	Factory	Haiti	Factory	Haiti
44	40	45	38	39	43
50	37	32	40	52	38
48	39	47	35	45	38
44	35	40	38	37	38
42	35	38	34	38	41
47	41	41	35	44	40
49	37	43	37	43	35
50	37	40	34	39	38
39	34	37	40	44	36

- (a) Set up hypotheses to examine the question of interest to these researchers.
 (b) Perform the significance test and summarize your results.
 (c) Find 95% confidence intervals for the mean at the factory, for the mean five months later in Haiti, and for the change.
- 7.40 **Design of controls.** The design of controls and instruments has a large effect on how easily people can use them. A student project investigated this effect by asking 25 right-handed students to turn a knob (with their right hands) that moved an indicator by screw action. There were two identical instruments, one with a right-hand thread (the knob turns clockwise) and the other with a left-hand thread (the knob turns counterclockwise). The table below gives the times required (in seconds) to move the indicator a fixed distance:¹²

Subject	Right thread	Left thread	Subject	Right thread	Left thread
1	113	137	14	107	87
2	105	105	15	118	166
3	130	133	16	103	146
4	101	108	17	111	123
5	138	115	18	104	135
6	118	170	19	111	112
7	87	103	20	89	93
8	116	145	21	78	76
9	75	78	22	100	116
10	96	107	23	89	78
11	122	84	24	85	101
12	103	148	25	88	123
13	116	147			

- (a) Each of the 25 students used both instruments. Discuss briefly how the experiment should be arranged and how randomization should be used.
 (b) The project hoped to show that right-handed people find right-hand threads easier to use. State the appropriate H_0 and H_a about the mean time required to complete the task.

- (c) Carry out a test of your hypotheses. Give the P -value and report your conclusions.
- 7.41 **Is the difference important?** Give a 90% confidence interval for the mean time advantage of right-hand over left-hand threads in the setting of the previous exercise. Do you think that the time saved would be of practical importance if the task were performed many times—for example, by an assembly-line worker? To help answer this question, find the mean time for right-hand threads as a percent of the mean time for left-hand threads.
- 7.42 **Executives learn Spanish.** The table below gives the pretest and posttest scores on the MLA listening test in Spanish for 20 executives who received intensive training in Spanish.¹³ The setting is identical to the one described in Example 7.7.
- | Subject | Pretest | Posttest | Subject | Pretest | Posttest |
|---------|---------|----------|---------|---------|----------|
| 1 | 30 | 29 | 11 | 30 | 32 |
| 2 | 28 | 30 | 12 | 29 | 28 |
| 3 | 31 | 32 | 13 | 31 | 34 |
| 4 | 26 | 30 | 14 | 29 | 32 |
| 5 | 20 | 16 | 15 | 34 | 32 |
| 6 | 30 | 25 | 16 | 20 | 27 |
| 7 | 34 | 31 | 17 | 26 | 28 |
| 8 | 15 | 18 | 18 | 25 | 29 |
| 9 | 28 | 33 | 19 | 31 | 32 |
| 10 | 20 | 25 | 20 | 29 | 32 |
- (a) We hope to show that the training improves listening skills. State an appropriate H_0 and H_a . Describe in words the parameters that appear in your hypotheses.
- (b) Make a graphical check for outliers or strong skewness in the data that you will use in your statistical test, and report your conclusions on the validity of the test.
- (c) Carry out a test. Can you reject H_0 at the 5% significance level? At the 1% significance level?
- (d) Give a 90% confidence interval for the mean increase in listening score due to the intensive training.
- 7.43 **A field trial.** An agricultural field trial compares the yield of two varieties of tomatoes for commercial use. The researchers divide in half each of 10 small plots of land in different locations and plant each tomato variety on one half of each plot. After harvest, they compare the yields in pounds per plant at each location. The 10 differences (Variety A – Variety B) give the following statistics: $\bar{x} = 0.34$ and $s = 0.83$. Is there convincing evidence that Variety A has the higher mean yield? State H_0 and H_a , and give a P -value to answer this question.
- 7.44 **Which design?** The following situations all require inference about a mean or means. Identify each as (1) a single sample, (2) matched pairs, or (3) two independent samples. The procedures of this section apply to cases (1) and (2). We will learn procedures for (3) in the next section.
- (a) A marketing researcher wants to learn whether a new type of packaging will be more attractive to consumers. The new packaging design is

shown to a collection of potential customers who give it a rating of 1 to 10 for attractiveness. The old design is shown to a different group of customers who rate its attractiveness. The attractiveness scores for the two packaging designs are compared.

- (b) Another marketing researcher approaches the same problem differently. She shows both packaging designs to a group of potential customers and asks each to give an attractiveness score for both designs. She then compares the scores.

7.45 Which design? The following situations all require inference about a mean or means. Identify each as (1) a single sample, (2) matched pairs, or (3) two independent samples. The procedures of this section apply to cases (1) and (2). We will learn procedures for (3) in the next section.

- (a) To evaluate a new analytical method, a chemist obtains a reference specimen of known concentration from the National Institute of Standards and Technology. She then makes 20 measurements of the concentration of this specimen with the new method and checks for bias by comparing the mean result with the known concentration.
- (b) Another chemist is evaluating the same new method. He has no reference specimen, but a familiar analytic method is available. He wants to know if the new and old methods agree. He takes a specimen of unknown concentration and measures the concentration 10 times with the new method and 10 times with the old method.

7.46 Unemployment rates. Table 1.1 (page 10) gives the unemployment rates in each of the 50 states. It does not make sense to use the *t* procedures (or any other statistical procedures) on these data to give a 95% confidence interval for the mean unemployment rate in the population of the American states. Explain why not.

*The following exercises concern the optional material in the sections on the power of the *t* test and on non-Normal populations.*

7.47 Credit card fees. The bank in Exercise 7.36 tested a new idea on a sample of 250 customers. Suppose that the bank wanted to be quite certain of detecting a mean increase of $\mu = \$100$ in the credit card amount charged, at the $\alpha = 0.01$ significance level. Perhaps a sample of only $n = 50$ customers would accomplish this. Find the approximate power of the test with $n = 50$ against the alternative $\mu = \$100$ as follows:

- (a) What is the *t* critical value for the one-sided test with $\alpha = 0.01$ and $n = 50$?
- (b) Write the criterion for rejecting $H_0: \mu = 0$ in terms of the *t* statistic. Then take $s = 108$ (an estimate based on the data in Exercise 7.36) and state the rejection criterion in terms of \bar{x} .
- (c) Assume that $\mu = 100$ (the given alternative) and that $\sigma = 108$ (an estimate from the data in Exercise 7.36). The approximate power is the probability of the event you found in (b), calculated under these assumptions. Find the power. Would you recommend that the bank do a test on 50 customers, or should more customers be included?

7.48 A field trial. The tomato experts who carried out the field trial described in Exercise 7.43 suspect that the relative lack of significance there is due to low power. They would like to be able to detect a mean difference in yields of 0.5 pound per plant at the 0.05 significance level. Based on the previous

- study, use 0.83 as an estimate of both the population σ and the value of s in future samples.
- (a) What is the power of the test from Exercise 7.43 with $n = 10$ against the alternative $\mu = 0.5$?
- (b) If the sample size is increased to $n = 25$ plots of land, what will be the power against the same alternative?
- 7.49 **Design of controls.** Apply the sign test to the data in Exercise 7.40 to assess whether the subjects can complete a task with a right-hand thread significantly faster than with a left-hand thread.
- (a) State the hypotheses two ways, in terms of a population median and in terms of the probability of completing the task faster with a right-hand thread.
- (b) Carry out the sign test. Find the approximate P -value using the Normal approximation to the Binomial distributions, and report your conclusion.
- 7.50 **Learning Spanish.** Use the sign test to assess whether the intensive language training of Exercise 7.42 improves Spanish listening skills. State the hypotheses, give the P -value using the Binomial table (Table C), and report your conclusion.

7.2 Comparing Two Means

How do small businesses that fail differ from those that succeed? Business school researchers compare two samples of firms started in 2000, one sample of failed businesses and one of firms that are still going after two years. This study *compares two random samples*, one from each of two different populations. Which of two incentive packages will lead to higher use of a bank's credit cards? The bank designs a *randomized comparative experiment to compare two treatments*. Credit card customers are assigned at random to receive one or the other incentive offer. After six months, the bank compares the amounts charged. *Two-sample problems* such as these are among the most common situations encountered in statistical practice.

TWO-SAMPLE PROBLEMS

- The goal of inference is to compare the responses in two groups.
- Each group is considered to be a sample from a distinct population.
- The responses in each group are independent of those in the other group.

You must carefully distinguish two-sample problems from the matched pairs designs studied earlier. In two-sample problems, there is no matching of the units in the two samples, and the two samples may be of different sizes. Inference procedures for two-sample data differ from those for matched pairs.

We can present two-sample data graphically by a back-to-back stemplot (for small samples) or by side-by-side boxplots (for larger samples). Now we will apply the ideas of formal inference in this setting. When both population distributions are symmetric, and especially when they are at least approximately Normal, a comparison of the mean responses in the two populations is most often the goal of inference.

We have two independent samples, from two distinct populations (such as failed businesses and successful businesses). We measure the same variable (such as initial capital) in both samples. We will call the variable x_1 in the first population and x_2 in the second because the variable may have different distributions in the two populations. Here is the notation that we will use to describe the two populations:

Population	Variable	Mean	Standard deviation
1	x_1	μ_1	σ_1
2	x_2	μ_2	σ_2

We want to compare the two population means, either by giving a confidence interval for $\mu_1 - \mu_2$ or by testing the hypothesis of no difference, $H_0: \mu_1 = \mu_2$. We base inference on two independent SRSs, one from each population. Here is the notation that describes the samples:

Population	Sample size	Sample mean	Sample standard deviation
1	n_1	\bar{x}_1	s_1
2	n_2	\bar{x}_2	s_2

Throughout this section, the subscripts 1 and 2 show the population to which a parameter or a sample statistic refers.

The two-sample z statistic

The natural estimator of the difference $\mu_1 - \mu_2$ is the difference between the sample means, $\bar{x}_1 - \bar{x}_2$. If we are to base inference on this statistic, we must know its sampling distribution. Our knowledge of probability is equal to the task. First, the mean of the difference $\bar{x}_1 - \bar{x}_2$ is the difference of the means $\mu_1 - \mu_2$. This follows from the addition rule for means (page 266) and the fact that the mean of any \bar{x} is the same as the mean of the population. Because the samples are independent, their sample means \bar{x}_1 and \bar{x}_2 are independent random variables. The addition rule for variances (page 272) says that the variance of the difference $\bar{x}_1 - \bar{x}_2$ is the sum of their variances, which is

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

We now know the mean and variance of the distribution of $\bar{x}_1 - \bar{x}_2$ in terms of the parameters of the two populations. If the two population distributions are both Normal, then the distribution of $\bar{x}_1 - \bar{x}_2$ is also Normal. This is true because each sample mean alone is Normally distributed and because a difference of independent Normal random variables is also Normal.

Any Normal random variable has the $N(0, 1)$ distribution when standardized. We have arrived at a new z statistic.

TWO-SAMPLE z STATISTIC

Suppose that \bar{x}_1 is the mean of an SRS of size n_1 drawn from an $N(\mu_1, \sigma_1)$ population and that \bar{x}_2 is the mean of an independent SRS of size n_2 drawn from an $N(\mu_2, \sigma_2)$ population. Then the **two-sample z statistic**

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has the standard Normal $N(0, 1)$ sampling distribution.

In the unlikely event that both population standard deviations are known, the two-sample z statistic is the basis for inference about $\mu_1 - \mu_2$. Exact z procedures are seldom used, however, because σ_1 and σ_2 are rarely known. In Chapter 6, we discussed the one-sample z procedures to introduce the ideas of inference. Now we pass immediately to the more useful t procedures.

The two-sample t procedures

In practice, the two population standard deviations σ_1 and σ_2 are not known. We estimate them by the sample standard deviations s_1 and s_2 from our two samples. Following the pattern of the one-sample case, we substitute the standard errors $s_i/\sqrt{n_i}$ for the standard deviations $\sigma_i/\sqrt{n_i}$ in the two-sample z statistic. The result is the *two-sample t statistic*:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Unfortunately, this statistic does *not* have a t distribution. A t distribution replaces an $N(0, 1)$ distribution only when a single standard deviation (σ) in a z statistic is replaced by an estimate (s). In this case, we replaced two standard deviations (σ_1 and σ_2) by their estimates (s_1 and s_2), which does not produce a statistic having a t distribution.

Nonetheless, we can approximate the distribution of the two-sample t statistic by using the $t(k)$ distribution with an approximation for the degrees of freedom k . We use these approximations to find approximate values of

Satterthwaite approximation

t^* for confidence intervals and to find approximate P -values for significance tests. There are two procedures used in practice:

1. Use an approximation known as the **Satterthwaite approximation** to calculate a value of k from the data. In general, this k will not be a whole number.
2. Use degrees of freedom k equal to the smaller of $n_1 - 1$ and $n_2 - 1$.

Most statistical software uses the Satterthwaite approximation for two-sample problems unless the user requests another method. Use of this approximation without software is a bit complicated; we will give the details later in this section. If you are not using software, we recommend the second approximation. This approximation is appealing because it is conservative.¹⁴ That is, margins of error for confidence intervals are a bit larger than they need to be, so the true confidence level is larger than C ; P -values are a bit smaller than the exact truth, so we are a little less likely to reject H_0 when it is true. In practice, the choice of approximation almost never makes a difference in our practical conclusion.

The two-sample t significance test

THE TWO-SAMPLE t SIGNIFICANCE TEST

Draw an SRS of size n_1 from a Normal population with unknown mean μ_1 and an independent SRS of size n_2 from another Normal population with unknown mean μ_2 . To test the hypothesis $H_0: \mu_1 = \mu_2$, compute the **two-sample t statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and use P -values or critical values for the $t(k)$ distribution, where the degrees of freedom k are either approximated by software or are the smaller of $n_1 - 1$ and $n_2 - 1$.

EXAMPLE 7.11

Is our product effective?

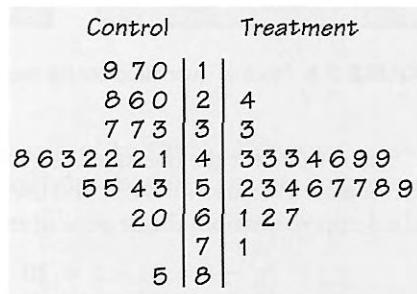
A company that sells educational materials reports statistical studies to convince customers that its materials improve learning. One new product supplies “directed reading activities” for classroom use. These activities should improve the reading ability of elementary school pupils.

A consultant arranges for a third-grade class of 21 students to take part in these activities for an eight-week period. A control classroom of 23 third-graders follows the same curriculum without the activities. At the end of the eight weeks, all students are given a Degree of Reading Power (DRP) test, which measures the aspects of reading ability that the treatment is designed to improve. The data appear in Table 7.3.¹⁵

TABLE 7.3 DRP scores for third-graders

Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

First examine the data:



A back-to-back stemplot suggests that there is a mild outlier in the control group but no deviation from Normality serious enough to forbid use of t procedures. Separate Normal quantile plots for both groups (Figure 7.8) confirm that both are approximately Normal. The scores of the treatment group appear to be somewhat higher than those of the control group. The summary statistics are

Group	n	\bar{x}	s
Treatment	21	51.48	11.01
Control	23	41.52	17.15

Because we hope to show that the treatment (Group 1) is better than the control (Group 2), the hypotheses are

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 > \mu_2$$

The two-sample t test statistic is

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{51.48 - 41.52}{\sqrt{\frac{11.01^2}{21} + \frac{17.15^2}{23}}} = 2.31 \end{aligned}$$

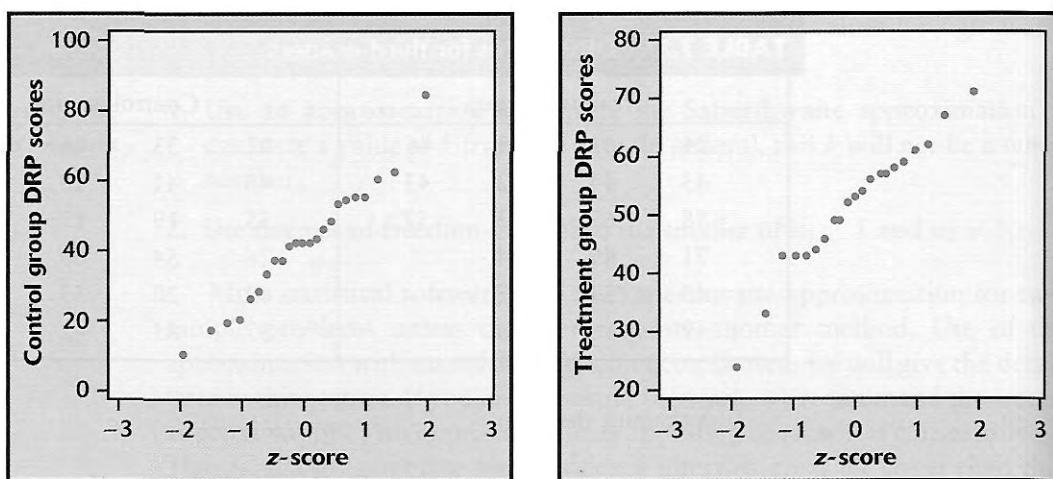


FIGURE 7.8 Normal quantile plots for the DRP scores in Table 7.3.

The P -value for the one-sided test is $P(T \geq 2.31)$. Software gives the approximate P -value as 0.0132 and uses 37.9 as the degrees of freedom. For the second approximation, the degrees of freedom k are equal to the smaller of

$$n_1 - 1 = 21 - 1 = 20 \text{ and } n_2 - 1 = 23 - 1 = 22$$

Comparing $t = 2.31$ with the entries in Table D for 20 degrees of freedom, we see that P lies between 0.02 and 0.01. The data strongly suggest that directed reading activity improves the DRP score ($t = 2.31$, df = 20, $P < 0.02$).

df = 20		
p	0.02	0.01
t^*	2.197	2.528

Some software gives P -values only for the two-sided alternative. In this case, divide the reported value by 2 to get the one-sided P -value. Be sure to check that the means differ in the direction specified by the alternative hypothesis.

- APPLY YOUR KNOWLEDGE**
- 7.51 **Selling on the Internet.** You want to compare the daily sales for two different designs of Web pages for your Internet business. You assign the next 60 days to either design A or design B, 30 days to each.
- Would you use a one-sided or a two-sided significance test for this problem? Explain your choice.
 - If you use Table D to find the critical value, what are the degrees of freedom?
 - The t statistic for comparing the mean sales is 3.15. If you use Table D, what P -value would you report? What would you conclude?
- 7.52 If you perform the significance test in the previous exercise using level $\alpha = 0.05$, how large (positive or negative) must the t statistic be to reject the null hypothesis that the two designs give the same average sales?
- 7.53 **Selling on the Internet.** You realize that day of the week may affect online sales. To compare two Web page designs, you choose two successive weeks in the middle of a month. You flip a coin to assign one Monday to the first design and the other Monday to the second. You repeat this for each of the seven days of the week. You now have 7 sales amounts for each design. It

is *incorrect* to use the two-sample t test to see whether the mean sales differ for the two designs. Carefully explain why.

The two-sample t confidence interval

The same ideas that we used for the two-sample t significance tests also apply to give us *two-sample t confidence intervals*. We can use either software or the conservative approach with Table D to approximate the value of t^* .

THE TWO-SAMPLE t CONFIDENCE INTERVAL

Draw an SRS of size n_1 from a Normal population with unknown mean μ_1 and an independent SRS of size n_2 from another Normal population with unknown mean μ_2 . The **confidence interval for $\mu_1 - \mu_2$** given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

has confidence level at least C no matter what the population standard deviations may be. The margin of error is

$$t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Here, t^* is the value for the $t(k)$ density curve with area C between $-t^*$ and t^* . The value of the degrees of freedom k is approximated by software or we use the smaller of $n_1 - 1$ and $n_2 - 1$.

To complete the analysis of the DRP scores we examined in Example 7.11, we need to describe the size of the treatment effect. We do this with a confidence interval for the difference between the treatment group and the control group means.

EXAMPLE 7.12

How much improvement?

We will find a 95% confidence interval for the mean improvement in the entire population of third-graders. The interval is

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) &\pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (51.48 - 41.52) \pm t^* \sqrt{\frac{11.01^2}{21} + \frac{17.15^2}{23}} \\ &= 9.96 \pm (t^* \times 4.31) \end{aligned}$$

Using software, the degrees of freedom are 37.9 and $t^* = 2.025$. This approximation gives

$$9.96 \pm (2.025 \times 4.31) = 9.96 \pm 8.72 = (1.2, 18.7)$$

The conservative approach uses the $t(20)$ distribution. Table D gives $t^* = 2.086$. With this approximation we have

$$9.96 \pm (2.086 \times 4.31) = 9.96 \pm 8.99 = (1.0, 18.9)$$

The conservative approach does give a larger interval than the more accurate approximation used by software. However, the difference is rather small.

We estimate the mean improvement in DRP scores to be about 10 points, but with a margin of error of almost 9 points. Although we have good evidence of some improvement, the data do not allow a very precise estimate of the size of the average improvement.

The design of the DRP study is not ideal. Random assignment of students was not possible in a school environment, so existing third-grade classes were used. The effect of the reading programs is therefore confounded with any other differences between the two classes. The classes were chosen to be as similar as possible in variables such as the social and economic status of the students. Pretesting showed that the two classes were on the average quite similar in reading ability at the beginning of the experiment. To avoid the effect of two different teachers, the same teacher taught reading in both classes during the eight-week period of the experiment. We can therefore be somewhat confident that the two-sample test is detecting the effect of the treatment and not some other difference between the classes. This example is typical of many situations in which an experiment is carried out but randomization is not possible.

APPLY YOUR KNOWLEDGE

- 7.54 **Flat screens?** The purchasing department has suggested that all new computer monitors for your company should be flat screens. You want data to assure you that employees will like the new screens. The next 20 employees needing a new computer are the subjects for an experiment.
- Label the employees 01 to 20. Randomly choose 10 to receive flat screens. The remaining 10 get standard monitors.
 - After a month of use, employees express their satisfaction with their new monitors by responding to the statement “I like my new monitor” on a scale from 1 to 5, where 1 represents “strongly disagree,” 2 is “disagree,” 3 is “neutral,” 4 is “agree,” and 5 stands for “strongly agree.” The employees with the flat screens have average satisfaction 4.6 with standard deviation 0.7. The employees with the standard monitors have average 3.2 with standard deviation 1.8. Give a 95% confidence interval for the difference in the mean satisfaction scores for all employees.
 - Would you reject the null hypothesis that the mean satisfaction for the two types of monitors is the same versus the two-sided alternative at significance level 0.05? Use your confidence interval to answer this question. Explain why you do not need to calculate the test statistic.
- 7.55 **Why randomize?** A coworker suggested that you give the flat screens to the next 10 employees who need new screens and the standard monitor to the following 10. Explain why your randomized design is better.

Robustness of the two-sample procedures

The two-sample t procedures are more robust than the one-sample t methods. When the sizes of the two samples are equal and the distributions of the two

populations being compared have similar shapes, probability values from the t table are quite accurate for a broad range of distributions when the sample sizes are as small as $n_1 = n_2 = 5$.¹⁶ When the two population distributions have different shapes, larger samples are needed. The guidelines given on page 447 for the use of one-sample t procedures can be adapted to two-sample procedures by replacing “sample size” with the “sum of the sample sizes” $n_1 + n_2$. These guidelines are rather conservative, especially when the two samples are of equal size. In planning a two-sample study, you should usually choose equal sample sizes. The two-sample t procedures are most robust against non-Normality in this case, and the conservative probability values are most accurate.

EXAMPLE 7.13**Wheat prices**

The U.S. Department of Agriculture (USDA) uses sample surveys to produce important economic estimates. One pilot study estimated wheat prices in July and in September using independent samples of wheat producers in the two months. Here are the summary statistics, in dollars per bushel:¹⁷

Month	n	\bar{x}	s
September	45	\$3.61	\$0.19
July	90	\$2.95	\$0.22

The September prices are higher on the average. But we have data from only a few producers each month. Can we conclude that national average prices in July and September are not the same? Or are these differences merely what we would expect to see due to random variation?

Because we did not specify a direction for the difference before looking at the data, we choose a two-sided alternative. The hypotheses are

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

Because the samples are moderately large, we can confidently use the t procedures even though we lack the detailed data and so cannot verify the Normality condition.

The two-sample t statistic is

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{3.61 - 2.95}{\sqrt{\frac{(0.19)^2}{45} + \frac{(0.22)^2}{90}}} \\ &= 18.03 \end{aligned}$$

	df = 40
<i>p</i>	0.0005
<i>t</i> *	3.551

The conservative approach finds the *P*-value by comparing 18.03 to critical values for the $t(44)$ distribution because the smaller sample has 45 observations. We must double the table tail area *p* because the alternative is two-sided.

Table D does not have entries for 44 degrees of freedom. When this happens, we use the next smaller degrees of freedom. Our calculated value of *t* is larger than the *p* = 0.0005 entry in the table. Doubling 0.0005, we conclude that the *P*-value is less than 0.001. The data give conclusive evidence that the mean wheat prices were higher in September than they were in July ($t = 18.03$, df = 44, $P < 0.001$).

In this example the exact *P*-value is very small because $t = 18$ says that the observed mean is 18 standard deviations above the hypothesized mean. This is so unlikely that the probability is zero for all practical purposes. The difference in mean prices is not only highly significant but large enough (66 cents per bushel) to be important to producers.

In this and other examples, we can choose which population to label 1 and which to label 2. After inspecting the data, we chose September as Population 1 because this choice makes the *t* statistic a positive number. This avoids any possible confusion from reporting a negative value for *t*. Choosing the population labels is *not* the same as choosing a one-sided alternative after looking at the data. Choosing hypotheses after seeing a result in the data is a violation of sound statistical practice.

Inference for small samples

Small samples require special care. We do not have enough observations to examine the distribution shapes, and only extreme outliers stand out. The power of significance tests tends to be low, and the margins of error of confidence intervals tend to be large. Despite these difficulties, we can often draw important conclusions from studies with small sample sizes. If the size of an effect is as large as it was in the wheat price example, it should still be evident even if the *n*'s are small.

EXAMPLE 7.14

More about wheat prices

In the setting of Example 7.13, a quick survey collects prices from only 5 producers each month. The data are

Month	Price of wheat (\$/bushel)				
September	\$3.5900	\$3.6150	\$3.5950	\$3.5725	\$3.5825
July	\$2.9200	\$2.9675	\$2.9175	\$2.9250	\$2.9325

The prices are reported to the nearest quarter of a cent. First, examine the distributions with a back-to-back stemplot after rounding each price to the nearest cent.

September	July
2.9	2 2 3 3 7
3.0	
3.1	
3.2	
3.3	
3.4	
9 8 7	3.5
2 0	3.6

The pattern is clear. There is little variation among prices within each month, and the distributions for the two months are far apart relative to the within-month variation.

A significance test can confirm that the difference between months is too large to easily arise just by chance. We test

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

The price is higher in September ($t = 56.99$, $df = 7.55$, $P < 0.0001$). The difference in sample means is 65.9 cents.

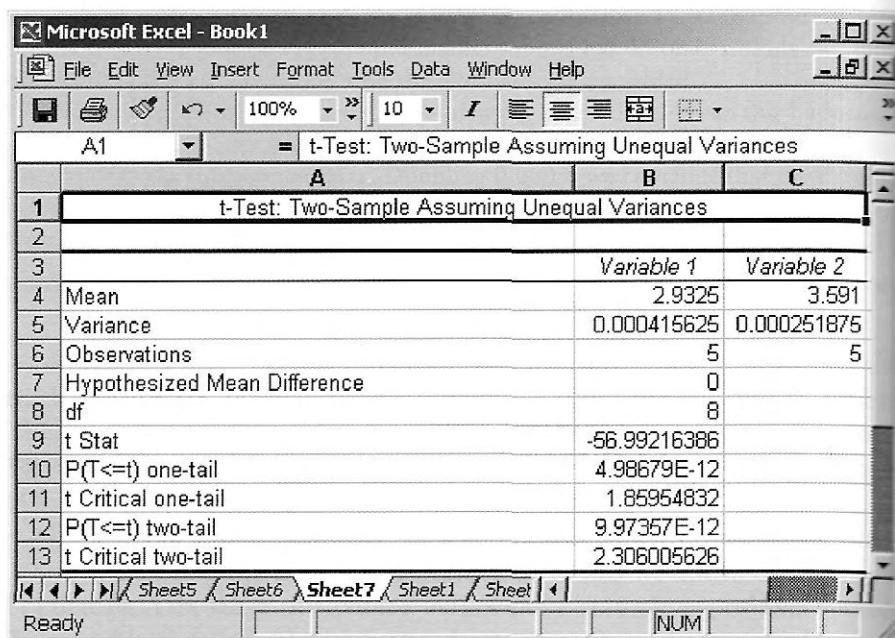
Figure 7.9 gives outputs for this analysis from several software systems. Although the formats differ, the basic information is the same. All report the sample sizes, the sample means and standard deviations (or variances), the t statistic, and its P -value. All agree that the P -value is very small, though some give more detail than others. Software often labels the groups in alphabetical order. In this example, July is then the first population and $t = -56.99$, the negative of our result. Always check the means first and report the statistic (you may need to change the sign) in an appropriate way. Be sure to also mention the size of the effect you observed, such as “The mean price for September was 65.9 cents higher than in July.”

SPSS and SAS report the results of *two t* procedures: a special procedure that assumes that the two population variances are equal and the general two-sample procedure that we have just studied. We don’t recommend the “equal-variances” procedures, but we describe them later, in the section on pooled two-sample t procedures.

Satterthwaite approximation for the degrees of freedom*

We noted earlier that the two-sample t statistic does not have an exact t distribution. Moreover, the exact distribution changes as the unknown

*This material can be omitted unless you are using statistical software and wish to understand what the software does.



SPSS						
Group Statistics						
	MONTH	N	Mean	Std. Deviation	Std. Error Mean	
PRICE	JULY	5	2.932500	2.03869E-02	9.11729E-03	
	SEPT	5	3.591000	1.58706E-02	7.09753E-03	

Independent Samples Test						
t-test for Equality of Means						
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
PRICE	Equal variances assumed	-56.99	8	.000	-.6585	1.155E-02
	Equal variances not assumed	-56.99	7.546	.000	-.6585	1.155E-02

FIGURE 7.9 Excel and SPSS output for Example 7.14 (continued).

Minitab**Two Sample T-Test and Confidence Interval**

Two sample T for Price

MONTH	N	Mean	StDev	SE MEAN
JULY	5	2.9325	0.0204	0.0091
SEPT	5	3.5910	0.0159	0.0071

95% CI for mu (JULY) - mu (SEPT): (-0.6858, -0.6312)
T-Test mu (JULY) = mu (SEPT) (vs not =): T = -56.99 P = 0.0000 DF = 7

SAS

The TTEST Procedure

Variable	month	Statistics					
		Lower CL		Upper CL		Mean	Std Err
		N	Mean	Mean	Mean		
price	july	5	2.9072	2.9325	2.9578	0.0091	
price	sept	5	3.5713	3.591	3.6107	0.0071	
price	Diff (1-2)		-0.685	-0.659	-0.632	0.0116	
T-Tests							
Variable	Method	Variances		DF	t Value	Pr > t	
price	Pooled	Equal		8	-56.99	< .0001	
price	Satterthwaite	Unequal		7.55	-56.99	< .0001	

FIGURE 7.9 (continued) Minitab and SAS output for Example 7.14.

population standard deviations σ_1 and σ_2 change. However, the distribution can be approximated by a t distribution with degrees of freedom given by

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2} \right)^2}$$

Most statistical software uses this approximation. It was discovered by a statistician named Satterthwaite, and his name is frequently used to identify the approximation on output. The Satterthwaite approximation is quite accurate when both sample sizes n_1 and n_2 are 5 or larger.

EXAMPLE 7.15

Output for the wheat prices example with samples of size 5 appears in Figure 7.9. SPSS and SAS report the degrees of freedom for the Satterthwaite approximation as 7.546 and 7.55. Excel and Minitab round this to a whole number. All of these df's are larger than the conservative $df = 4$ that we would use in the absence of

software. In this example, the t statistic is so large that all choices lead to the same practical conclusion.



The number df given by the Satterthwaite approximation is always at least as large as the smaller of $n_1 - 1$ and $n_2 - 1$. On the other hand, df is never larger than the sum $n_1 + n_2 - 2$ of the two individual degrees of freedom. The number of degrees of freedom is generally not a whole number. There is a t distribution with any positive degrees of freedom, even though Table D contains entries only for whole-number degrees of freedom. Because of this and the need to calculate df , we do not recommend using the approximation unless a computer is doing the arithmetic. With software, however, the more accurate procedures are painless.

APPLY YOUR KNOWLEDGE

- 7.56 **Calculating the Satterthwaite df .** The SPSS output in Figure 7.9 gives $df = 7.546$. Use the sample standard deviations in that output and the Satterthwaite formula to verify this value.
- 7.57 **Compare the output.** Figure 7.9 gives the output from four software systems for the two-sample t procedures. Make tables to compare the outputs, giving the numerical values reported by the software.
 - (a) How do they report the means?
 - (b) How do they report variability for each of the groups?
 - (c) How do they report the test statistic, degrees of freedom, and the P value?
 - (d) Do they give a confidence interval for the mean difference?
 - (e) Write a short summary comparing the four outputs from the viewpoint of a user of statistics. Which do you like the best? Which do you like the least?
- 7.58 **Can you do better?** Design your own output for a two-sample t analysis and illustrate it with the results given in Figure 7.9. Pay particular attention to the use of labels and how numbers are rounded.

The pooled two-sample t procedures*

There is one situation in which a t statistic for comparing two means has exactly a t distribution. Suppose that the two Normal population distributions have the *same* standard deviation. In this case we need substitute only a single standard error in a z statistic, and the resulting t statistic has a t distribution. We will develop the z statistic first, as usual, and from it the t statistic.

Call the common—but still unknown—standard deviation of both populations σ . Both sample variances s_1^2 and s_2^2 estimate σ^2 . The best way to combine these two estimates is to average them with weights equal to their

*This section is optional, but you should read it if you plan to study Chapters 15 and 16 on analysis of variance.

degrees of freedom. This gives more weight to the information from the larger sample. The resulting estimator of σ^2 is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

pooled estimator
of σ^2

This is called the pooled estimator of σ^2 because it combines the information in both samples.

When both populations have variance σ^2 , the addition rule for variances says that $\bar{x}_1 - \bar{x}_2$ has variance equal to the sum of the individual variances, which is

$$\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

The standardized difference of means in this equal-variance case is therefore

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

This is a special two-sample z statistic for the case in which the populations have the same σ . Replacing the unknown σ by the estimate s_p gives a t statistic. The degrees of freedom are $n_1 + n_2 - 2$, the sum of the degrees of freedom of the two sample variances. This statistic is the basis of the pooled two-sample t inference procedures.

THE POOLED TWO-SAMPLE t PROCEDURES

Draw an SRS of size n_1 from a Normal population with unknown mean μ_1 and an independent SRS of size n_2 from another Normal population with unknown mean μ_2 . Suppose that the two populations have the same unknown standard deviation. A level C confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Here t^* is the value for the $t(n_1 + n_2 - 2)$ density curve with area C between $-t^*$ and t^* .

To test the hypothesis $H_0: \mu_1 = \mu_2$, compute the pooled two-sample t statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

and use P -values from the $t(n_1 + n_2 - 2)$ distribution.

TABLE 7.4 Ratio of current assets to current liabilities

Healthy firms						Failed firms		
1.50	0.10	1.76	1.14	1.84	2.21	0.82	0.89	1.31
2.08	1.43	0.68	3.15	1.24	2.03	0.05	0.83	0.90
2.23	2.50	2.02	1.44	1.39	1.64	1.68	0.99	0.62
0.89	0.23	1.20	2.16	1.80	1.87	0.91	0.52	1.45
1.91	1.67	1.87	1.21	2.05	1.06	1.16	1.32	1.17
0.93	2.17	2.61	3.05	1.52	1.93	0.42	0.48	0.93
1.95	2.61	1.11	0.95	0.96	2.25	0.88	1.10	0.23
2.73	1.56	2.73	0.90	2.12	1.42	1.11	0.19	0.13
1.62	1.76	2.22	2.80	1.85	0.96	2.03	0.51	1.12
1.71	1.02	2.50	1.55	1.69	1.64	0.92	0.26	1.15
1.03	1.80	0.67	2.44	2.30	2.21	0.13	0.88	0.09
1.96	1.81							

CASE 7.2**HEALTHY COMPANIES VERSUS FAILED COMPANIES**

In what ways are companies that fail different from those that continue to do business? To answer this question, one study compared various characteristics of 68 healthy and 33 failed firms.¹⁸ One of the variables was the ratio of current assets to current liabilities. Roughly speaking, this is the amount that the firm is worth divided by what it owes. The data appear in Table 7.4.



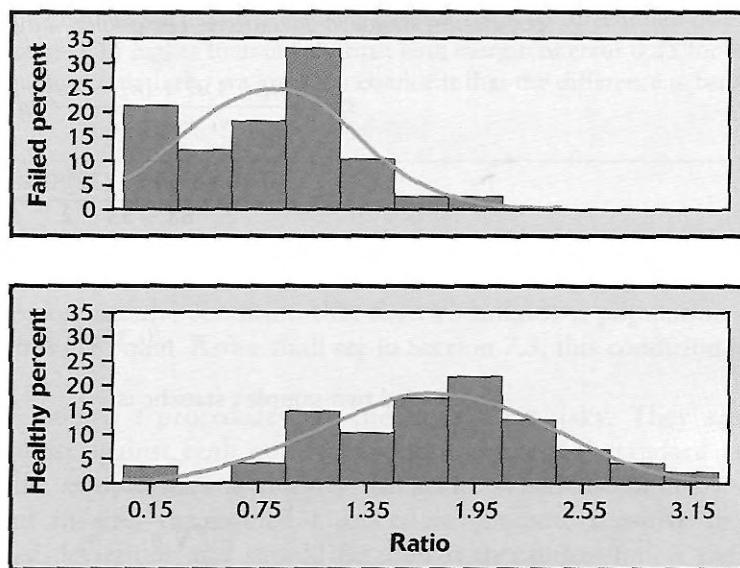
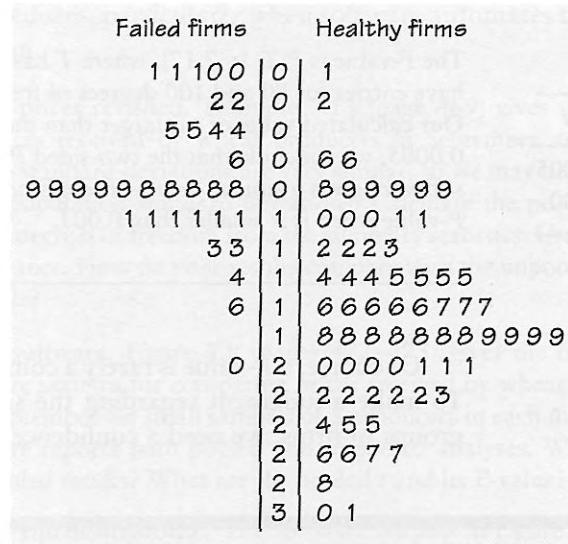
As usual, we first examine the data. Histograms for the two groups of firms are given in Figure 7.10. Normal curves with mean and standard deviation equal to the sample values are superimposed on the histograms. The distribution for the healthy firms looks more Normal than the distribution for the failed firms. It appears that there may actually be two subgroups of data for the failed firms. This is seen more clearly in the back-to-back stemplot given in Figure 7.11. However, there are no outliers or strong departures from Normality that will prevent us from using the *t* procedures for these data. Let's compare the mean current assets to current liabilities ratio for the two groups of firms using a significance test.

EXAMPLE 7.16**CASE 7.2****Do mean asset/liability ratios differ?**

Take Group 1 to be the firms that were healthy and Group 2 to be those that failed. The question of interest is whether or not the mean ratio of current assets to current liabilities is different for the two groups. We therefore test

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

**FIGURE 7.10** Histograms for assets to liabilities ratios, for Example 7.16.**FIGURE 7.11** Back-to-back stemplot for assets to liabilities ratios, for Example 7.16.

Here are the summary statistics:

Group	Firms	n	\bar{x}	s
1	Healthy	68	1.7256	0.6393
2	Failed	33	0.8236	0.4811

The sample standard deviations are fairly close. A difference this large is not particularly unusual even in samples this large. We are willing to assume equal

population standard deviations. The pooled sample variance is

$$\begin{aligned}s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\&= \frac{(67)(0.6393)^2 + (32)(0.4811)^2}{68 + 33 - 2} = 0.35141\end{aligned}$$

so that

$$s_p = \sqrt{0.35141} = 0.5928$$

The pooled two-sample t statistic is

$$\begin{aligned}t &= \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\&= \frac{1.7256 - 0.8236}{0.5928 \sqrt{\frac{1}{68} + \frac{1}{33}}} = 7.17\end{aligned}$$

The P -value is $P(T \geq 7.17)$, where T has the $t(99)$ distribution. In Table D we have entries for 80 and 100 degrees of freedom. We will use the entries for 100. Our calculated value of t is larger than the $p = 0.0005$ entry in the table. Doubling 0.0005, we conclude that the two-sided P -value is less than 0.001. Statistical software gives a similar result—there is no practical need to report the exact P -value when it is smaller than 0.001.

df = 100
p
t^*

0.0005
3.300

Of course, a P -value is rarely a complete summary of a statistical analysis. To make a judgment regarding the size of the difference between the two groups of firms, we need a confidence interval.

EXAMPLE 7.17



How different are mean asset/liability ratios?

The difference in mean current assets to current liabilities ratios for healthy versus failed firms is

$$\bar{x}_1 - \bar{x}_2 = 1.7256 - 0.8236 = 1.151$$

For a 95% margin of error we will use the critical value $t^* = 1.984$ from the $t(100)$ distribution. The margin of error is

$$\begin{aligned}t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &= (1.984)(0.5928) \sqrt{\frac{1}{68} + \frac{1}{33}} \\&= 0.249\end{aligned}$$

We report that the successful firms have current assets to current liabilities ratios that average 1.15 higher than failed firms, with margin of error 0.25 for 95% confidence. Alternatively, we are 95% confident that the difference is between 0.90 and 1.40.

The pooled two-sample t procedures are anchored in statistical theory and have long been the standard version of the two-sample t in textbooks. But they require the condition that the two unknown population standard deviations are equal. As we shall see in Section 7.3, this condition is hard to verify.

The pooled t procedures are therefore a bit risky. They are reasonably robust against both non-Normality and unequal standard deviations when the sample sizes are nearly the same. When the samples are quite different in size, the pooled t procedures become sensitive to unequal standard deviations and should be used with caution unless the samples are large. Unequal standard deviations are quite common. In particular, it is common for the spread of data to increase when the center moves up, as in the data given in Case 7.2. We recommend regular use of the unpooled t procedures, particularly when software automates the Satterthwaite approximation.

APPLY YOUR KNOWLEDGE

- 7.59 **Wheat prices revisited.** Example 7.13 (page 469) gives summary statistics for prices received by wheat producers in September and July. The two sample standard deviations are very similar, so we may be willing to assume equal population standard deviations. Calculate the pooled t test statistic and its degrees of freedom from the summary statistics. Use Table D to assess significance. How do your results compare with the unpooled analysis in the example?
- 7.60 **Using software.** Figure 7.9 (pages 472–473) gives the outputs from four software systems for comparing prices received by wheat producers in July and September for small samples of 5 producers in each month. Some of the software reports both pooled and unpooled analyses. Which outputs give the pooled results? What are the pooled t and its P -value?
- 7.61 **Pooled equals unpooled?** The software outputs in Figure 7.9 give the *same value* for the pooled and unpooled t statistics. Do some simple algebra to show that this is always true when the two sample sizes n_1 and n_2 are the same. In other cases, the two t statistics usually differ.

SECTION 7.2 SUMMARY

- Significance tests and confidence intervals for the difference of the means μ_1 and μ_2 of two Normal populations are based on the difference $\bar{x}_1 - \bar{x}_2$ of the sample means from two independent SRSs. Because of the central limit theorem, the resulting procedures are approximately correct for other population distributions when the sample sizes are large.

- When independent SRSs of sizes n_1 and n_2 are drawn from two Normal populations with parameters μ_1 , σ_1 and μ_2 , σ_2 the two-sample z statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has the $N(0, 1)$ distribution.

- The two-sample t statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

does *not* have a t distribution. However, software can give accurate P -values and critical values using the **Satterthwaite approximation**.

- Conservative inference procedures for comparing μ_1 and μ_2 use the two-sample t statistic and the $t(k)$ distribution with degrees of freedom k equal to the smaller of $n_1 - 1$ and $n_2 - 1$. Use this method unless you are using software.
- An approximate level C confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Here, t^* is the value for the $t(k)$ density curve with area C between $-t^*$ and t^* , where k either is found by the Satterthwaite approximation or is the smaller of $n_1 - 1$ and $n_2 - 1$. The margin of error is

$$t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Significance tests for $H_0: \mu_1 = \mu_2$ are based on the two-sample t statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- The P -value is approximated using the $t(k)$ distribution, where k either is found by the Satterthwaite approximation or is the smaller of $n_1 - 1$ and $n_2 - 1$.
- The guidelines for practical use of two-sample t procedures are similar to those for one-sample t procedures. Equal sample sizes are recommended.

- If we can assume that the two populations have equal variances, pooled two-sample t procedures can be used. These are based on the pooled estimator

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

of the unknown common variance and the $t(n_1 + n_2 - 2)$ distribution.

9.2

SECTION 7.2 EXERCISES

In exercises that call for two-sample t procedures, you may use either of the two approximations for the degrees of freedom that we have discussed: the value given by your software or the smaller of $n_1 - 1$ and $n_2 - 1$. Be sure to state clearly which approximation you have used.

- 7.62 Healthy companies versus failed companies. Table 7.4 gives data for 68 healthy firms and 33 comparable firms that failed. The variable recorded is the ratio of a firm's current assets to its current liabilities. We expect healthy firms to have a higher ratio of assets to liabilities on the average. Do the data give good evidence in favor of this expectation? By how much on the average does the ratio for healthy firms exceed that for failed firms (use 99% confidence)? (Examples 7.16 and 7.17 analyze these data under the special assumption that the two populations of firms have the same standard deviation. In practice, we prefer not to make this assumption.)
- 7.63 How much for an extra bedroom? Tom wants to compare the cost of one- and two-bedroom apartments in the area of your campus. He collects data for a random sample of 10 advertisements of each type. Here are the rents for the two-bedroom apartments (in dollars per month):
- 595, 500, 580, 650, 675, 675, 750, 500, 495, 670
- Here are the rents for the one-bedroom apartments:
- 500, 650, 600, 505, 450, 550, 515, 495, 650, 395
- Find a 95% confidence interval for the additional cost of a second bedroom.
- 7.64 Is the difference significant? Tom wonders if two-bedroom apartments rent for significantly more than one-bedroom apartments. Use the data in the previous exercise to find out.
- (a) State appropriate null and alternative hypotheses.
 - (b) Report the test statistic, its degrees of freedom, and the P -value. What do you conclude?
 - (c) Can you conclude that every one-bedroom apartment costs less than every two-bedroom apartment?
 - (d) In the previous exercise you found a confidence interval. In this exercise you performed a significance test. Which do you think is more useful to someone planning to rent an apartment? Why?
- 7.65 Effect of storage on a product. Does bread lose vitamins when stored? Researchers prepared loaves of bread with flour fortified with a known



CASE 7.2

amount of vitamins. After baking, they measured the vitamin C content of two loaves. Another two loaves were baked at the same time, stored for three days, and then the vitamin C content was measured. The units are milligrams of vitamin C per hundred grams of flour (mg/100 g). Here are the data.¹⁹

Immediately after baking:	47.62	49.79
Three days after baking:	21.25	22.34

- (a) Does bread lose vitamin C when it is stored for three days? Use a two-sample t test to answer this question. (State the hypotheses, give the test statistic, degrees of freedom, and P -value, and state your conclusion in nontechnical language.)
- (b) Give a 90% confidence interval for the amount of vitamin C lost.

- 7.66 **Study design matters!** The researchers in Exercise 7.65 might have measured the same two loaves of bread immediately after baking and again after three days. Suppose that the data given had come from this study design. (The values are for first loaf and second loaf from right to left.)
- (a) Explain carefully why your analysis from Exercise 7.65 is *not correct* now, even though the data are the same.
- (b) Redo the analysis for the design based on measuring the same loaves twice.

- 7.67 **Another ingredient.** The researchers of Exercise 7.65 also measured the amount of vitamin E (in mg/100 g of flour) in the same loaves. Here are the data:

Immediately after baking:	94.6	96.0
Three days after baking:	97.4	94.3

- (a) Does bread lose vitamin E in short-term storage? State hypotheses for a statistical test, give the test statistic, degrees of freedom, and P -value, and state your conclusion.
- (b) Give a 90% confidence interval for the amount of vitamin E lost.

- 7.68 **Are the samples too small?** Exercises 7.65 and 7.67 are based on samples of just two loaves of bread. Some people claim that significance tests with very small samples never lead to rejection of the null hypothesis. Discuss this claim using the results of these two exercises.

- 7.69 **Interpreting software output.** You use statistical software to perform a significance test of the null hypothesis that two means are equal. The software reports P -values for the two-sided alternative. Your alternative is that the first mean is greater than the second mean.
- (a) The software reports $t = 2.07$ with a P -value of 0.06. Would you reject H_0 with $\alpha = 0.05$? Explain your answer.
- (b) The software reports $t = -2.07$ with a P -value of 0.06. Would you reject H_0 with $\alpha = 0.05$? Explain your answer.

- 7.70 **Piano lessons.** Do piano lessons improve the spatial-temporal reasoning of preschool children? We examined this question in Exercises 7.34 and 7.35 (page 456) by analyzing the change in spatial-temporal reasoning of 34 preschool children after six months of piano lessons. Here we examine the same question by comparing the changes of those students with the changes

of 44 children in a control group.²⁰ Here are the data for the children who took piano lessons:

2	5	7	-2	2	7	4	1	0	7	3	4	3	4	9	4	5
2	9	6	0	3	6	-1	3	4	6	7	-2	7	-3	3	4	4

The control group scores are

1	-1	0	1	-4	0	0	1	0	-1	0	1	1	-3	-2	4	-1
2	4	2	2	2	-3	-3	0	2	0	-1	3	-1	5	-1	7	0
4	0	2	1	-6	0	2	-1	0	-2							

- (a) Display the data and summarize the distributions.
- (b) Make a table with the sample size, the mean, the standard deviation, and the standard error of the mean for each of the two groups.
- (c) Translate the question of interest into hypotheses, test them, and summarize your conclusions.

7.71 **Piano lessons, continued.** The previous exercise gives data from a study of the effects of piano lessons. Give a 95% confidence interval that describes the comparison between the children who took piano lessons and the controls.

7.72 **Comparing several statistical approaches.** Exercises 7.34 and 7.35 (page 456) and Exercises 7.70 and 7.71 all address the effects of piano lessons on spatial-temporal reasoning. Discuss the relative merits of each approach. (You need not actually do all four exercises.)

7.73 **Fitness and ego.** Employers sometimes seem to prefer executives who appear physically fit, despite the legal troubles that may result. Employers may also favor certain personality characteristics. Fitness and personality are related. In one study, middle-aged college faculty who had volunteered for a fitness program were divided into low-fitness and high-fitness groups based on a physical examination. The subjects then took the Cattell Sixteen Personality Factor Questionnaire.²¹ Here are the data for the “ego strength” personality factor:

Low fitness			High fitness		
4.99	5.53	3.12	6.68	5.93	5.71
4.24	4.12	3.77	6.42	7.08	6.20
4.74	5.10	5.09	7.32	6.37	6.04
4.93	4.47	5.40	6.38	6.53	6.51
4.16	5.30		6.16	6.68	

- (a) Is the difference in mean ego strength significant at the 5% level? At the 1% level? Be sure to state H_0 and H_a .
- (b) You should hesitate to generalize these results to the population of all middle-aged men. Explain why.
- (c) You should also hesitate to conclude that increasing fitness *causes* an increase in ego strength. Explain why.

7.74 **Compare the effectiveness of two products.** In a study of cereal leaf beetle damage on oats, researchers measured the number of beetle larvae per stem

in small plots of oats after randomly applying one of two treatments: no pesticide or Malathion at the rate of 0.25 pound per acre. Here are the data.²²

Control:	2	4	3	4	2	3	3	5	3	2	6	3	4
Treatment:	0	1	1	2	1	2	1	1	2	1	1	1	

- (a) Is there significant evidence at the 1% level that the mean number of larvae per stem is reduced by Malathion? Be sure to state H_0 and H_a .
- (b) These data are far from Normal. Why? The researchers nonetheless used t procedures. Although we might prefer a different approach, t procedures are probably reasonably accurate here. Why?
- 7.75 **Study design matters!** In Exercise 7.73 you analyzed data on the ego strength of high-fitness and low-fitness participants in an campus fitness program. Suppose that instead you had data on the ego strengths of the *same* men before and after six months in the program. You wonder if the program has affected their ego scores. Explain carefully how the statistical procedures you would use would differ from those you applied in Exercise 7.73.
- 7.76 **Cocaine use and low birth weight.** Does cocaine use by pregnant women cause their babies to have low birth weight? To study this question, birth weights of babies of women who tested positive for cocaine during a drug-screening test were compared with the birth weights for women who either tested negative or were not tested, a group we call “other.”²³ Here are the summary statistics. The birth weights are measured in grams.
- | Group | n | \bar{x} | s |
|---------------|------|-----------|-----|
| Positive test | 134 | 2733 | 599 |
| Other | 5974 | 3118 | 672 |
- (a) Formulate appropriate hypotheses and carry out the test of significance for these data.
- (b) Give a 95% confidence interval for the mean difference in birth weights.
- (c) Discuss the limitations of the study design. What do you believe can be concluded from this study?
- 7.77 **Sales of small appliances.** A market research firm supplies manufacturers with estimates of the retail sales of their products from samples of retail stores. Marketing managers are prone to look at the estimate and ignore sampling error. Suppose that an SRS of 75 stores this month shows mean sales of 52 units of a small appliance, with standard deviation 13 units. During the same month last year, an SRS of 53 stores gave mean sales of 49 units, with standard deviation 11 units. An increase from 49 to 52 is a rise of 6%. The marketing manager is happy, because sales are up 6%.
- (a) Use the two-sample t procedure to give a 95% confidence interval for the difference in mean number of units sold at all retail stores.
- (b) Explain in language that the manager can understand why he cannot be confident that sales rose by 6%, and that in fact sales may even have dropped.

- 7.78 **Compare two marketing strategies.** A bank compares two proposals to increase the amount that its credit card customers charge on their cards. (The bank earns a percentage of the amount charged, paid by the stores that accept the card.) Proposal A offers to eliminate the annual fee for customers who charge \$2400 or more during the year. Proposal B offers a small percent of the total amount charged as a cash rebate at the end of the year. The bank offers each proposal to an SRS of 150 of its existing credit card customers. At the end of the year, the total amount charged by each customer is recorded. Here are the summary statistics:

Group	<i>n</i>	\bar{x}	<i>s</i>
A	150	\$1987	\$392
B	150	\$2056	\$413

- (a) Do the data show a significant difference between the mean amounts charged by customers offered the two plans? Give the null and alternative hypotheses, and calculate the two-sample *t* statistic. Obtain the *P*-value (either approximately from Table D or more accurately from software). State your practical conclusions.
- (b) The distributions of amounts charged are skewed to the right, but outliers are prevented by the limits that the bank imposes on credit balances. Do you think that skewness threatens the validity of the test that you used in (a)? Explain your answer.
- 7.79 **Study habits.** The Survey of Study Habits and Attitudes (SSHA) is a psychological test designed to measure the motivation, study habits, and attitudes toward learning of college students. These factors, along with ability, are important in explaining success in school. Scores on the SSHA range from 0 to 200. A selective private college gives the SSHA to an SRS of both male and female first-year students. The data for the women are as follows:

154	109	137	115	152	140	154	178	101
103	126	126	137	165	165	129	200	148

Here are the scores of the men:

108	140	114	91	180	115	126	92	169	146
109	132	75	88	113	151	70	115	187	104

- (a) Examine each sample graphically, with special attention to outliers and skewness. Is use of a *t* procedure acceptable for these data?
- (b) Most studies have found that the mean SSHA score for men is lower than the mean score in a comparable group of women. Test this supposition here. That is, state hypotheses, carry out the test and obtain a *P*-value, and give your conclusions.
- (c) Give a 90% confidence interval for the mean difference between the SSHA scores of male and female first-year students at this college.
- 7.80 **Summer earnings of college students.** College financial aid offices expect students to use summer earnings to help pay for college. But how large are these earnings? One college studied this question by asking a sample

of students how much they earned.²⁴ Omitting students who were not employed, 1296 responses were received. Here are the data in summary form:

Group	<i>n</i>	\bar{x}	<i>s</i>
Males	675	\$3297.91	\$2394.65
Females	621	\$2380.68	\$1815.55

- (a) Use the two-sample *t* procedures to give a 90% confidence interval for the difference between the mean summer earnings of male and female students.
- (b) The distribution of earnings is strongly skewed to the right. Nevertheless, use of *t* procedures is justified. Why?
- (c) Once the sample size was decided, the sample was chosen by taking every *k*th name from an alphabetical list of undergraduates. Is it reasonable to consider the samples as SRSs chosen from the male and female undergraduate populations?
- (d) What other information about the study would you request before accepting the results as describing all undergraduates?

*The following exercises concern optional material on the Satterthwaite approximation and the pooled two-sample *t* procedures.*

- 7.81 **Satterthwaite approximation.** Example 7.11 (page 464) reports an analysis comparing a new with a traditional method for teaching reading. Starting from the computer's results for \bar{x}_i and s_i , verify that the Satterthwaite approximation for the degrees of freedom is 37.9.
- 7.82 **Pooled procedures.** Exercise 7.54 (page 468) compares flat-screen monitors with standard screens. Reanalyze the data using the pooled procedure. Does the conclusion depend on the choice of the method? The standard deviations are quite different for these data, so we do not recommend use of the pooled procedures in this case.
- 7.83 **The advantage of pooling.** The analysis of the loss of vitamin C when bread is stored (Exercise 7.65, page 481) is a rather extreme case. There are only two observations per condition (immediately after baking and three days later). When the samples are so small, we have very little information to make a judgment about whether the population standard deviations are equal. The potential gain from pooling is large when the sample sizes are very small. Assume that we will perform a two-sided test using the 5% significance level.
 - (a) Find the critical value for the unpooled *t* test statistic that does not assume equal variances. Use the minimum of $n_1 - 1$ and $n_2 - 1$ for the degrees of freedom.
 - (b) Find the critical value for the pooled *t* test statistic.
 - (c) How does comparing these critical values show an advantage of the pooled test?
- 7.84 **The advantage of pooling.** Suppose that in the setting of the previous exercise you are interested in 95% confidence intervals for the difference rather than significance testing. Find the widths of the intervals for the two procedures (assuming or not assuming equal standard deviations). How do they compare?

7.3 Optional Topics in Comparing Distributions*

In this section we discuss two topics that are related to the procedures we have learned for inference about population means. First, it is natural to ask if we can do inference about population spread as well. The answer is yes, but there are many cautions. We also show how to find the power for the two-sample t test. This is a bit technical, but necessary if you plan to design studies.

Inference for population spread

The two most basic descriptive features of a distribution are its center and spread. In a Normal population, we measure center and spread by the mean and the standard deviation. We have described procedures for inference about population means for Normal populations and found that these procedures are often useful for non-Normal populations as well. It is natural to turn next to inference about the standard deviations of Normal populations. Our recommendation here is short and clear: Don't do it without expert advice.

There are indeed inference procedures appropriate for the standard deviations of Normal populations. We will describe the most common such procedure, the F test for comparing the spread of two Normal populations. Unlike the t procedures for means, the F test and other procedures for standard deviations are extremely sensitive to non-Normal distributions.²⁵ This lack of robustness does not improve in large samples. It is difficult in practice to tell whether a significant F -value is evidence of unequal population spreads or simply evidence that the populations are not Normal.

The deeper difficulty that underlies the very poor robustness of Normal population procedures for inference about spread already appeared in our work on describing data. The standard deviation is a natural measure of spread for Normal distributions but not for distributions in general. In fact, because skewed distributions have unequally spread tails, no single numerical measure is adequate to describe the spread of a skewed distribution. Thus, the standard deviation is not always a useful parameter, and even when it is, the results of inference about it are not trustworthy. Consequently, we do not recommend use of inference about population standard deviations in basic statistical practice.²⁶

It was once common to test equality of standard deviations as a preliminary to performing the pooled two-sample t test for equality of two population means. It is better practice to check the distributions graphically, with special attention to skewness and outliers, and to use the more general two-sample t that does not require equal standard deviations.

Chapters 15 and 16 discuss procedures called “analysis of variance” for comparing several means. These procedures are extensions of the pooled t test and require that the populations have a common standard deviation.

*This section can be omitted without loss of continuity.

Fortunately, like the t test, analysis of variance comparisons of means are quite robust. (Analysis of variance uses F statistics, but these are not the same as the F statistic for comparing two population standard deviations.) Formal tests for the equality of two standard deviations are not very robust and will often give misleading results. In the words of one distinguished statistician, “To make a preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!”²⁷

The F test for equality of spread

Because of the limited usefulness of procedures for inference about the standard deviations of Normal distributions, we will present only one such procedure. Suppose that we have independent SRSs from two Normal populations, a sample of size n_1 from $N(\mu_1, \sigma_1)$ and a sample of size n_2 from $N(\mu_2, \sigma_2)$. The population means and standard deviations are all unknown. The hypothesis of equal spread,

$$H_0: \sigma_1 = \sigma_2$$

$$H_a: \sigma_1 \neq \sigma_2$$

is tested by a simple statistic, the ratio of the two sample variances.

THE F STATISTIC AND F DISTRIBUTIONS

When s_1^2 and s_2^2 are sample variances from independent SRSs of sizes n_1 and n_2 drawn from Normal populations, the F statistic

$$F = \frac{s_1^2}{s_2^2}$$

has the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom when $H_0: \sigma_1 = \sigma_2$ is true.

F distributions

The F distributions are a family of distributions with two parameters, the degrees of freedom of the sample variances in the numerator and denominator of the F statistic.* The numerator degrees of freedom are always mentioned first. Interchanging the degrees of freedom changes the distribution, so the order is important. Our brief notation will be $F(j, k)$ for the F distribution with j degrees of freedom in the numerator and k in the denominator. The F distributions are not symmetric but are right-skewed. The density curve in Figure 7.12 illustrates the shape. Because sample variances cannot be negative, the F statistic takes only positive values and the F distribution has no probability below 0. The peak of the F density curve is near 1; values far

*The F distributions are another of R. A. Fisher's contributions to statistics and are called F in his honor. Fisher introduced F statistics for comparing several means. We will meet these useful statistics in Chapters 15 and 16.

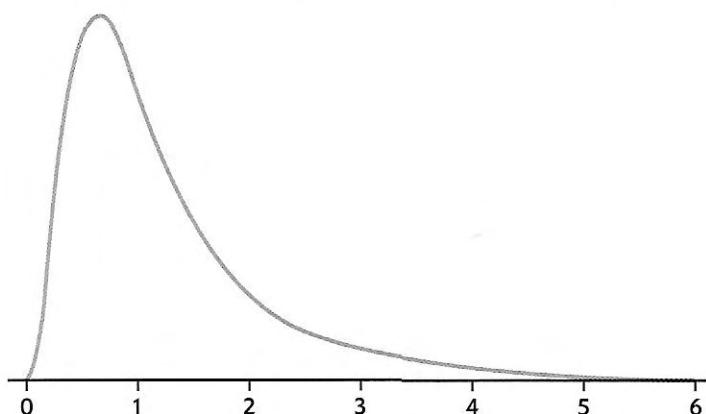


FIGURE 7.12 The density curve for the $F(9,10)$ distribution. The F distributions are skewed to the right.

from 1 in either direction provide evidence against the hypothesis of equal standard deviations.

Tables of F critical values are awkward, because a separate table is needed for every pair of degrees of freedom j and k . Table E in the back of the book gives upper p critical values of the F distributions for $p = 0.10, 0.05, 0.025, 0.01$, and 0.001 . For example, these critical values for the $F(9, 10)$ distribution shown in Figure 7.12 are

p	0.10	0.05	0.025	0.01	0.001
F^*	1.48	1.65	1.82	2.03	2.55

The skewness of the F distributions causes additional complications. In the symmetric Normal and t distributions, the point with probability 0.05 below it is just the negative of the point with probability 0.05 above it. This is not true for F distributions. We therefore require either tables of both the upper and lower tails or some way of getting by without lower-tail critical values. Statistical software that eliminates the need for tables is plainly very convenient. If you do not use statistical software, arrange the F test as follows:

1. Take the test statistic to be

$$F = \frac{\text{larger } s^2}{\text{smaller } s^2}$$

This amounts to naming the populations so that s_1^2 is the larger of the observed sample variances. The resulting F is always 1 or greater.

2. Compare the value of F with critical values from Table E. Then *double* the significance levels from the table to obtain the significance level for the two-sided F test.

The idea is that we calculate the probability in the upper tail and double to obtain the probability of all ratios on either side of 1 that are at least

as improbable as that observed. Remember that the order of the degrees of freedom is important in using Table E.

EXAMPLE 7.18



Comparing healthy and failed firms

Case 7.2 (page 476) recounts a study that compared current assets to current liability ratios for successful and failed firms. There we used the pooled two-sample t procedures. Because these procedures require equal population standard deviations, it is tempting to first test

$$H_0: \sigma_1 = \sigma_2$$

$$H_a: \sigma_1 \neq \sigma_2$$

The 68 healthy firms had a standard deviation of $s = 0.6393$ while the 33 failed firms had a standard deviation of $s = 0.4811$. The F test statistic is therefore

$$F = \frac{\text{larger } s^2}{\text{smaller } s^2} = \frac{0.6393^2}{0.4811^2} = 1.77$$

We compare the calculated value $F = 1.77$ with critical points for the $F(67, 32)$ distribution. One set of entries in Table E that is close to this distribution is the $F(60, 30)$ distribution. Our observed value $F = 1.77$ falls between the critical values $F^* = 1.65$ and $F^* = 1.82$, corresponding to tail areas 0.05 and 0.025. The two-sided P -value therefore lies between 0.10 and 0.05. The standard deviations are significantly different at the 10% level but not at the 5% level.

In most significance tests, we insist on quite strong evidence before we abandon a null hypothesis. That's because we are trying to convince ourselves and others that the effect described by the alternative hypothesis really is present in the population. In this example, however, the null hypothesis $H_0: \sigma_1 = \sigma_2$ is a requirement for using the pooled t to compare two means. Evidence much weaker than the usual significance levels of 5% or 1% should lead you to question H_0 and abandon the pooled t for the more general t that we have recommended.

Statistical output often includes the results of this test with the output for the two-sample t test. Here is the output from SAS:

Equality of Variances						
Variable	Method	Num DF	Den DF	F Value	Pr > F	
ratio	Folded F	67	32	1.77	0.0792	

We see that the exact P -value, if the populations are Normal, is 0.0792. When we examined the histograms and the stemplot (Figures 7.10 and 7.11) for these data, we noted that the data for the failed firms did not look particularly Normal but we were not concerned about that fact because we were interested in comparing the means. For the question of equality of variances, we are not in the same situation. To the extent that our populations are not Normal, we are less confident in our conclusion that the standard deviations are not significantly different.

noncentral t distribution

The power of the two-sample *t* test

The two-sample *t* test is one of the most used statistical procedures. Unfortunately, because of inadequate planning, users frequently fail to find evidence for the effects that they believe to be present. Power calculations should be part of the planning of any statistical study.

In an optional part of Section 7.1, we learned how to approximate the power of the one-sample *t* test. The basic idea is the same for the two-sample case, but we give the exact method rather than an approximation. The exact power calculation involves a new distribution, the **noncentral *t* distribution**. This calculation is not practical by hand but is easy with software that calculates probabilities for this new distribution.

We consider only the common case where the null hypothesis is “no difference,” $\mu_1 - \mu_2 = 0$. We illustrate the calculation for the pooled two-sample *t* test. A simple modification is needed when we do not pool. The unknown parameters in the pooled *t* setting are μ_1 , μ_2 , and a single common standard deviation σ . To find the power for the pooled two-sample *t* test, follow these steps.

Step 1. Specify these quantities:

- an alternative that you consider important to detect; that is, a value for $\mu_1 - \mu_2$;
- the sample sizes, n_1 and n_2 ;
- a fixed significance level, α , often $\alpha = 0.05$;
- an estimate of the standard deviation, σ , from a pilot study or previous studies under similar conditions.

Step 2. Find the degrees of freedom $df = n_1 + n_2 - 2$ and the value of t^* that will lead to rejecting H_0 at your chosen level α .

noncentrality parameter

Step 3. Calculate the noncentrality parameter

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Step 4. The power is the probability that a noncentral *t* random variable with degrees of freedom df and noncentrality parameter δ will be less than t^* . Use software to calculate this probability. In SAS, the command is `1-PROBT(tstar, df, delta)`. If you do not have software that can perform this calculation, you can approximate the power as the probability that a standard Normal random variable is greater than $t^* - \delta$, that is, $P(Z > t^* - \delta)$. Use Table A or software for standard Normal probabilities.

Note that the denominator in the noncentrality parameter,

$$\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is our guess at the standard error for the difference in the sample means. Therefore, if we wanted to assess a possible study in terms of the margin of error for the estimated difference, we would examine t^* times this quantity.

If we do not assume that the standard deviations are equal, we need to guess both standard deviations and then combine these to get an estimate of the standard error:

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

This guess is then used in the denominator of the noncentrality parameter. Use the conservative value, the smaller of $n_1 - 1$ and $n_2 - 1$, for the degrees of freedom.

EXAMPLE 7.19



Healthy versus failed firms

In Case 7.2 we compared the ratio of current assets to current liabilities for 68 successful and 33 failed firms. Using the pooled two-sample procedure, the difference was statistically significant ($t = 7.71$, $df = 99$, $P < 0.001$). Now that this study is several years old, we are planning a similar study to determine if the finding continues to hold.

Should our new sample have similar numbers of firms? Or could we save resources by using smaller samples and still be able to declare that the successful and failed firms are different? To answer this question, we do a power calculation.

Step 1. We want to be able to detect a difference in the means that is about the same as the value that we observed in our previous study. So in our calculations we will use $\mu_1 - \mu_2 = 1.15$. We are willing to assume that the standard deviations will be about the same as in the earlier study, so we take the standard deviation for each of the two groups of firms to be the pooled value from our previous study, $\sigma = 0.5928$.

We need only two pieces of additional information: a significance level α and the sample sizes n_1 and n_2 . For the first, we will choose the standard value $\alpha = 0.05$. For the sample sizes we want to try several different values. Let's start with $n_1 = 15$ and $n_2 = 15$.

Step 2. The degrees of freedom are $n_1 + n_2 - 2 = 28$. The critical value is $t^* = 2.048$, the value from Table D for a two-sided $\alpha = 0.05$ significance test based on 28 degrees of freedom.

Step 3. The noncentrality parameter is

$$\delta = \frac{1.15}{0.5928 \sqrt{\frac{1}{15} + \frac{1}{15}}} = \frac{1.15}{0.2165} = 5.31$$

Step 4. Software gives the power as 0.99922, over 99.9%. The Normal approximation is very accurate:

$$P(Z > t^* - \delta) = P(Z > -3.262) = 0.99945$$

If we repeat the calculation with $n_1 = 10$ and $n_2 = 10$, we still have power greater than 0.98. A difference of means as large as that found in the earlier study can be detected with quite small samples.

A different, and perhaps more important, issue is the margin of error for the estimated difference. For $n_1 = 15$ and $n_2 = 15$, it is

$$t^* \times 0.5928 \sqrt{\frac{1}{15} + \frac{1}{15}} = 2.048 \times 5.31 = 0.44$$

Repeating the calculation for $n_1 = 10$ and $n_2 = 10$ gives 0.56. If we want to reduce the margin of error for the difference in the means of asset/liability ratios, we will need larger sample sizes.

SECTION 7.3 SUMMARY

- Inference procedures for comparing the standard deviations of two Normal populations are based on the *F* statistic, which is the ratio of sample variances:

$$F = \frac{s_1^2}{s_2^2}$$

- When an SRS of size n_1 is drawn from the x_1 population and an independent SRS of size n_2 is drawn from the x_2 population, the *F* statistic has the *F* distribution $F(n_1 - 1, n_2 - 1)$ if the two population standard deviations σ_1 and σ_2 are in fact equal.
- The *F* test for equality of standard deviations tests $H_0: \sigma_1 = \sigma_2$ versus $H_a: \sigma_1 \neq \sigma_2$ using the statistic

$$F = \frac{\text{larger } s^2}{\text{smaller } s^2}$$

and doubles the upper-tail probability to obtain the *P*-value.

- The *t* procedures are quite **robust** when the distributions are not Normal. The *F* tests and other procedures for inference about the spread of one or more Normal distributions are not robust. They are so strongly affected by non-Normality that we do not recommend them for regular use.
- The power of the two-sample *t* test is found by first finding the critical value for the significance test, the degrees of freedom, and the **noncentrality parameter** for the alternative of interest. These are used to calculate the power from a **noncentral t distribution**. A Normal approximation works quite well. Calculating margins of error for various study designs and conditions is an alternative procedure for evaluating designs.

SECTION 7.3 EXERCISES

In all exercises calling for use of the F test, assume that both population distributions are very close to Normal. The actual data are not always sufficiently Normal to justify use of the F test.

- 7.85 The F statistic $F = s_1^2/s_2^2$ is calculated from samples of size $n_1 = 10$ and $n_2 = 21$. (Remember that n_1 is the numerator sample size.)
- What is the upper 5% critical value for this F ?
 - In a test of equality of standard deviations against the two-sided alternative, this statistic has the value $F = 2.45$. Is this value significant at the 10% level? Is it significant at the 5% level?
- 7.86 The F statistic for equality of standard deviations based on samples of sizes $n_1 = 21$ and $n_2 = 26$ takes the value $F = 2.88$.
- Is this significant evidence of unequal population standard deviations at the 5% level?
 - Use Table E to give an upper and a lower bound for the P -value.
- 7.87 **Piano lessons.** In Exercise 7.70 (page 482) we examined the effect of piano lessons on spatial-temporal reasoning. Do the data provide evidence that would cause us to suspect that the standard deviation of the children who took piano lessons is different from that of the controls? Set up the hypotheses, perform the significance test, and summarize the results.
- 7.88 **Wheat prices.** Example 7.14 (page 470) describes a USDA survey used to estimate wheat prices in July and September. Calculate the two sample standard deviations. Perform the test for equality of standard deviations and summarize your conclusion.
- 7.89 **Vitamin C loss in storage.** Exercise 7.65 (page 481) presents data on the loss of vitamin C when bread is stored. Two loaves were measured immediately after baking and another two loaves were measured after three days of storage. These are very small sample sizes.
- Use Table E to find the value that the ratio of variances would have to exceed for us to reject (at the 5% level) the null hypothesis that the standard deviations are equal. What does this suggest about the power of the test?
 - Perform the test and state your conclusion.
- 7.90 **Vitamin E loss in storage.** Exercise 7.67 (page 482) gives data on the loss of vitamin E when bread is stored. Two loaves were measured immediately after baking and another two loaves were measured after three days of storage. These are very small sample sizes.
- Use Table E to find the value that the ratio of variances would have to exceed for us to reject (at the 5% level) the null hypothesis that the standard deviations are equal. What does this suggest about the power of the test?
 - Perform the test and state your conclusion.
- 7.91 **Study habits of men and women.** Return to the SSNA data in Exercise 7.79 (page 485). SSNA scores are generally less variable among women than among men. We want to know whether this is true for this college.
- State H_0 and H_a . Note that H_a is one-sided.
 - Because Table E contains only upper critical values for F , a one-sided test requires that the numerator s^2 in the F statistic belongs to the group that H_a claims to have the larger σ . Calculate this F .

- 9.0
- (c) Compare F to the entries in Table E (no doubling of p) to obtain the P -value. Be sure the degrees of freedom are in the proper order. What do you conclude about the variation in SSHA scores?
- 7.92 **Cocaine use and birth weight: power.** Exercise 7.76 (page 484) summarizes data on cocaine use and birth weight. The study has been criticized because of several design problems. Suppose that you are designing a new study. Based on the results in Exercise 7.76, you think that the true difference in mean birth weights may be about 300 grams (g). A difference this large is clinically important. For planning purposes assume that you will have 100 women in each group and that the common standard deviation is 650 g, a guess that is between the two standard deviations in Exercise 7.76. If you use a pooled two-sample t test with significance level 0.05, what is the power of the test for this design?
- 7.93 **Power, continued.** Repeat the power calculation in the previous exercise, for 25, 50, 75, and 125 women in each group. Summarize your power study, including the results for 100 women per group that you found in the previous exercise. A graph of power against sample size will help.
- 7.94 **Margins of error.** For each of the sample sizes considered in the previous two exercises, estimate the margin of error for the 95% confidence interval for the difference in mean weights. Display these results with a graph or a sketch.
- 7.95 **Ego strength: power.** You want to compare the ego strengths of MBA students who plan to seek work at consulting firms and those who favor manufacturing firms. Based on the data from Exercise 7.73, you will use $\sigma = 0.7$ for planning purposes. The pooled two-sample t test with $\alpha = 0.01$ will be used to make the comparison. You judge a difference of 0.5 points to be of interest.
- Find the power for the design with 20 MBA students in each group.
 - The power in part (a) is not acceptable. Redo the calculations for 30 students in each group and $\alpha = 0.05$.

STATISTICS IN SUMMARY

This chapter presents t tests and confidence intervals for inference about the mean of a single population and for comparing the means of two populations. The one-sample t procedures do inference about one mean, and the two-sample t procedures compare two means. Matched pairs studies use one-sample procedures because you first create a single sample by taking the differences in the responses within each pair. These t procedures are among the most common methods of statistical inference.

The t procedures require that the data be random samples and that the distribution of the population or populations be Normal. One reason for the wide use of t procedures is that they are not very strongly affected by lack of Normality. If you can't regard your data as a random sample, however, the results of inference may be of little value.

The chapter exercises are important in this and later chapters. You must now recognize problem settings and decide which of the methods presented

in the chapter fits. In this chapter, you must recognize one-sample studies, matched pairs studies, and two-sample studies. Here are the most important skills you should have after reading this chapter.

A. RECOGNITION

1. Recognize when a problem requires inference about a mean or comparing two means.
2. Recognize from the design of a study whether one-sample, matched pairs, or two-sample procedures are needed.

B. ONE-SAMPLE t PROCEDURES

1. Recognize when the t procedures are appropriate in practice, in particular that they are quite robust against lack of Normality but are influenced by outliers.
2. Also recognize when the design of the study, outliers, or a small sample from a skewed distribution makes the t procedures risky.
3. Use t to obtain a confidence interval at a stated level of confidence for the mean μ of a population.
4. Carry out a t test for the hypothesis that a population mean μ has a specified value against either a one-sided or a two-sided alternative. Use Table D of t distribution critical values to approximate the P -value or carry out a fixed α test.
5. Recognize matched pairs data and use the t procedures to obtain confidence intervals and to perform tests of significance for such data.

C. TWO-SAMPLE t PROCEDURES

1. Recognize when the two-sample t procedures are appropriate in practice.
2. Give a confidence interval for the difference between two means. Use the two-sample t statistic with conservative degrees of freedom if you do not have statistical software. Use software if you have it.
3. Test the hypothesis that two populations have equal means against either a one-sided or a two-sided alternative. Use the two-sample t test with conservative degrees of freedom if you do not have statistical software. Use software if you have it.
4. Know that procedures for comparing the standard deviations of two Normal populations are available, but that these procedures are risky because they are not at all robust against non-Normal distributions.

CHAPTER 7 REVIEW EXERCISES

- 7.96 *Insulation failures.* A manufacturer of electric motors tests insulation at a high temperature (250°C) and records the number of hours until the insulation fails. The data for 5 specimens are²⁸

300 324 372 372 444

The small sample size makes judgment from the data difficult, but engineering experience suggests that the logarithm of the failure time will have a Normal distribution. Take the logarithms of the 5 observations, and use *t* procedures to give a 90% confidence interval for the mean of the log failure time for insulation of this type.

- 7.97 **Comparing earnings of bank employees.** Banks employ many workers paid by the hour as tellers and data clerks and in other capacities. Table 1.8 (page 31) presents the annual earnings for a random sample of hourly workers at National Bank. The population of all such workers is in the data set *hourly.dat*, described in detail in the Data Appendix. Suppose that we are interested only in the question of whether or not there is an apparent difference in the salaries of men and women. We will therefore combine the sample data for the two races. Use the data in Table 1.8 for an analysis that compares the earnings of men and women. Include a graphical summary, the results of a significance test, and a confidence interval. Summarize your conclusions. Does the finding of a statistically significant difference mean that the bank discriminates?
- 7.98 **Competitive prices?** A retailer entered into an exclusive agreement with a supplier who guaranteed to provide all products at competitive prices. The retailer eventually began to purchase supplies from other vendors who offered better prices. The original supplier filed a legal action claiming violation of the agreement. In defense, the retailer had an audit performed on a random sample of invoices. For each audited invoice, all purchases made from other suppliers were examined and the prices were compared with those offered by the original supplier. For each invoice, the percent of purchases for which the alternate supplier offered a lower price than the original supplier was recorded. Here are the data.²⁹

100	0	0	100	33	45	100	34	78
100	77	33	100	69	100	89	100	100
100	100	100	100	100	100	100	100	

Report the average of the percents with a 95% margin of error. Do the sample invoices suggest that the original supplier's prices are not competitive on the average?

- 7.99 **Behavior of pet owners.** On the morning of March 5, 1996, a train with 14 tankers of propane derailed near the center of the small Wisconsin town of Weyauwega. Six of the tankers were ruptured and burning when the 1700 residents were ordered to evacuate the town. Researchers study disasters like this so that effective relief efforts can be designed for future disasters. About half of the households with pets did not evacuate all of their pets. A study conducted after the derailment focused on problems associated with retrieval of the pets after the evacuation and characteristics of the pet owners. One of the scales measured "commitment to adult animals." The people who evacuated some or all of their pets were compared with those who did not evacuate any of their pets. Higher scores indicate that the pet owner is more likely to take actions that benefit the pet. Here are the data summaries.³⁰

Group	<i>n</i>	\bar{x}	<i>s</i>
Evacuated all or some pets	116	7.95	3.62
Did not evacuate any pets	125	6.26	3.56

Analyze the data and prepare a short report describing the results.

- 7.100 **Weight-loss programs.** In a study of the effectiveness of weight-loss programs, 47 subjects who were at least 20% overweight took part in a group support program for 10 weeks. Private weighings determined each subject's weight at the beginning of the program and 6 months after the program's end. A *t* test was used to assess the significance of the average weight loss. The paper reporting the study said, "The subjects lost a significant amount of weight over time, $t(46) = 4.68, p < 0.01$."³¹ It is common to report the results of statistical tests in this abbreviated style.
- Which *t* test did the study use?
 - Explain to someone who knows no statistics but is interested in weight-loss programs what the practical conclusion is.
 - The paper follows the tradition of reporting significance only at fixed levels such as $\alpha = 0.01$. In fact, the results are more significant than " $p < 0.01$ " suggests. What can you say about the *P*-value of the test?

- 7.101 **Preservatives in meat products.** Nitrites are often added to meat products as preservatives. In a study of the effect of these chemicals on bacteria, the rate of uptake of a radiolabeled amino acid was measured for a number of cultures of bacteria, some growing in a medium to which nitrites had been added. Here are the summary statistics from this study.

Group	<i>n</i>	\bar{x}	<i>s</i>
Nitrite	30	7880	1115
Control	30	8112	1250

Carry out a test of the research hypothesis that nitrites decrease amino acid uptake, and report your results.

- 7.102 **Testing job applicants.** The one-hole test is used to test the manipulative skill of job applicants. This test requires subjects to grasp a pin, move it to a hole, insert it, and return for another pin. The score on the test is the number of pins inserted in a fixed time interval. One study compared male college students with experienced female industrial workers. Here are the data for the first minute of the test.³²

Group	<i>n</i>	\bar{x}	<i>s</i>
Students	750	35.12	4.31
Workers	412	37.32	3.83

- We expect that the experienced workers will outperform the students, at least during the first minute, before learning occurs. State the hypotheses

for a statistical test of this expectation and perform the test. Give a P -value and state your conclusions.

- (b) The distribution of scores is slightly skewed to the left. Explain why the procedure you used in (a) is nonetheless acceptable.
- (c) One purpose of the study was to develop performance norms for job applicants. Based on the data above, what is the range that covers the middle 95% of experienced workers? (Be careful! This is not the same as a 95% confidence interval for the mean score of experienced workers.)
- (d) The five-number summary of the distribution of scores among the workers is

23 33.5 37 40.5 46

for the first minute and

32 39 44 49 59

for the fifteenth minute of the test. Display these summaries graphically, and describe briefly the differences between the distributions of scores in the first and fifteenth minute.

- 7.103 Occupation and diet.** Do various occupational groups differ in their diets? A British study of this question compared 98 drivers and 83 conductors of London double-decker buses. The conductors' jobs require more physical activity. The article reporting the study gives the data as "mean daily consumption (\pm se)." ³³ Some of the study results appear below:

	Drivers	Conductors
Total calories	2821 ± 44	2844 ± 48
Alcohol (grams)	0.24 ± 0.06	0.39 ± 0.11

- (a) What does "se" stand for? Give \bar{x} and s for each of the four sets of measurements.
- (b) Is there significant evidence at the 5% level that conductors consume more calories per day than do drivers? Use a t test to give a P -value, and then assess significance.
- (c) How significant is the observed difference in mean alcohol consumption? Use a t test to obtain the P -value.

- 7.104 Occupation and diet, continued.** Use the data in the previous exercises to give two confidence intervals:

- (a) A 90% confidence interval for the mean daily alcohol consumption of London double-decker bus conductors.
- (b) An 80% confidence interval for the difference in mean daily alcohol consumption between drivers and conductors.

- 7.105 The pooled test.** Use of the pooled two-sample t test is justified in part (b) of Exercise 7.103. Explain why. Find the P -value for the pooled t statistic and compare with your result in Exercise 7.103.

- 7.106 Conditions for inference.** The report cited in Exercise 7.103 says that the distribution of alcohol consumption among the individuals studied is "grossly skew."

- (a) Do you think that this skewness prevents the use of the two-sample t test for equality of means? Explain your answer.
- (b) Do you think that the skewness of the distributions prevents the use of the F test for equality of standard deviations? Explain your answer.
- 7.107 **Conditions for inference.** Table 1.13 (page 82) gives the populations of all 58 counties in the state of California. Is it proper to apply the one-sample t method to these data to give a 95% confidence interval for the mean population of a California county? Explain your answer.
- 7.108 **Male and female CS students.** Is there a difference between the average SAT scores of male and female computer science students? The CSDATA data set, described in the Data Appendix, gives the math (SATM) and verbal (SATV) scores for a group of 224 computer science majors. The variable SEX indicates whether each individual is male or female.
- (a) Compare the two distributions graphically, and then use software to compare the average SATM scores of males and females. Is it appropriate to use the pooled t test for this comparison? Write a brief summary of your results and conclusions. Refer to both versions of the t test and also to the F test for equality of standard deviations. Include a 99% confidence interval for the difference in the means.
- (b) The students in the CSDATA data set are all computer science majors who entered a major university during a particular year. To what extent do you think that your results would generalize to (i) computer science students entering in different years, (ii) computer science majors at other colleges and universities, and (iii) college students in general?
- 7.109 **Ego strengths of MBA graduates: power.** In Exercise 7.95 (page 495) you found the power for a study designed to compare the “ego strengths” of two groups of MBA students. Now you must design a study to compare MBA graduates who reached partner in a large consulting firm with those who joined the firm but failed to become partners.
- Assume the same value of $\sigma = 0.07$ and use $\alpha = 0.05$. You are planning to have 20 subjects in each group. Calculate the power of the pooled two-sample t test that compares the mean ego strengths of these two groups of MBA graduates for several values of the true difference. Include values that have a very small chance of being detected and some that are virtually certain to be seen in your sample. Plot the power versus the true difference and write a short summary of what you have found.
- 7.110 **t approaches z .** As the degrees of freedom increase, the t distributions get closer and closer to the $N(0, 1)$, or z , distribution. One way to see this is to look at how the critical value t^* for a 95% confidence interval changes with the degrees of freedom. Make a plot with degrees of freedom from 2 to 100 on the x axis and t^* on the y axis. Draw a horizontal line on the plot corresponding to the value of $z^* = 1.96$. Summarize the main features of the plot.
- 7.111 **Sample size and margin of error.** The margin of error for the one-sample t confidence interval depends on the confidence level, the standard deviation, and the sample size. Fix the confidence level at 95% and the standard error at $s = 1$ to examine the effect of the sample size n . Find the margin of error for sample sizes of 5 to 100 by 5's. That is, let $n = 5, 10, 15, \dots, 100$. Plot the margins of error versus the sample size and summarize the relationship.



CHAPTER 7 CASE STUDY EXERCISES

9.21

CASE STUDY 7.1: Architectural firms. Table 1.3 (page 14) and data set *cse_01.dat* give characteristics of 24 large Indianapolis area architectural firms. Make a table giving the mean, the standard deviation, the 95% confidence interval, and the five-number summary for the variables 1998 billings, 1997 billings, the number of licensed architects employed, the number of licensed engineers employed, and the number of full-time equivalent staff.

Classify each firm as “old” or “new” based on whether or not they began doing business in the area before 1970. Compare the means of the old and new firms for the variables 1998 billings, 1997 billings, the number of licensed architects employed, the number of licensed engineers employed, and the number of full-time equivalent staff. Be sure to state whether or not you use the pooled procedures and why. Give your results with graphical and numerical summaries. Then write a short paragraph explaining any differences that you have found.

CASE STUDY 7.2: Three or four bedrooms. How much more would you expect to pay for a home that has four bedrooms than for a home that has three? Here are data for West Lafayette, Indiana.³⁴ These are the selling prices (in dollars) that the owners of the homes are asking.

Four-bedroom homes:

121,900	139,900	157,000	159,900	176,900
224,900	235,000	245,000	294,000	

Three-bedroom homes:

65,500	79,900	79,900	79,900	82,900	87,900	94,000
97,500	105,000	111,900	116,900	117,900	119,900	122,900
124,000	125,000	126,900	127,900	127,900	127,900	132,900
145,000	145,500	157,500	194,000	205,900	259,900	265,000

Plot the selling prices for the two sets of homes and describe the two distributions. Test the null hypothesis that the mean asking prices for the two sets of homes are equal versus the two-sided alternative. Give the test statistic with degrees of freedom, the *P*-value, and your conclusion. Would you consider using a one-sided alternative for this analysis? Explain why or why not. Give a 95% confidence interval for the difference in mean selling prices. These data are not SRSs from a population. Give a justification for use of the two-sample *t* procedures in this case.

Go to the Web site www.realtor.com and select two geographical areas of interest to you. You will compare the prices of similar types of homes in these two areas. State clearly how you define the areas and the type of homes. For example, you can use city names or zip codes to define the area and you can select single-family homes or condominiums. We view these homes as representative of the asking prices of homes for these areas at the time of your search. If the search gives a large number of homes, select a random sample. Be sure to explain exactly how you do this. Use the methods you have learned in this chapter to compare the selling prices. Be sure to include a graphical summary.

Prelude

Jury selection

In a small border county in Texas, a so-called key man system was used to select citizens to serve on juries. Under this procedure, lists of persons in the county were compiled by jury commissioners and these were interviewed under oath by the county judge, who determined whether or not they were qualified. Qualifications included literacy and "good moral character." This system was challenged in a case called *Castaneda v. Partida* on the basis that it reduced the proportion of Hispanic citizens serving on juries. A statistical comparison was at the heart of the case: 79% of the county's adult population had Spanish surnames, but only 39% of the grand jurors did. This comparison gave a z statistic of 29 with an extremely small P -value. The United States Supreme Court concluded that this disparity established a *prima facie* case.¹ Procedures such as the key man system are no longer used to select juries in the United States. The Court interpreted the z statistic as the number of standard deviations by which the two percents differed: 79% differed from 39% by 29 standard deviations. The decision stated that a difference *greater than 2 or 3 standard deviations* is sufficient. This criterion is sometimes called the *Castaneda rule*. These values of z correspond to P -values of 0.0456 and 0.0026.

Excel Output

CHAPTER

8

9.24

Inference for Proportions

- 8.1 Inference for a Single Proportion**
- 8.2 Comparing Two Proportions**

Introduction

Some statistical studies concern variables measured in a scale of equal units, such as dollars or grams. Chapter 7 discussed inference about the mean of variables like these. Other studies record *categorical variables*, such as the race or occupation of a person, the make of a car, or the type of complaint received from a customer. When we record categorical variables, our data consist of *counts* or of *percents* obtained from counts.

Many statistical studies produce counts rather than measurements. An opinion poll asks a sample of adults whether they approve of the president's conduct of his office; the data are the counts of "Yes," "No," and "No opinion." Insurance customers either renew their policy or drop coverage. A researcher classifies each of a sample of students according to gender and whether or not they are frequent binge drinkers; the data are the counts of students in each of four categories. The parameters we want to do inference about in these settings are *population proportions*. Just as in the case of inference about population means, we may be concerned with a single population or with comparing two populations. Inference about one or two proportions is very similar to inference about means, which we discussed in Chapter 7. In particular, inference for both means and proportions is based on sampling distributions that are approximately Normal.

We begin in Section 8.1 with inference about a single population proportion. Section 8.2 concerns methods for comparing two proportions.

8.1

Inference for a Single Proportion



CASE 8.1

WORK STRESS AND PERSONAL LIFE

"The nature of work is changing at whirlwind speed. Perhaps now more than ever before, job stress poses a threat to the health of workers and, in turn, to the health of organizations."² So says the National Institute for Occupational Safety and Health. Employers are concerned about the effect of stress on their employees. Stress can lower morale and efficiency and increase medical costs. A large survey of restaurant employees found that 75% reported that work stress had a negative impact on their personal lives.³

The human resources manager of a chain of restaurants is concerned that work stress may be affecting the chain's employees. She asks a random sample of 100 employees to respond Yes or No to the question "Does work stress have a negative impact on your personal life?" Of these, 68 say "Yes." The *parameter* of interest is the proportion of the chain's employees who would answer "Yes" if asked. This is a *population proportion*, which we

population proportion

Excel Output

call p . The *statistic* used to estimate this unknown parameter is the **sample proportion**

$$\hat{p} = \frac{68}{100} = 0.68$$

The *count* of “Yes” answers in the sample of Case 8.1 is $X = 68$. We will regularly use X to stand for sample counts and $\hat{p} = X/n$ to denote sample proportions. The sample proportion \hat{p} in Case 8.1 is a discrete random variable that can take the values 0, 1/100, 2/100, ..., 99/100, or 1.* If the sample size n is very small, we must base tests and confidence intervals for p on the discrete distribution of \hat{p} . Discrete distributions are a bit awkward to work with.⁴ Fortunately, we can approximate the distribution of \hat{p} by a Normal distribution when the sample size n is large.

SAMPLING DISTRIBUTION OF A SAMPLE PROPORTION

Choose an SRS of size n from a large population that contains population proportion p of “successes.” Let \hat{p} be the **sample proportion** of successes,

$$\hat{p} = \frac{\text{count of successes in the sample}}{n} = \frac{X}{n}$$

Then:

- As the sample size increases, the sampling distribution of \hat{p} becomes **approximately Normal**.
- The **mean** of the sampling distribution is p .
- The **standard deviation** of the sampling distribution is

$$\sqrt{\frac{p(1-p)}{n}}$$

Figure 8.1 summarizes these facts in a form that recalls the idea of sampling distributions. We will consider only inference procedures based on this Normal approximation. These procedures are similar to those for inference about the mean of a Normal distribution. We will see, however, that there are a few extra details involved, caused by the added difficulty in approximating the discrete distribution of \hat{p} by a continuous Normal distribution.

*In many cases, a probability model for \hat{p} can be based on the Binomial distributions for counts, discussed in the optional Chapter 5.

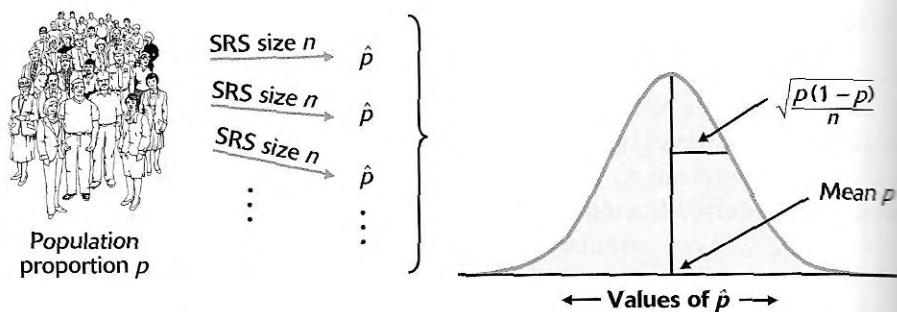


FIGURE 8.1 Draw a large SRS from a population in which the proportion p are successes. The sampling distribution of the sample proportion \hat{p} of successes has approximately a Normal distribution.

Confidence interval for a single proportion

The sample proportion $\hat{p} = X/n$ is the natural estimator of the population proportion p . The traditional confidence interval for p is based on the Normal approximation to the distribution of \hat{p} . Unfortunately, modern computer studies reveal that confidence intervals based on this statistic can be quite inaccurate, even for large samples. We can get a hint of the difficulties by asking what we would do if the sample survey in Case 8.1 found that *none* of the 100 employees in the sample said job stress hurt their personal lives. Then because $\hat{p} = 0$, the standard error based on this estimate is also 0. That is, the traditional method says that we are certain that no employees in the entire restaurant chain are bothered by work stress. This isn't plausible.

Both computer studies and careful mathematics show that we can do better by moving the sample proportion \hat{p} slightly away from 0 and 1. There are several ways to do this. We will use a simple adjustment that works very well in practice. The adjustment is based on the following idea: act as if we have 4 additional observations, 2 of which are successes and 2 of which are failures. The new sample size is $n + 4$ and the count of successes is $X + 2$. The estimator of the population proportion based on this *add 2 successes and 2 failures* rule is

$$\tilde{p} = \frac{X + 2}{n + 4}$$

Wilson estimate

Because this estimate was first suggested by Edwin Bidwell Wilson in 1927 (though rarely used in practice until recently), we call it the **Wilson estimate**. We base a confidence interval on the z statistic obtained by standardizing the Wilson estimate \tilde{p} . Because \tilde{p} is the sample proportion for our modified sample of size $n + 4$, it isn't surprising that the distribution of \tilde{p} is close to the Normal distribution with mean p and standard deviation $\sqrt{p(1-p)/(n+4)}$. To get a confidence interval, we estimate p by \tilde{p} in this standard deviation to get the standard error of \tilde{p} . Here is the final result:

feel output

CONFIDENCE INTERVAL FOR A POPULATION PROPORTION

Choose an SRS of size n from a large population with unknown proportion p of successes. The **Wilson estimate** of the population proportion is

$$\tilde{p} = \frac{X + 2}{n + 4}$$

The standard error of \tilde{p} is

$$SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}}$$

An approximate level C confidence interval for p is

$$\tilde{p} \pm z^* SE_{\tilde{p}}$$

where z^* is the value for the standard Normal density curve with area C between $-z^*$ and z^* . The margin of error is

$$m = z^* SE_{\tilde{p}}$$

Use this interval when the sample size is at least $n = 5$ and the confidence level is 90%, 95%, or 99%.

EXAMPLE 8.1**Estimating the effect of work stress**

The sample survey in Case 8.1 found that 68 of a sample of 100 employees agreed that work stress had a negative impact on their personal lives. That is, the sample size is $n = 100$ and the count of successes is $X = 68$. The Wilson estimate of the proportion of all employees affected by work stress is

$$\begin{aligned}\tilde{p} &= \frac{X + 2}{n + 4} \\ &= \frac{68 + 2}{100 + 4} = \frac{70}{104} = 0.6731\end{aligned}$$

The standard error is

$$\begin{aligned}SE_{\tilde{p}} &= \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}} \\ &= \sqrt{\frac{0.6731(1 - 0.6731)}{104}} = 0.0460\end{aligned}$$

The z critical value for 95% confidence is $z^* = 1.96$, so the confidence interval is

$$\begin{aligned}\tilde{p} \pm z^* SE_{\tilde{p}} &= 0.6731 \pm (1.96)(0.0460) \\ &= 0.673 \pm 0.090\end{aligned}$$

We are 95% confident that between 58.3% and 76.3% of the restaurant chain's employees feel that work stress is damaging their personal lives.

Remember that the margin of error in any confidence interval includes only random sampling error. If employees hesitate to respond honestly,

Minitab

Sample	X	N	Sample p	95.0 % CI
1	70	104	0.673077	(0.582923, 0.763231)

SAS**Binomial Proportion for y = 0**

Proportion	0.6731
ASE	0.0460
95% Lower Conf Limit	0.5829
95% Upper Conf Limit	0.7632

Exact Conf Limits

95% Lower Conf Limit	0.5741
95% Upper Conf Limit	0.7619

Sample Size = 104

FIGURE 8.2 Minitab and SAS output for the confidence interval in Example 8.1.

for example, your estimate is likely to miss by more than the margin of error.

Because the calculations for statistical inference for a single proportion are relatively straightforward, many software packages do not include them. Moreover, newer methods such as the interval based on the Wilson estimate have only slowly begun to appear in software. Figure 8.2 gives output from Minitab and SAS for the data in Case 8.1. We added 2 successes and 2 failures to the actual data; this forces the software to use the Wilson estimate. That's why both outputs give the sample size as 104. As usual, the output reports more digits than are useful. When you use software, be sure to think about how many digits are meaningful for your purposes. SAS gives the standard error next to the label ASE, which stands for asymptotic standard error. The SAS output also includes an alternative interval based on an "exact" method. Recent research suggests that the Wilson estimate method performs better than this alternative.

APPLY YOUR KNOWLEDGE

- 8.1 **Will the upgrade be profitable?** To profitably produce a planned upgrade of a software product you make, you must charge customers \$100. Are your customers willing to pay this much? You contact a random sample of 40 customers and find that 11 would pay \$100 for the upgrade. Find a 95% confidence interval for the proportion of all of your customers (the population) who would be willing to buy the upgrade for \$100.
- 8.2 **Holiday shopping.** A poll of 811 adults aged 18 or older asked about purchases that they intended to make for the upcoming holiday season. One of the questions asked about what kind of gift they intended to buy for the person on whom they would spend the most. Clothing was the first choice of 487 people. Give a 99% confidence interval for the proportion of people in this population who intend to buy clothing as their first choice.
- 8.3 **New product sales.** Yesterday, your top salesperson called on 5 customers and obtained orders for your new product from all 5. Suppose

that it is reasonable to view these 5 customers as a random sample of all of her customers.

- Give the Wilson estimate of the proportion of her customers who would buy the new product. Notice that we don't estimate that all customers will buy, even though all 5 in the sample did.
- Give the margin of error for 95% confidence. (You may see that the upper endpoint of the confidence interval is greater than 1. In that case, take the upper endpoint to be 1.)
- Do the results apply to all of your sales force? Explain why or why not.

Significance test for a single proportion

We know that the sample proportion $\hat{p} = X/n$ is approximately Normal, with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$. For confidence intervals, we used the Wilson estimate and estimated the standard deviation from the data. When performing a significance test, however, the null hypothesis specifies a value for p , which we will call p_0 . When we calculate P-values, we act as if the hypothesized p were actually true. When we test $H_0: p = p_0$, we substitute p_0 for p in the expression for $\sigma_{\hat{p}}$ and then standardize \hat{p} . Here are the details.

LARGE-SAMPLE TEST FOR A POPULATION PROPORTION

Choose an SRS of size n from a large population with unknown proportion p of successes. To test the hypothesis $H_0: p = p_0$, compute the z statistic

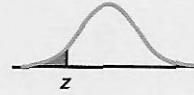
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

In terms of a standard Normal random variable Z , the approximate P-value for a test of H_0 against

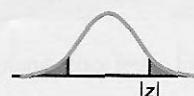
$$H_a: p > p_0 \text{ is } P(Z \geq z)$$



$$H_a: p < p_0 \text{ is } P(Z \leq z)$$



$$H_a: p \neq p_0 \text{ is } 2P(Z \geq |z|)$$



Use this test when the expected number of successes np_0 and the expected number of failures $n(1 - p_0)$ are both greater than 10.

We call this test a “large-sample test” because it is based on a Normal approximation to the sampling distribution of \hat{p} that becomes more accurate as the sample size increases. For small samples, or if the population is less than 10 times as large as the sample, consult an expert for other procedures.

EXAMPLE 8.2
Work stress

A national survey of restaurant employees found that 75% said that work stress had a negative impact on their personal lives. A sample of 100 employees of a restaurant chain finds that 68 answer “Yes” when asked, “Does work stress have a negative impact on your personal life?” Is this good reason to think that the proportion of all employees of this chain who would say “Yes” differs from the national proportion $p_0 = 0.75$?

To answer this question, we test

$$\begin{aligned} H_0: p &= 0.75 \\ H_a: p &\neq 0.75 \end{aligned}$$

The expected numbers of “Yes” and “No” responses are $100 \times 0.75 = 75$ and $100 \times 0.25 = 25$. Both are greater than 10, so we can use the z test. The test statistic is

$$\begin{aligned} z &= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \\ &= \frac{0.68 - 0.75}{\sqrt{\frac{(0.75)(0.25)}{100}}} = 1.62 \end{aligned}$$

From Table A we find $P(Z \geq 1.62) = 1 - 0.9474$, or 0.0526. The P -value is the area in both tails, $P = 2 \times 0.0526 = 0.1052$. Figure 8.3 displays the P -value as

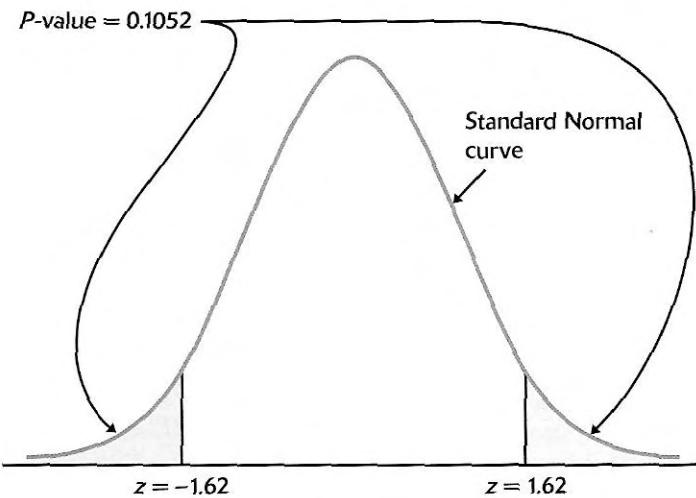


FIGURE 8.3 The P -value for Example 8.2.

an area under the standard Normal curve. We conclude that the chain restaurant data are compatible with the survey results ($\hat{p} = 0.68$, $z = 1.62$, $P = 0.11$).

Minitab and SAS output for the analysis in Example 8.2 appear in Figure 8.4. (Example 8.2 finds a slightly different P -value because we rounded the z statistic to two decimal places to use Table A.) Note that for the significance test, we did not alter the data set by adding additional cases as we did to generate the confidence interval output in Figure 8.2.

In Example 8.2, we counted “Yes” responses. Would we get the same results if we instead counted the people who said “No?” We would. If we count No’s as successes, both the sample and the population proportions are proportions of No’s. The null hypothesis is then the claim that the population proportion is 0.25 rather than the 0.75 used for Yes’s in Example 8.2. You will find that the z test statistic is unchanged except for its sign and that the P -value remains the same.

As usual, the confidence interval in Example 8.1 is more informative than the significance test in Example 8.2. The confidence interval gives the range of population proportions p that are consistent with the sample data. Because the 95% confidence interval is (0.583, 0.763), we would not be surprised if the truth were 60% or 75%. The test simply says that we don’t have good evidence against the single value 75% that served as p_0 . Significance tests for a single population proportion are not common in practice, because we rarely want to compare data with a precise value p_0 . The confidence interval, however, is used quite often.

Minitab

Test of $p = 0.75$ vs $p \neq 0.75$

Success = 1

Variable	X	N	Sample p	Z-Value	P-Value
C2	68	100	0.680000	-1.62	0.106

SAS

Test of H_0 : Proportion = 0.75

ASE under H_0	0.0433
Z	-1.6166
One-sided Pr < Z	0.0530
Two-sided Pr > Z	0.1060

Sample Size = 100

FIGURE 8.4 Minitab and SAS output for the significance test in Example 8.2.

APPLY YOUR KNOWLEDGE

- 8.4 A profitable upgrade? In Exercise 8.1 we found that 11 customers from a random sample of 40 would be willing to buy a software upgrade that costs \$100. If the upgrade is to be profitable, you will need to sell it to

more than 20% of your customers. Do the sample data give good evidence that more than 20% are willing to buy?

- Formulate this problem as a hypothesis test. Give the null and alternative hypotheses. Will you use a one-sided or a two-sided alternative? Why?
- Carry out the significance test. Report the test statistic and the P -value.
- Should you proceed with plans to produce and market the upgrade?

8.5

Yes or no? Case 8.1 describes a survey of 100 employees, 68 of whom answered “Yes” to a question on work stress. That is, 32 of the sample of 100 answered “No.”

- Example 8.1 gives a 95% confidence interval for the proportion of all employees who would answer “Yes.” Give the 95% confidence interval for the proportion who would answer “No.” Explain carefully why your result agrees with Example 8.1.
- Example 8.2 tests the null hypothesis that the proportion of “Yes” among all employees is 0.75. Carry out a two-sided test of the null hypothesis that the proportion of “No” is 0.25. Explain carefully why your result agrees with Example 8.2.



Choosing a sample size

In Chapter 6, we showed how to choose the sample size n to obtain a confidence interval with specified margin of error m for a Normal mean. Because we are using a Normal approximation for inference about a population proportion, sample size selection proceeds in much the same way.

Recall that the margin of error for the confidence interval for a population proportion is

$$m = z^* \text{SE}_{\tilde{p}} = z^* \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}}$$

Choosing a confidence level C fixes the critical value z^* . The margin of error also depends on the value of \tilde{p} and the sample size n . Because we don't know the value of \tilde{p} until we gather the data, we must guess a value to use in the calculations. We will call the guessed value p^* . Here are two ways to get p^* :

- Use the sample estimate from a pilot study or from similar studies done earlier.
- Use $p^* = 0.5$. Because the margin of error is largest when $\tilde{p} = 0.5$, this choice gives a sample size that is somewhat larger than we really need for the confidence level we choose. It is a safe choice no matter what the data later show.

Once we have chosen p^* and the margin of error m that we want, we can find the n we need to achieve this margin of error. Here is the result.

SAMPLE SIZE FOR DESIRED MARGIN OF ERROR

The level C confidence interval for a proportion p will have a margin of error approximately equal to a specified value m when the sample size satisfies

$$n + 4 = \left(\frac{z^*}{m} \right)^2 p^*(1 - p^*)$$

Here z^* is the critical value for confidence C , and p^* is a guessed value for the proportion of successes in the future sample.

The margin of error will be less than or equal to m if p^* is chosen to be 0.5. The sample size required is then given by

$$n + 4 = \left(\frac{z^*}{2m} \right)^2$$

The value of n obtained by this method is not particularly sensitive to the choice of p^* as long as p^* is not too far from 0.5. However, if your actual sample turns out to have \hat{p} smaller than about 0.3 or larger than about 0.7, the sample size based on $p^* = 0.5$ may be much larger than needed.

EXAMPLE 8.3**Planning a sample of customers**

Your company has received complaints about its customer support service. You intend to hire a consulting company to carry out a sample survey of customers. Before contacting the consultant, you want some idea of the sample size you will have to pay for. One critical question is the degree of satisfaction with your customer service, measured on a five-point scale. You want to estimate the proportion p of your customers who are satisfied (that is, who choose either “satisfied” or “very satisfied,” the two highest levels on the five-point scale).

You want to estimate p with 95% confidence and a margin of error less than or equal to 3%, or 0.03. For planning purposes, you are willing to use $p^* = 0.5$. To find the sample size required,

$$n + 4 = \left(\frac{z^*}{2m} \right)^2 = \left[\frac{1.96}{(2)(0.03)} \right]^2 = 1067.1$$

Round up to get $n + 4 = 1068$, or $n = 1064$. (Always round up. Rounding down would give a margin of error slightly greater than 0.03.)

Similarly, for a 2.5% margin of error we have (after rounding up)

$$n + 4 = \left[\frac{1.96}{(2)(0.025)} \right]^2 = 1537$$

and for a 2% margin of error,

$$n + 4 = \left[\frac{1.96}{(2)(0.02)} \right]^2 = 2401$$

News reports frequently describe the results of surveys with sample sizes between 1000 and 1500 and a margin of error of about 3%. These surveys generally use sampling procedures more complicated than simple random sampling, so the calculation of confidence intervals is more involved than what we have studied in this section. The calculations in Example 8.3 nonetheless show in principle how such surveys are planned.

In practice, many factors influence the choice of a sample size. The following case study illustrates one set of factors.



MARKETING CHRISTMAS TREES

An association of Christmas tree growers in Indiana sponsored a sample survey of Indiana households to help improve the marketing of Christmas trees.⁷ The researchers decided to use a telephone survey and estimated that each telephone interview would take about 2 minutes. Nine trained students in agribusiness marketing were to make the phone calls between 1:00 p.m. and 8:00 p.m. on a Sunday. After discussing problems related to people not being at home or being unwilling to answer the questions, the survey team proposed a sample size of 500. Several of the questions asked demographic information about the household. The key questions of interest had responses of “Yes” or “No,” for example, “Did you have a Christmas tree last year?” The primary purpose of the survey was to estimate various sample proportions for Indiana households. An important issue in designing the survey was therefore whether the proposed sample size of $n = 500$ would be adequate to provide the sponsors of the survey with the information they required.

To address this question, we calculate the margins of error of 95% confidence intervals for various values of \tilde{p} .

EXAMPLE 8.4



Margins of error

In the Christmas tree market survey, the margin of error of a 95% confidence interval for any value of \tilde{p} and $n = 500$ is

$$\begin{aligned} m &= z^* \text{SE}_{\tilde{p}} \\ &= 1.96 \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{504}} \end{aligned}$$

The results for various values of \tilde{p} are

\tilde{p}	m	\tilde{p}	m
0.05	0.019	0.60	0.043
0.10	0.026	0.70	0.040
0.20	0.035	0.80	0.035
0.30	0.040	0.90	0.026
0.40	0.043	0.95	0.019
0.50	0.044		

The survey team judged these margins of error to be acceptable, and they used a sample size of 500 in their survey.

The table in Example 8.4 illustrates two points. First, the margins of error for $\tilde{p} = 0.05$ and $\tilde{p} = 0.95$ are the same. The margins of error will always be the same for \tilde{p} and $1 - \tilde{p}$. This is a direct consequence of the form of the confidence interval. Second, the margin of error varies only between 0.040 and 0.044 as \tilde{p} varies from 0.3 to 0.7, and the margin of error is greatest when $\tilde{p} = 0.5$, as we claimed earlier. It is true in general that the margin of error will vary relatively little for values of \tilde{p} between 0.3 and 0.7. Therefore, when planning a study, it is not necessary to have a very precise guess for p . If $p^* = 0.5$ is used and the observed \tilde{p} is between 0.3 and 0.7, the actual interval will be a little shorter than needed, but the difference will be quite small.

APPLY YOUR KNOWLEDGE

- 8.6 **Is there interest in a new product?** One of your employees has suggested that your company develop a new product. You decide to take a random sample of your customers and ask whether or not there is interest in the new product. The response is on a 1 to 5 scale, with 1 indicating “definitely would not purchase”; 2, “probably would not purchase”; 3, “not sure”; 4, “probably would purchase”; and 5, “definitely would purchase.” For an initial analysis, you will record the responses 1, 2, and 3 as “No” and 4 and 5 as “Yes.” What sample size would you use if you wanted the 95% margin of error to be 0.10 or less?
- 8.7 **More information is needed.** Refer to the previous exercise. Suppose that after reviewing the results of the previous survey, you proceeded with preliminary development of the product. Now you are at the stage where you need to decide whether or not to make a major investment to produce and market the product. You will use another random sample of your customers, but now you want the margin of error to be smaller. What sample size would you use if you wanted the 95% margin of error to be 0.05 or less?

SECTION 8.1 SUMMARY

- Inference about a population proportion p from an SRS of size n is based on the sample proportion $\hat{p} = X/n$ and the Wilson estimate $\tilde{p} = (X + 2)/(n + 4)$, which is just the sample proportion when we add 2 successes and 2 failures to the data. When n is large, \hat{p} has approximately the Normal distribution with mean p and standard deviation $\sqrt{p(1 - p)/n}$, and \tilde{p} is approximately Normal with mean p and standard deviation $\sqrt{p(1 - p)/(n + 4)}$.
- An approximate level C confidence interval for p is

$$\tilde{p} \pm z^* \text{SE}_{\tilde{p}}$$

where z^* is the value for the standard Normal density curve with area C between $-z^*$ and z^* , and the **standard error of \tilde{p}** is

$$\text{SE}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}}$$

- The sample size required to obtain a confidence interval of approximate margin of error m for a proportion is found from

$$n + 4 = \left(\frac{z^*}{m}\right)^2 p^*(1 - p^*)$$

where p^* is a guessed value for the proportion, and z^* is the standard Normal critical value for the desired level of confidence. To ensure that the margin of error of the interval is less than or equal to m no matter what \tilde{p} may be, use

$$n + 4 = \left(\frac{z^*}{2m}\right)^2$$

- Tests of $H_0: p = p_0$ are based on the z statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

with P -values calculated from the $N(0, 1)$ distribution.

SECTION 8.1 EXERCISES

- 8.8 **Do job applicants lie?** When trying to hire managers and executives, companies sometimes verify the academic credentials described by the applicants. One company that performs these checks summarized its findings for a six-month period. Of the 84 applicants whose credentials were checked, 15 lied about having a degree.⁸
- Find the Wilson estimate of the proportion of applicants who lie about having a degree and the standard error of this estimate.
 - Consider these data to be a random sample of credentials from a large collection of similar applicants. Give a 90% confidence interval for the true proportion of applicants who lie about having a degree.
- 8.9 Suppose that 9 of the 84 applicants checked in the previous exercise lied about their major. Can we conclude that a total of $24 = 15 + 9$ applicants lied about having a degree or about their major? Explain your answer.
- 8.10 **Christmas tree marketing.** One question in the Christmas tree market survey described in Case 8.2 was “Did you have a Christmas tree last year?” Of the 500 respondents, 421 answered “Yes.”
- Find the Wilson estimate of the population proportion and its standard error.
 - Give a 95% confidence interval for the proportion of Indiana households that had a Christmas tree last year.
- 8.11 **Shipping the orders on time.** As part of a quality improvement program, your mail-order company is studying the process of filling customer orders.



9.24

According to company standards, an order is shipped on time if it is sent within 3 working days of the time it is received. You select an SRS of 100 of the 5000 orders received in the past month for an audit. The audit reveals that 86 of these orders were shipped on time. Find a 95% confidence interval for the true proportion of the month's orders that were shipped on time.

- 8.12 **Power companies and trimming trees.** Large trees growing near power lines can cause power failures during storms when their branches fall on the lines. Power companies spend a great deal of time and money trimming and removing trees to prevent this problem. Researchers are developing hormone and chemical treatments that will stunt or slow tree growth. If the treatment is too severe, however, the tree will die. In one series of laboratory experiments on 216 sycamore trees, 41 trees died. Give a 99% confidence interval for the proportion of sycamore trees that would be expected to die from this particular treatment.
- 8.13 **Financial goals of college students.** In recent years over 70% of first-year college students responding to a national survey have identified "being well-off financially" as an important personal goal. A state university finds that 132 of an SRS of 200 of its first-year students say that this goal is important. Give a 95% confidence interval for the proportion of all first-year students at the university who would identify being well-off as an important personal goal.
- 8.14 **Alcohol abuse on campus.** College presidents have described alcohol abuse as the number one problem on campus. How common is it? A survey of 17,096 students in U.S. four-year colleges collected information on drinking behavior and alcohol-related problems.⁹ The researchers defined "frequent binge drinking" as having five or more drinks in a row three or more times in the past two weeks. According to this definition, 3314 students were classified as frequent binge drinkers. Find a 99% confidence interval for the proportion of frequent binge drinkers in this population.
- 8.15 **Bicycle accidents and alcohol.** In the United States approximately 900 people die in bicycle accidents each year. One study examined the records of 1711 bicyclists aged 15 or older who were fatally injured in bicycle accidents between 1987 and 1991 and were tested for alcohol.¹⁰ Of these, 542 tested positive for alcohol (blood alcohol concentration of 0.01% or higher).
- Summarize the data with appropriate descriptive statistics.
 - To do statistical inference for these data, we think of p as the probability that a tested bicycle rider is positive for alcohol. Find a 95% confidence interval for p .
 - Can you conclude from your statistical analysis of this study that alcohol causes fatal bicycle accidents?
- 8.16 **What proportion were legally drunk?** The study mentioned in the previous exercise found that 386 bicyclists had blood alcohol levels above 0.10%, a level defining legally drunk in many states. Give a 95% confidence interval for the proportion who were legally drunk according to this criterion.
- 8.17 **Can we use the z test?** In each of the following cases, is the sample large enough to permit safe use of the z test? (The population is very large.)

- (a) $n = 10$ and $H_0: p = 0.4$.
 (b) $n = 100$ and $H_0: p = 0.6$.
 (c) $n = 1000$ and $H_0: p = 0.996$.
 (d) $n = 500$ and $H_0: p = 0.3$.
- 8.18 **Checking the demographics of a sample.** Of the 500 households that responded to the Christmas tree marketing survey, 38% were from rural areas (including small towns), and the other 62% were from urban areas (including suburbs). According to the census, 36% of Indiana households are in rural areas, and the remaining 64% are in urban areas. Let p be the proportion of rural respondents. Set up hypotheses about p_0 and perform a test of significance to examine how well the sample represents the state in regard to rural versus urban residence. Summarize your results. CASE 8.2 
- 8.19 In the previous exercise we arbitrarily chose to state the hypotheses in terms of the proportion of rural respondents. We could as easily have used the proportion of *urban* respondents.
 (a) Write hypotheses in terms of the proportion of urban residents to examine how well the sample represents the state in regard to rural versus urban residence.
 (b) Perform the test of significance and summarize the results.
 (c) Compare your results with the results of the previous exercise. Summarize and generalize your conclusion.
- 8.20 **Vouchers for schools?** A national opinion poll found that 44% of all American adults agree that parents should be given vouchers good for education at any public or private school of their choice. The result was based on a small sample. How large an SRS is required to obtain a margin of error of ± 0.03 (that is, $\pm 3\%$) in a 95% confidence interval? (Use the previous poll's result to obtain the guessed value p^* .)
- 8.21 **Insect infestations.** An entomologist samples a field for egg masses of a harmful insect by placing a yard-square frame at random locations and carefully examining the ground within the frame. An SRS of 75 locations selected from a county's pastureland found egg masses in 13 locations. Give a 95% confidence interval for the proportion of all possible locations that are infested. CASE 8.2 
- 8.22 **Profile of the survey respondents.** Of the 500 respondents in the Christmas tree market survey of Case 8.2, 44% had no children at home and 56% had at least one child at home. The corresponding figures for the most recent census are 48% with no children and 52% with at least one child. Test the null hypothesis that the telephone survey technique has a probability of selecting a household with no children that is equal to the value obtained by the census. Give the z statistic and the P -value. What do you conclude? CASE 8.2 
- 8.23 **Mathematician tosses coin 10,000 times!** The South African mathematician John Kerrich, while a prisoner of war during World War II, tossed a coin 10,000 times and obtained 5067 heads.
 (a) Is this significant evidence at the 5% level that the probability that Kerrich's coin comes up heads is not 0.5?

- (b) Give a 95% confidence interval to see what probabilities of heads are roughly consistent with Kerrich's result.
- 8.24 Instant versus fresh-brewed coffee. A matched pairs experiment compares the taste of instant coffee with fresh-brewed coffee. Each subject tastes two unmarked cups of coffee, one of each type, in random order and states which he or she prefers. Of the 50 subjects who participate in the study, 19 prefer the instant coffee and the other 31 prefer fresh-brewed. Take p to be the proportion of the population that prefers fresh-brewed coffee.
- Test the claim that a majority of people prefer the taste of fresh-brewed coffee. Report the z statistic and its P -value. Is your result significant at the 5% level? What is your practical conclusion?
 - Find a 90% confidence interval for p .
- 8.25 Free throws. Leroy, a starting player for a major college basketball team, made only 38.4% of his free throws last season. During the summer he worked on developing a softer shot in the hope of improving his free-throw accuracy. In the first eight games of this season, Leroy made 25 free throws in 40 attempts. Let p be his probability of making each free throw that he shoots this season.
- State the null hypothesis H_0 that Leroy's free-throw probability has remained the same as last year and the alternative H_a that his work in the summer resulted in a higher probability of success.
 - Calculate the z statistic for testing H_0 versus H_a .
 - Find the P -value. Do you accept or reject H_0 at the $\alpha = 0.05$ significance level?
 - Give a 90% confidence interval for Leroy's free-throw success probability for the new season. Are you convinced that he is now a better free-throw shooter than last season?
 - What assumptions are needed for the validity of the test and confidence interval calculations that you performed?
- 8.26 Student employment. You want to estimate the proportion of students at your college or university who are employed for 10 or more hours per week while classes are in session. You plan to present your results by a 95% confidence interval. Using the guessed value $p^* = 0.35$, find the sample size required if the interval is to have an approximate margin of error of $m = 0.05$.
- 8.27 High-income households on a mailing list. Land's Beginning sells merchandise through the mail. It is considering buying a list of addresses from a magazine. The magazine claims that at least 20% of its subscribers have high incomes (that is, household income in excess of \$100,000). Land's Beginning would like to estimate the proportion of high-income people on the list. Verifying income is difficult, but another company offers this service. Land's Beginning will pay to verify the incomes of an SRS of people on the magazine's list. They would like the margin of error of the 95% confidence interval for the proportion to be 0.05 or less. Use the guessed value $p^* = 0.2$ to find the required sample size.
- 8.28 Change the specs. Refer to the previous exercise. For each of the following variations on the design specifications, state whether the required sample size will be higher, lower, or the same as that found above.

- (a) Use a 99% confidence interval.
 - (b) Change the allowable margin of error to 0.01.
 - (c) Use a planning value of $p^* = 0.15$.
 - (d) Use a different company to do the income verification.
- 8.29 **Start a student nightclub?** A student organization wants to start a nightclub for students under the age of 21. To assess support for this proposal, the organization will select an SRS of students and ask each respondent if he or she would patronize this type of establishment. About 70% of the student body are expected to respond favorably. What sample size is required to obtain a 90% confidence interval with an approximate margin of error of 0.04? Suppose that 50% of the sample responds favorably. Calculate the margin of error of the 90% confidence interval.
- 8.30 **Are the customers dissatisfied?** An automobile manufacturer would like to know what proportion of its customers are dissatisfied with the service received from their local dealer. The customer relations department will survey a random sample of customers and compute a 99% confidence interval for the proportion that are dissatisfied. From past studies, they believe that this proportion will be about 0.2. Find the sample size needed if the margin of error of the confidence interval is to be about 0.015. Suppose 10% of the sample say that they are dissatisfied. What is the margin of error of the 99% confidence interval?
- 8.31 **Increase student fees?** You have been asked to survey students at a large college to determine the proportion that favors an increase in student fees to support an expansion of the student newspaper. Each student will be asked whether he or she is in favor of the proposed increase. Using records provided by the registrar you can select a random sample of students from the college. After careful consideration of your resources, you decide that it is reasonable to conduct a study with a sample of 100 students. Construct a table of the margins of error for 95% confidence when \hat{p} takes the values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9.
- 8.32 **Justify the cost of the survey.** A former editor of the student newspaper agrees to underwrite the study in the previous exercise because she believes the results will demonstrate that most students support an increase in fees. She is willing to provide funds for a sample of size 500. Write a short summary for your benefactor of why the increased sample size will provide better results.

8.2

Comparing Two Proportions

Because comparative studies are so common, we often want to compare the proportions of two groups (such as men and women) that have some characteristic. We call the two groups being compared Population 1 and Population 2, and the two population proportions of “successes” p_1 and p_2 . The data consist of two independent SRSs. The sample sizes are n_1 for Population 1 and n_2 for Population 2. The proportion of successes in

each sample estimates the corresponding population proportion. Here is the notation we will use in this section:

Population	Population proportion	Sample size	Count of successes	Sample proportion
1	p_1	n_1	X_1	$\hat{p}_1 = X_1/n_1$
2	p_2	n_2	X_2	$\hat{p}_2 = X_2/n_2$

To compare the two unknown population proportions, start with the observed difference between the two sample proportions,

$$D = \hat{p}_1 - \hat{p}_2$$

When both sample sizes are sufficiently large, the sampling distribution of the difference D is approximately Normal. What are the mean and the standard deviation of D ? Each of the two \hat{p} 's has the mean and standard deviation given in the box on page 505. Because the two samples are independent, the two \hat{p} 's are also independent. We can apply the rules for means and variances of sums of random variables. Here is the result, which is summarized in Figure 8.5.

SAMPLING DISTRIBUTION OF $\hat{p}_1 - \hat{p}_2$

Choose independent SRSs of sizes n_1 and n_2 from two populations with proportions p_1 and p_2 of successes. Let $D = \hat{p}_1 - \hat{p}_2$ be the difference between the two sample proportions of successes. Then

- As both sample sizes increase, the sampling distribution of D becomes approximately Normal.
- The mean of the sampling distribution is $p_1 - p_2$.
- The standard deviation of the sampling distribution is

$$\sigma_D = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

APPLY YOUR KNOWLEDGE

- 8.33 Rules for means and variances. It is quite easy to verify the mean and standard deviation of the difference D .
- What are the means and standard deviations of the two sample proportions \hat{p}_1 and \hat{p}_2 ? (Look at the box on page 505 if you need to review this.)
 - Use the addition rule for means of random variables: What is the mean of $D = \hat{p}_1 - \hat{p}_2$?
 - The two samples are independent. Use the addition rule for variances of random variables: What is the variance of D ?

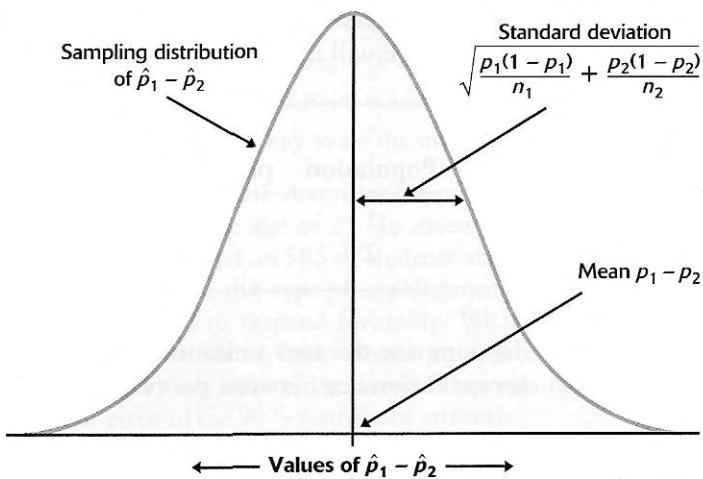


FIGURE 8.5 The sampling distribution of the difference of two sample proportions is approximately Normal. The mean and standard deviation are found from the two population proportions of successes, p_1 and p_2 .

Confidence intervals

Just as in the case of estimating a single proportion, a small modification of the sample proportions greatly improves the accuracy of confidence intervals.¹¹ As before, add 2 successes and 2 failures to the actual data, and now divide them equally between the two samples. That is, *add 1 success and 1 failure to each sample*. We will again call the estimates produced by adding hypothetical observations **Wilson estimates**. The Wilson estimates of the two population proportions are

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} \quad \text{and} \quad \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2}$$

The estimated difference between the populations is

$$\tilde{D} = \tilde{p}_1 - \tilde{p}_2$$

The standard deviation of \tilde{D} is approximately

$$\sigma_{\tilde{D}} = \sqrt{\frac{p_1(1-p_1)}{n_1 + 2} + \frac{p_2(1-p_2)}{n_2 + 2}}$$

This is just the result given in the box for D (page 521), adjusted to the sizes of the modified samples.

To obtain a confidence interval for $p_1 - p_2$, we once again replace the unknown parameters in the standard deviation by estimates to obtain an estimated standard deviation, or standard error. Here is the confidence interval we want.

Wilson estimates

CONFIDENCE INTERVALS FOR COMPARING TWO PROPORTIONS

Choose an SRS of size n_1 from a large population having proportion p_1 of successes and an independent SRS of size n_2 from another population having proportion p_2 of successes. An approximate level C confidence interval for $p_1 - p_2$ is

$$(\tilde{p}_1 - \tilde{p}_2) \pm z^* \text{SE}_{\tilde{D}}$$

where

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} \quad \text{and} \quad \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2}$$

are the Wilson estimates of the population proportions, and the standard error of the difference is

$$\text{SE}_{\tilde{D}} = \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 + 2}}$$

and z^* is the value for the standard Normal density curve with area C between $-z^*$ and z^* . The margin of error is

$$m = \pm z^* \text{SE}_{\tilde{D}}$$

Use this method when both sample sizes are at least 10 and the confidence level is 90%, 95%, or 99%.

CASE 8.3



"NO SWEAT" GARMENT LABELS

Following complaints about the working conditions in some apparel factories both in the United States and abroad, a joint government and industry commission recommended in 1998 that companies that monitor and enforce proper standards be allowed to display a "No Sweat" label on their products. Does the presence of these labels influence consumer behavior?

A survey of U.S. residents aged 18 or older asked a series of questions about how likely they would be to purchase a garment under various conditions. For some conditions, it was stated that the garment had a "No Sweat" label; for others, there was no mention of such a label. On the basis of the responses, each person was classified as a "label user" or a "label nonuser."¹² About 16.5% of those surveyed were label users. One purpose of the study was to describe the demographic characteristics of users and nonusers.



EXAMPLE 8.5**Gender differences in label use**

The study in Case 8.3 suggested that there is a gender difference in the proportion of label users. Here is a summary of the data. We let X denote the number of label users.

Population	n	X	$\hat{p} = X/n$	$\tilde{p} = (X + 1)/(n + 2)$
1 (women)	296	63	0.213	0.215
2 (men)	251	27	0.108	0.111

In this table the \hat{p} column gives the sample proportions of label users, and the \tilde{p} column gives the Wilson estimates that we use to construct a confidence interval for the difference. First calculate the standard error of the observed difference:

$$\begin{aligned} \text{SE}_{\tilde{D}} &= \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 + 2}} \\ &= \sqrt{\frac{(0.215)(0.785)}{296 + 2} + \frac{(0.111)(0.889)}{251 + 2}} \\ &= 0.0308 \end{aligned}$$

The 95% confidence interval is

$$\begin{aligned} (\tilde{p}_1 - \tilde{p}_2) \pm z^* \text{SE}_{\tilde{D}} &= (0.215 - 0.111) \pm (1.96)(0.0308) \\ &= 0.104 \pm 0.060 \\ &= (0.04, 0.16) \end{aligned}$$

With 95% confidence we can say that the difference in the proportions is between 0.04 and 0.16. Alternatively, we can report that the women are about 10% more likely to be label users than men, with a 95% margin of error of 6%.

Minitab and SAS output for Example 8.5 appear in Figure 8.6. To use the Wilson estimates for this problem, we altered the data set by adding four cases, one user and one nonuser for each gender.

In surveys such as this, men and women are typically not sampled separately. The respondents to a single sample are divided after the fact into men and women. The sample sizes are then random and reflect the characteristics of the population sampled. Two-sample significance tests and confidence intervals are still approximately correct in this situation, even though the two sample sizes were not fixed in advance.

Minitab

Success = 1

C2	X	N	Sample p
1	64	298	0.214765
2	28	253	0.110672

Estimate for $p(1) - p(2)$: 0.104093
 95% CI for $p(1) - p(2)$: (0.0435263, 0.164660)

SAS

	Risk	ASE	(Asymptotic) 95% Confidence Limits	
Row 1	0.2148	0.0238	0.1681	0.2614
Row 2	0.1107	0.0197	0.0720	0.1493
Total	0.1670	0.0159	0.1358	0.1981
Difference	0.1041	0.0309	0.0435	0.1647

FIGURE 8.6 Minitab and SAS output for Example 8.5.

In Example 8.5 we chose women to be the first population. Had we chosen men as the first population, the estimate of the difference would be negative (-0.104). Because it is easier to discuss positive numbers, we generally choose the first population to be the one with the higher proportion. The choice doesn't affect the substance of the analysis.

APPLY YOUR KNOWLEDGE

- 8.34 Who gets stock options? Different kinds of companies compensate their key employees in different ways. Established companies may pay higher salaries, while new companies may offer stock options that will be valuable if the company succeeds. Do high-tech companies tend to offer stock options more often than other companies? One study looked at a random sample of 200 companies. Of these, 91 were listed in the *Directory of Public High Technology Corporations* and 109 were not listed. Treat these two groups as SRSs of high-tech and non-high-tech companies. Seventy-three of the high-tech companies and 75 of the non-high-tech companies offered incentive stock options to key employees.¹³ Give a 95% confidence interval for the difference in the proportions of the two types of companies that offer stock options.
- 8.35 Unhappy HMO customers. A study was designed to find reasons why patients leave a health maintenance organization (HMO).¹⁴ Patients were classified as to whether or not they had filed a complaint with the HMO. We want to compare the proportion of complainers who leave the HMO with the proportion of those who do not file complaints. In the year of the study, 639 patients filed complaints, and 54 of these patients left the HMO voluntarily. For comparison, the HMO chose an SRS of 743 patients who had not filed complaints. Twenty-two of these patients left voluntarily. Give an estimate of the difference in the two proportions with a 95% confidence interval.

Significance tests

Although we prefer to compare two proportions by giving a confidence interval for the difference between the two population proportions, it is sometimes useful to test the null hypothesis that the two population proportions are the same.

We standardize $D = \hat{p}_1 - \hat{p}_2$ by subtracting its mean $p_1 - p_2$ and then dividing by its standard deviation

$$\sigma_D = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

If n_1 and n_2 are large, the standardized difference is approximately $N(0, 1)$. To get a confidence interval, we used sample estimates in place of the unknown population proportions p_1 and p_2 in the expression for σ_D . Although this approach would lead to a valid significance test, we follow the more common practice of replacing the unknown σ_D with an estimate that takes into account the null hypothesis that $p_1 = p_2$. If these two proportions are equal, we can view all of the data as coming from a single population. Let p denote the common value of p_1 and p_2 . The standard deviation of $D = \hat{p}_1 - \hat{p}_2$ is then

$$\begin{aligned}\sigma_{Dp} &= \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} \\ &= \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\end{aligned}$$

The subscript on σ_{Dp} reminds us that this is the standard deviation under the special condition that the two populations share a common proportion p of successes.

We estimate the common value of p by the overall proportion of successes in the two samples:

$$\hat{p} = \frac{\text{number of successes in both samples}}{\text{number of observations in both samples}} = \frac{X_1 + X_2}{n_1 + n_2}$$

pooled estimate of p

This estimate of p is called the **pooled estimate** because it combines, or pools, the information from both samples.

To estimate the standard deviation of D , substitute \hat{p} for p in the expression for σ_{Dp} . The result is a standard error for D under the condition that the null hypothesis $H_0: p_1 = p_2$ is true. The test statistic uses this standard error to standardize the difference between the two sample proportions.

SIGNIFICANCE TESTS FOR COMPARING TWO PROPORTIONS

Choose an SRS of size n_1 from a large population having proportion p_1 of successes and an independent SRS of size n_2 from another population having proportion p_2 of successes. To test the hypothesis

$$H_0: p_1 = p_2$$

compute the z statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\text{SE}_{Dp}}$$

where the pooled standard error is

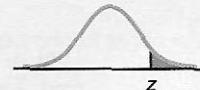
$$\text{SE}_{Dp} = \sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

based on the pooled estimate of the common proportion of successes,

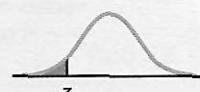
$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

In terms of a standard Normal random variable Z , the P -value for a test of H_0 against

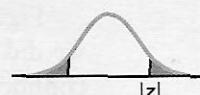
$$H_a: p_1 > p_2 \text{ is } P(Z \geq z)$$



$$H_a: p_1 < p_2 \text{ is } P(Z \leq z)$$



$$H_a: p_1 \neq p_2 \text{ is } 2P(|Z| \geq |z|)$$



Use this test when the number of successes and the number of failures in each of the samples is at least 5.

EXAMPLE 8.6



Men, women, and garment labels

Example 8.5 (page 524) presents survey data on whether consumers are “label users” who pay attention to label details when buying a shirt. Are men and women equally likely to be label users? Here is the data summary:

Population	n	X	$\hat{p} = X/n$
1 (women)	296	63	0.213
2 (men)	251	27	0.108

The sample proportions are certainly quite different, but we need a significance test to verify that the difference is too large to easily result from the play of chance in choosing the sample. Formally, we compare the proportions of label users in the two populations (women and men) by testing the hypotheses

$$H_0: p_1 = p_2$$

$$H_a: p_1 \neq p_2$$

The pooled estimate of the common value of p is

$$\hat{p} = \frac{63 + 27}{296 + 251} = \frac{90}{547} = 0.1645$$

This is just the proportion of label users in the entire sample.

The test statistic is calculated as follows:

$$\begin{aligned} \text{SE}_{Dp} &= \sqrt{(0.1645)(0.8355)\left(\frac{1}{296} + \frac{1}{251}\right)} = 0.03181 \\ z &= \frac{\hat{p}_1 - \hat{p}_2}{\text{SE}_{Dp}} = \frac{0.213 - 0.108}{0.03181} \\ &= 3.30 \end{aligned}$$

That is, the observed difference is more than 3 standard deviations away from zero. The P -value is $2P(Z \geq 3.30)$. From Table A we have $P = 2 \times 0.0005 = 0.0010$. Software gives $P = 0.0009$. We report: 21% of women are label users versus only 11% of men; the difference is statistically significant ($z = 3.30$, $P < 0.001$).

Figure 8.7 gives the Minitab and SAS outputs for Example 8.6. Note that we did not alter the data set by adding additional cases as we did for the confidence interval. Carefully examine the output to find all of the important pieces that you would need to report the results of the analysis and to draw a conclusion.

Many market researchers would expect the proportion of label users to be higher among women than among men. That is, we might choose the one-sided alternative $H_a: p_1 > p_2$. The P -value would be half of the value obtained for the two-sided test. Because the z statistic is so large, this distinction is of no practical importance.

- APPLY YOUR KNOWLEDGE**
- 8.36 Are the high-techs different? In Exercise 8.34 we looked at whether or not companies offered stock options to their employees. There, we compared high-tech companies with other companies using a 95% confidence interval. Let's now test the null hypothesis that the two types of companies are equally likely to offer this kind of benefit to their employees. In the sample, 73 of the 91 high-tech companies and 75 of the 109 other companies offered incentive stock options to key employees. State appropriate null and alternative

Minitab

Success = 1

C2	X	N	Sample p
1	63	296	0.212838
2	27	251	0.107570

Estimate for $p(1) - p(2)$: 0.105268Test for $p(1) - p(2) = 0$ (vs not = 0): $Z = 3.31$ P-Value = 0.001**SAS**Two Sample Test of Equality of Proportions
Sample Statistics

Value	- Frequencies of x for gen -	
	1	2
0	233	224
1	63	27

Hypothesis Test

Null hypothesis:

Proportion of $x(\text{gen} = 1)$ - Proportion of $x(\text{gen} = 2) = 0$

Alternative:

Proportion of $x(\text{gen} = 1)$ - Proportion of $x(\text{gen} = 2) \neq 0$

Value	- Proportions of x for gen -		Z	Prob > Z
	1	2		
1	0.2128	0.1076	3.31	0.0009

FIGURE 8.7 Minitab and SAS output for Example 8.6.

hypotheses, compute the test statistic, and report the P -value. Give a brief statement of your conclusion.

- 8.37 Are the unhappy HMO customers likely to leave? In Exercise 8.35 we examined data from a study designed to find reasons why patients leave a health maintenance organization (HMO). There we compared the proportion of complainers who leave the HMO with the proportion of noncomplainers who leave. In the year of the study, 639 patients filed complaints and 54 of these patients left the HMO voluntarily. For comparison, the HMO chose an SRS of 743 patients who had not filed complaints. Twenty-two of those patients left voluntarily. We expect a higher proportion of complainers to leave. Do the data support this expectation? State hypotheses, find the test statistic and its P -value, and state your conclusion.

relative risk

BEYOND THE BASICS: RELATIVE RISK

In Example 8.5 (page 524) we compared the proportions of women and men who are “label users” when they shop for a shirt by giving a confidence interval for the *difference* of proportions. We might alternatively choose to make this comparison by giving the *ratio* of the two proportions. This ratio is often called the **relative risk** (RR). A relative risk of 1 means that the proportions \hat{p}_1 and \hat{p}_2 are equal. Confidence intervals for relative risk apply the principles that we have studied, but the details are somewhat complicated. Fortunately, we can leave the details to software and concentrate on interpreting and communicating the results.

EXAMPLE 8.7



Relative risk for use of labels

The following table summarizes the data on the proportions of men and women who use labels when buying a shirt:

Population	n	X	$\hat{p} = X/n$
1 (women)	296	63	0.2128
2 (men)	251	27	0.1076

The relative risk for this sample is

$$\text{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{0.2128}{0.1076} = 1.98$$

Confidence intervals for the relative risk in the entire population of shoppers are based on this sample relative risk. Software (for example, PROC FREQ with the MEASURES option in SAS) gives a 95% confidence interval as 1.30 to 3.01. Our summary: Women are about twice as likely as men to use labels; the 95% confidence interval is (1.30, 3.01).

In Example 8.7 the confidence interval is clearly not symmetric about the estimate: that is, 1.98 is not the midpoint of 1.30 and 3.01. This is true in general.

Relative risk—that is, comparing proportions by a ratio rather than by a difference—is particularly useful when the proportions are small. This is often the case in epidemiology and medical statistics. Here is a typical epidemiological example.

EXAMPLE 8.8

Blood pressure and heart disease

There is much evidence that high blood pressure is associated with increased risk of death from cardiovascular disease. A major study of this association examined 3338 men with high blood pressure and 2676 men with low blood pressure. During the period of the study, 21 men in the low-blood-pressure and 55 in the high-blood-pressure group died from cardiovascular disease.¹⁵

As usual, we choose the group with the higher proportion to be Population 1. The proportion of the high-blood-pressure group who died is $\hat{p}_1 = 0.0165$. For men with low blood pressure, the proportion is $\hat{p}_2 = 0.0078$. The difference is statistically significant ($z = 2.98$, $P = 0.0014$), and the 95% confidence interval for the difference in population proportions is (0.0032, 0.0141).

Because the proportions of men who died are both small, relative risk communicates the results of this study more effectively. The relative risk for the sample is

$$\text{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{0.0165}{0.0078} = 2.1$$

Software gives the 95% confidence interval for the population as 1.3 to 3.5. That is, the risk of dying from cardiovascular disease is doubled for men with high blood pressure compared with men who have low blood pressure (RR = 2.1, 95% confidence interval = 1.3 to 3.5).

SECTION 8.2 SUMMARY

- The approximate level C confidence interval for the difference in two proportions $p_1 - p_2$ is

$$(\tilde{p}_1 - \tilde{p}_2) \pm z^* \text{SE}_{\tilde{D}}$$

where the Wilson estimates of the population proportions are

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} \quad \text{and} \quad \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2}$$

and the standard error of the difference is

$$\text{SE}_{\tilde{D}} = \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 + 2}}$$

- Significance tests of $H_0: p_1 = p_2$ use the z statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\text{SE}_{Dp}}$$

with P -values from the $N(0, 1)$ distribution. In this statistic,

$$\text{SE}_{Dp} = \sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

where \hat{p} is the pooled estimate of the common value of p_1 and p_2 ,

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

- Relative risk is the ratio of two sample proportions:

$$\text{RR} = \frac{\hat{p}_1}{\hat{p}_2}$$

Confidence intervals for relative risk are an alternative to confidence intervals for the difference when we want to compare two proportions.

SECTION 8.2 EXERCISES

- 8.38 **A corporate liability trial.** A major court case on liability for contamination of groundwater took place in the town of Woburn, Massachusetts. A town well in Woburn was contaminated by industrial chemicals. During the period that residents drank water from this well, there were 16 birth defects among 414 births. In years when the contaminated well was shut off and water was supplied from other wells, there were 3 birth defects among 228 births. The plaintiffs suing the firms responsible for the contamination claimed that these data show that the rate of birth defects was higher when the contaminated well was in use.¹⁶ How statistically significant is the evidence? Be sure to state what assumptions your analysis requires and to what extent these assumptions seem reasonable in this case.
- 8.39 **Support for a marketing program.** Twenty states have “corn checkoff” programs that divert a small part of the sale price of corn (typically one-half cent per bushel) to support corn product marketing and research. Some checkoff programs are voluntary and some are mandatory. The Indiana Department of Agriculture asked random samples of corn producers in each county whether they favored a mandatory program. In Tippecanoe County, 263 farmers were in favor of the program and 252 were not. In neighboring Benton County, 260 were in favor and 377 were not.
- Find the proportions of farmers in favor of the program in each of the two counties.
 - Find the standard error needed to compute a confidence interval for the difference in the proportions.
 - Compute a 99% confidence interval for the difference between the proportions of farmers favoring the program in Tippecanoe County and in Benton County. Do you think opinions differed in the two counties?
- 8.40 Refer to the survey of farmers described in the previous exercise.
- Formulate null and alternative hypotheses for comparing the proportions of farmers in the two counties who favor mandatory checkoff.
 - Combine the two samples and find the overall proportion of farmers who favor the corn checkoff program.
 - Find the standard error needed for testing the hypotheses you stated in (a).
 - Compute the z statistic and its P -value. What conclusion do you draw?
- 8.41 **Natural versus artificial Christmas trees.** In the Christmas tree survey introduced in Case 8.2 (page 514), respondents who had a tree during the holiday season were asked whether the tree was natural or artificial. Respondents were also asked if they lived in an urban area or in a rural area. Of the 421 households displaying a Christmas tree, 160 lived in rural areas and 261 were urban residents. The tree growers want to know if there is a difference in preference for natural trees versus artificial trees between urban and rural households. Here are the data:



Population	n	$X(\text{natural})$
1 (rural)	160	64
2 (urban)	261	89

- 9.21
- (a) Give the null and alternative hypotheses that are appropriate for this problem assuming that we have no prior information suggesting that one population would have a higher preference than the other.
 - (b) Test the null hypothesis. Give the test statistic and the P -value, and summarize the results.
 - (c) Give a 90% confidence interval for the difference in proportions.

- 8.42 **Summer employment of college students.** A university financial aid office polled an SRS of undergraduate students to study their summer employment. Not all students were employed the previous summer. Here are the results for men and women:

	Men	Women
Employed	718	593
Not employed	79	139
Total	797	732

- (a) Is there evidence that the proportion of male students employed during the summer differs from the proportion of female students who were employed? State H_0 and H_a , compute the test statistic, and give its P -value.
 - (b) Give a 99% confidence interval for the difference between the proportions of male and female students who were employed during the summer. Does the difference seem practically important to you?
- 8.43 **A hazardous work environment.** The power takeoff driveline on farm tractors can be a serious hazard to farmers. A shield covers the driveline on new tractors, but for a variety of reasons, the shield is often missing on older tractors. Two types of shield are the bolt-on and the flip-up. A study initiated by the National Safety Council took a sample of older tractors to examine the proportions of shields removed. The study found that 35 shields had been removed from the 83 tractors having bolt-on shields and that 15 had been removed from the 136 tractors with flip-up shields.¹⁷
- (a) Test the null hypothesis that there is no difference between the proportions of the two types of shields removed. Give the z statistic and the P -value. State your conclusion in words.
 - (b) Give a 90% confidence interval for the difference in the proportions of removed shields for the bolt-on and the flip-up types. Based on the data, what recommendation would you make about the type of shield to be used on new tractors?
- 8.44 **Gender bias in textbooks.** To what extent do textbooks on syntax (analysis of sentence structure) display gender bias? A study of this question sampled sentences from 10 texts.¹⁸ One part of the study examined the use of the words “girl,” “boy,” “man,” and “woman.” Call the first two words *juvenile* and the last two *adult*. Is the proportion of female references that are juvenile (“girl” rather than “woman”) equal to the proportion of male

references that are juvenile (“boy” rather than “man”)? Here are data from one of the texts:

Gender	<i>n</i>	<i>X(juvenile)</i>
Female	60	48
Male	132	52

- (a) Find the Wilson estimates of the proportions of juvenile references for females and for males.
- (b) Give a 95% confidence interval for the difference and briefly summarize what the data show.

- 8.45 **Bicycle accidents, alcohol, and gender.** In Exercise 8.15 (page 517) we examined the percent of fatally injured bicyclists tested for alcohol who tested positive. Here are the same data broken down by gender:

Gender	<i>n</i>	<i>X(tested positive)</i>
Female	191	27
Male	1520	515

- (a) Summarize the data by giving the Wilson estimates of the two population proportions and a 90% confidence interval for their difference.
- (b) The standard error $SE_{\tilde{D}}$ contains a contribution from each sample, $\tilde{p}_1(1 - \tilde{p}_1)/(n_1 + 2)$ and $\tilde{p}_2(1 - \tilde{p}_2)/(n_2 + 2)$. Which of these contributes the larger amount to the standard error of the difference? Explain why.

- 8.46 **Is the gender bias statistically significant?** Exercise 8.44 addresses a question about gender bias with a confidence interval. Set up the problem as a significance test. Carry out the test and summarize the results.

- 8.47 **Are the gender differences statistically significant?** The proportions of fatally injured female and male bicyclists were compared with a confidence interval in Exercise 8.45. Examine the same data with a test of significance.

- 8.48 **Lying by job applicants.** Is lying about credentials by job applicants changing? In Exercise 8.8 (page 516) we looked at the proportion of applicants who lied about having a degree in a six-month period. To see if there is a change over time, we can compare that period with the following six months. Here are the data:

Period	<i>n</i>	<i>X(lied)</i>
1	84	15
2	106	21

Use a 90% confidence interval to address the question of interest.

- 8.49 **Did the Yankees have a home field advantage?** In the 2000 regular baseball season, the World Series Champion New York Yankees played 80 games at home and 81 games away. They won 44 of their home games and 43 of the games played away. We can consider these games as samples from potentially large populations of games played at home and away. How much advantage does the Yankee home field provide?

- (a) Find the Wilson estimate of the proportion of wins for all home games.
Do the same for away games.
- (b) Find the standard error needed to compute a confidence interval for the difference in the proportions.
- (c) Compute a 90% confidence interval for the difference between the probability that the Yankees win at home and the probability that they win when on the road. Are you convinced that the Yankees were more likely to win at home in 2000?
- 8.50 **Have the lies increased?** Data on the proportion of applicants who lied about having a degree in two consecutive six-month periods appear in Exercise 8.48. Is there evidence of a change over time? State hypotheses, carry out a significance test, and summarize the results.
- 8.51 **Is it easier to win at home?** Return to the New York Yankees data in Exercise 8.49.
- Most people think that it is easier to win at home than away. State null and alternative hypotheses to test this supposition.
 - Combining all of the games played, what proportion did the Yankees win?
 - Find the standard error needed for testing that the probability of winning is the same at home and away.
 - Compute the z statistic and its P -value. What conclusion do you draw?
- 8.52 **What about the Mets?** In the 2000 World Series the New York Yankees played the New York Mets. Exercise 8.49 examines the Yankees' home and away victories. During the regular season the Mets won 55 of the 84 home games that they played and 39 of the 81 games that they played away. Perform the same analysis for the Mets as you did in Exercise 8.49, and write a short summary comparing these results with those you found for the Yankees.
- 8.53 **Chromosomes and crime.** A study of chromosomal abnormalities and criminality examined data on 4124 Danish males born in Copenhagen.¹⁹ The study used the penal registers maintained in the offices of the local police chiefs and classified each man as having a criminal record or not. Each was also classified as having the normal male XY chromosome pair or one of the abnormalities XYY or XXY. Of the 4096 men with normal chromosomes, 381 had criminal records, while 8 of the 28 men with chromosomal abnormalities had criminal records. Some experts believe that chromosomal abnormalities are associated with increased criminality. Do these data lend support to this belief? Report your analysis and draw a conclusion.
- 8.54 **Aspirin and stroke.** A clinical trial examined the effectiveness of aspirin in the treatment of cerebral ischemia (stroke).²⁰ Patients were randomized into treatment and control groups. The study was double-blind in the sense that neither the patients nor the physicians who evaluated the patients knew which patients received aspirin and which the placebo tablet. After six months of treatment, the attending physicians evaluated each patient's progress as either favorable or unfavorable. Of the 78 patients in the aspirin group, 63 had favorable outcomes; 43 of the 77 control patients had favorable outcomes.

- (a) Compute the Wilson estimates of the proportions of patients having favorable outcomes in the two populations.
- (b) Give a 95% confidence interval for the difference between the favorable proportions in the populations.
- (c) The physicians conducting the study believed from previous research that aspirin was likely to increase the chance of a favorable outcome. Carry out a significance test to confirm this conclusion. State hypotheses, find the P -value, and write a summary of your results.
- 8.55 **Products to control cockroaches.** The pesticide diazinon is in common use to treat infestations of the German cockroach, *Blattella germanica*. A study investigated the persistence of this pesticide on various types of surfaces. Researchers applied a 0.5% emulsion of diazinon to glass and plasterboard. After 14 days, they placed 18 cockroaches on each surface and recorded the number that died within 48 hours. On glass, 9 cockroaches died, while on plasterboard, 13 died.²¹
- (a) Find a 90% confidence interval for the difference in the two population proportions of dead cockroaches.
- (b) Chemical analysis of the residues of diazinon suggests that it may persist longer on plasterboard than on glass because it binds to the paper covering on the plasterboard. The researchers therefore expected the mortality rate to be greater on plasterboard than on glass. Conduct a significance test to assess the evidence that this is true.
- 8.56 **Effect of the sample size.** Return to the study of undergraduate student summer employment described in Exercise 8.42. Similar results from a smaller number of students may not have the same statistical significance. Specifically, suppose that 72 of 80 men surveyed were employed and 59 of 73 women surveyed were employed. The sample proportions are essentially the same as in the earlier exercise.
- (a) Compute the z statistic for these data and report the P -value. What do you conclude?
- (b) Compare the results of this significance test with your results in Exercise 8.42. What do you observe about the effect of the sample size on the results of these significance tests?
- 8.57 **More cockroaches.** Suppose that the experiment of Exercise 8.55 placed more cockroaches on each surface and observed similar mortality rates. Specifically, suppose that 36 cockroaches were placed on each surface and that 26 died on the plasterboard, while 18 died on the glass.
- (a) Compute the z statistic for these data and report its P -value. What do you conclude?
- (b) Compare the results of this significance test with those you gave in Exercise 8.55. What do you observe about the effect of the sample size on the results of these significance tests?

STATISTICS IN SUMMARY

Inference about population proportions is based on sample proportions. We rely on the fact that a sample proportion has a distribution that is close to Normal unless the sample is small. All the z procedures in this chapter work

well when the samples are large enough. You must check this before using them. Here are the things you should now be able to do.

A. RECOGNITION

1. Recognize from the design of a study whether one-sample, matched pairs, or two-sample procedures are needed.
2. Recognize what parameter or parameters an inference problem concerns. In particular, distinguish among settings that require inference about a mean, comparing two means, inference about a proportion, or comparing two proportions.
3. Calculate from sample counts the sample proportion or proportions and the Wilson estimates of population proportions.

B. INFERENCE ABOUT ONE PROPORTION

1. Use the z procedure to give a confidence interval for a population proportion p .
2. Use the z statistic to carry out a test of significance for the hypothesis $H_0: p = p_0$ about a population proportion p against either a one-sided or a two-sided alternative.
3. Check that you can safely use these z procedures in a particular setting.

C. COMPARING TWO PROPORTIONS

1. Use the two-sample z procedure to give a confidence interval for the difference $p_1 - p_2$ between proportions in two populations based on independent samples from the populations.
2. Use a z statistic to test the hypothesis $H_0: p_1 = p_2$ that proportions in two distinct populations are equal.
3. Check that you can safely use these z procedures in a particular setting.

Statistical inference always draws conclusions about one or more parameters of a population. When you think about doing inference, ask first what the population is and what parameter you are interested in. The t procedures of Chapter 7 allow us to give confidence intervals and carry out tests about population means. We use the z procedures of this chapter for inference about population proportions.

CHAPTER 8 REVIEW EXERCISES

- 8.58 Time to repair golf clubs. The Ping Company makes custom-built golf clubs and competes in the \$4 billion golf equipment industry. To improve its business processes, Ping decided to seek ISO 9001 certification.²² As part of this process, a study of the time it took to repair golf clubs sent to the company by mail determined that 16% of orders were sent back to the customers in 5 days or less. Ping examined the processing of repair orders and made changes. Following the changes, 90% of orders were completed

within 5 days. Assume that each of the estimated percents is based on a random sample of 200 orders.

- (a) How many orders were completed in 5 days or less before the changes? Give a 95% confidence interval for the proportion of orders completed in this time.
- (b) Do the same for orders after the changes.
- (c) Give a 95% confidence interval for the improvement. Express this both for a difference in proportions and for a difference in percents.

8.59 **Demographics of Internet users.** To devise effective marketing strategies it is helpful to know the characteristics of your customers. A study compared demographic characteristics of people who use the Internet for travel arrangements and of people who do not.²³ Of 1132 Internet users, 643 had completed college. Among the 852 nonusers, 349 had completed college.

- (a) Do users and nonusers differ significantly in the proportion of college graduates?
- (b) Give a 95% confidence interval for the difference in the proportions.

8.60 **Income of Internet users.** The study mentioned in the previous exercise also asked about income. Among Internet users, 493 reported income of less than \$50,000 and 378 reported income of \$50,000 or more. (Not everyone answered the income question.) The corresponding numbers for nonusers were 477 and 200. Perform a significance test to compare the incomes of users with nonusers and also give an estimate of the difference in proportions with a 95% margin of error.

8.61 **Nonresponse.** Refer to the previous two exercises. Give the total number of users and the total number of nonusers for the analysis of education. Do the same for the analysis of income. The difference is due to respondents who chose “Rather not say” for the income question. Give the proportions of “Rather not say” individuals for users and nonusers. Perform a significance test to compare these and give a 95% confidence interval for the difference. People are often reluctant to provide information about their income. Do you think that this type of nonresponse for the income is a serious limitation for this study?

8.62 **Brand loyalty and the Chicago Cubs.** According to literature on brand loyalty, consumers who are loyal to a brand are likely to consistently select the same product. This type of consistency may come from a positive childhood association. To examine brand loyalty among fans of the Chicago Cubs, 371 Cubs fans among patrons of a restaurant located in Wrigleyville were surveyed before a game at Wrigley Field, the Cubs home field.²⁴ The respondents were classified as “die-hard fans” or “less loyal fans.” Of the 134 die-hard fans, 90.3% reported that they watched or listened to Cubs games when they were children. Among the 237 less loyal fans, 67.9% said that they watched or listened as children.

- (a) Find the numbers of die-hard Cubs fans who watched or listened to games when they were children. Do the same for the less loyal fans.
- (b) Use a significance test to compare the die-hard fans with the less loyal fans with respect to their childhood experiences of the team.
- (c) Express the results with a 95% confidence interval for the difference in proportions.

Excel Output

9.21

- 8.63 **Brand loyalty in action.** The study mentioned in the previous exercise found that two-thirds of the die-hard fans attended Cubs games at least once a month, but only 20% of the less loyal fans attended this often. Analyze these data using a significance test and a confidence interval. Write a short summary of your findings.
- 8.64 **Credit cards.** A Gallup Poll used telephone interviews to survey a sample of 1025 U.S. residents over the age of 18 regarding their use of credit cards.²⁵ The poll reported that 76% of Americans said that they had at least one credit card. Give the 95% margin of error for this estimate.
- 8.65 **Do they pay off the monthly balance?** The Gallup Poll in the previous exercise reported that 41% of those who have credit cards do not pay the full balance each month. Find the number of people in the survey who said that they had at least one credit card, using the information in the previous exercise. Combine this number with the reported 41% to give a margin of error for the proportion of credit card holders who do not pay their full balance.
- 8.66 **Frequent lottery players.** A study of state lotteries included a random-digit-dialing (RDD) survey conducted by the National Opinion Research Center (NORC). The survey asked 2406 adults about their lottery spending.²⁶ A total of 248 individuals were classified as “heavy” players. Of these, 152 were male. The study notes that 48.5% of U.S. adults are male. Use a significance test to compare the proportion of males among heavy lottery players with the proportion of males in the U.S. adult population and write a short summary of your results. For this analysis, assume that the 248 heavy lottery players are a random sample of all heavy lottery players and that the margin of error for the 48.5% estimate of the percent of males in the U.S. adult population is so small that it can be neglected.
- 8.67 **Frequent lottery players, continued.** Use a confidence interval to give an alternative analysis for the previous exercise.
- 8.68 **Ability of children to distinguish new products.** Many new products are targeted toward children. The choice behavior of children with regard to new products is of interest to companies that design marketing strategies for these products. As part of one study, children in different age groups were compared on their ability to sort new products into the correct product category (milk or juice in this case).²⁷ Here are some of the data:

Age group	<i>n</i>	Number who sorted correctly
4- to 5-year-olds	50	10
6- to 7-year-olds	53	28

Test the null hypothesis that the two age groups are equally skilled at sorting. Justify your choice of an alternative hypothesis. Also, give a 90% confidence interval for the difference. Summarize your results with a short paragraph.

- 8.69 **Does the new process give a better product?** Eleven percent of the products produced by an industrial process over the past several months fail to conform to the specifications. The company modifies the process in an

attempt to reduce the rate of nonconformities. In a trial run, the modified process produces 16 nonconforming items out of a total of 300 produced. Do these results demonstrate that the modification is effective? Support your conclusion with a clear statement of your assumptions and the results of your statistical calculations.

- 8.70 **How much is the improvement?** In the setting of the previous exercise, give a 95% confidence interval for the proportion of nonconforming items for the modified process. Then, taking $p_0 = 0.11$ to be the old proportion and p the proportion for the modified process, give a 95% confidence interval for $p - p_0$.
- 8.71 **Binge drinking by men and women.** In Exercise 8.14 (page 517) we estimated the proportion of college students who engage in frequent binge drinking. Are there student characteristics related to this behavior? For example, how similar is frequent binge drinking among men and women? Here are counts of frequent binge drinkers by gender:

Population	<i>n</i>	X
1 (men)	7,180	1,630
2 (women)	9,916	1,684
Total	17,096	3,314

Write a short report on the size and the statistical significance of the male-female difference in frequent binge drinking.

- 8.72 **Students change their majors.** In a random sample of 950 students from a large public university, it was found that 444 of the students changed majors during their college years.
- Give a 99% confidence interval for the proportion of students at this university who change majors.
 - Express your results from (a) in terms of the *percent* of students who change majors.
 - University officials are more interested in the *number* of students who change majors than in the proportion. The university has 30,000 undergraduate students. Convert your confidence interval in (a) to a confidence interval for the number of students who change majors during their college years.
- 8.73 **Gender and top students.** Many colleges that once enrolled only male or only female students have become coeducational. Some administrators and alumni were concerned that the academic standards of the institutions would decrease with the change. One formerly all-male college undertook a study of the first class to contain women. The class consisted of 851 students, 214 of whom were women. An examination of first-semester grades revealed that 15 of the top 30 students were female.
- What is the proportion of women in the class? Call this value p_0 .
 - Are women more likely to be top students than their proportion in the class would suggest? Test the null hypothesis that the proportion of women in the top 30 students has the value p_0 . What do you conclude?

- 8.74 **Race and diet.** Seventh-Day Adventists avoid alcohol and tobacco, and the church recommends a vegetarian diet. Adventists are therefore a useful group for studies of the effects of a meatless diet. A random sample of Seventh-Day Adventists were interviewed at a national convention for a study on blood pressure and diet.²⁸ Because blacks in the general population have higher average blood pressure than whites, the study took race into account. The 312 Adventists in the sample were categorized by race and whether or not they were vegetarians. Here are the data:

	Black	White
Vegetarian	42	135
Not vegetarian	47	88

Are the proportions of vegetarians the same among all black and white Seventh-Day Adventists who attended this meeting? Analyze the data, paying particular attention to this question. Summarize your analysis and conclusions. What can you infer about the proportions of vegetarians among black and white Seventh-Day Adventists in general? What about blacks and whites in general?

- 8.75 **Blood pressure and the risk of death.** In Example 8.8 (page 530) we discussed a study that examined the association between high blood pressure and increased risk of death from cardiovascular disease. There were 2676 men with low blood pressure and 3338 men with high blood pressure. In the low-blood-pressure group, 21 men died from cardiovascular disease; in the high-blood-pressure group, 55 died.
- Verify the calculations in Example 8.8 by computing the 95% confidence interval for the difference in proportions.
 - Do the study data confirm that death rates are higher among men with high blood pressure? State hypotheses, carry out a significance test, and give your conclusions.
- 8.76 **Aspirin and heart disease.** A large experiment evaluated the effects of aspirin on cardiovascular disease. The subjects were 5139 male British medical doctors. The doctors were randomly assigned to two groups. One group of 3429 doctors took one aspirin daily, and the other group did not take aspirin. After 6 years, there were 148 deaths from heart attack or stroke in the first group and 79 in the second group. A similar experiment used male American medical doctors as subjects. These doctors were also randomly assigned to one of two groups. The 11,037 doctors in the first group took one aspirin every other day, and the 11,034 doctors in the second group took no aspirin. After nearly 5 years, there were 104 deaths from heart attacks in the first group and 189 in the second.²⁹ Analyze the data from these two studies and summarize the results. How do the conclusions of the two studies differ, and why?
- 8.77 **Choosing sample sizes.** For a single proportion the margin of error of a confidence interval is largest for any given sample size n and confidence level C when $\tilde{p} = 0.5$. This led us to use $p^* = 0.5$ for planning purposes. A similar result is true for the two-sample problem. The margin of error

of the confidence interval for the difference between two proportions is largest when $\tilde{p}_1 = \tilde{p}_2 = 0.5$. Use these conservative values in the following calculations, and assume that the sample sizes n_1 and n_2 have the common value n . Calculate the margins of error of the 99% confidence intervals for the difference in two proportions for the following choices of n : 10, 30, 50, 100, 200, and 500. Present the results in a table or with a graph. Summarize your conclusions.

- 8.78 **Choosing sample sizes, continued.** As the previous problem noted, using the guessed value 0.5 for both \tilde{p}_1 and \tilde{p}_2 gives a conservative margin of error in confidence intervals for the difference between two population proportions. You are planning a survey and will calculate a 95% confidence interval for the difference in two proportions when the data are collected. You would like the margin of error of the interval to be less than or equal to 0.05. You will use the same sample size n for both populations.
- How large a value of n is needed?
 - Give a general formula for n in terms of the desired margin of error and the critical value z^* .
- 8.79 **Unequal sample sizes.** You are planning a survey in which a 90% confidence interval for the difference between two proportions will present the results. You will use the conservative guessed value 0.5 for \tilde{p}_1 and \tilde{p}_2 in your planning. You would like the margin of error of the confidence interval to be less than or equal to 0.1. It is very difficult to sample from the first population, so that it will be impossible for you to obtain more than 20 observations from this population. Taking $n_1 = 20$, can you find a value of n_2 that will guarantee the desired margin of error? If so, report the value; if not, explain why not.
- 8.80 **Evaluate a proposal.** You are asked to evaluate a proposal for an experiment on the effects of aspirin on cardiovascular disease similar to the experiments described in Exercise 8.76. The researchers will randomly assign subjects to a treatment group or to a control group. The proposed sample sizes are 200 for each group. Write a short evaluation of this proposal, using any relevant information from Exercise 8.76.
- 8.81 **Statistics and the law.** *Castaneda v. Partida* is an important court case in which statistical methods were used as part of a legal argument. When reviewing this case, the Supreme Court used the phrase “two or three standard deviations” as a criterion for statistical significance. This Supreme Court review has served as the basis for many subsequent applications of statistical methods in legal settings. (The two or three standard deviations referred to by the Court are values of the z statistic and correspond to P -values of approximately 0.05 and 0.0026.) In *Castaneda* the plaintiff alleged that the method for selecting juries in a county in Texas was biased against Mexican Americans.³⁰ For the period of time at issue, there were 181,535 persons eligible for jury duty, of whom 143,611 were Mexican Americans. Of the 870 people selected for jury duty, 339 were Mexican Americans.
- What proportion of eligible voters were Mexican Americans? Let this value be p_0 .

- (b) Let p be the probability that a randomly selected juror is a Mexican American. The null hypothesis to be tested is $H_0: p = p_0$. Find the value of \hat{p} for this problem, compute the z statistic, and find the P -value. What do you conclude? (A finding of statistical significance in this circumstance does not constitute a proof of discrimination. It can be used, however, to establish a *prima facie* case. The burden of proof then shifts to the defense.)
- (c) We can reformulate this exercise as a two-sample problem. Here we wish to compare the proportion of Mexican Americans among those selected as jurors with the proportion of Mexican Americans among those not selected as jurors. Let p_1 be the probability that a randomly selected juror is a Mexican American, and let p_2 be the probability that a randomly selected nonjuror is a Mexican American. Find the z statistic and its P -value. How do your answers compare with your results in (b)?

9.211



CHAPTER 8 CASE STUDY EXERCISES

CASE STUDY 8.1: Gender bias in textbooks. Exercise 8.44 (page 533) reports a study of gender bias in 10 syntax textbooks. Here are the counts of “girl,” “woman,” “boy,” and “man” for all of the texts. The data in Exercise 8.44 are for text number 6.

	Text number									
	1	2	3	4	5	6	7	8	9	10
Girl	2	5	25	11	2	48	38	5	48	13
Woman	3	2	31	65	1	12	2	13	24	5
Boy	7	18	14	19	12	52	70	6	128	32
Man	27	45	51	138	31	80	2	27	48	95

Analyze the data and write a report summarizing your conclusions. The researchers who conducted the study note that the authors of texts 8, 9, and 10 are women, while the other seven texts were written by men. Do you see any pattern that suggests that the gender of the author is associated with the results?

CASE STUDY 8.2: Sample size, P -value, and the margin of error. In this case study we examine the effects of the sample size on the significance test and the confidence interval for comparing two proportions. For each calculation, suppose that $\hat{p}_1 = 0.6$ and $\hat{p}_2 = 0.4$, and take n to be the common value of n_1 and n_2 . Use the z statistic to test $H_0: p_1 = p_2$ versus the alternative $H_a: p_1 \neq p_2$. Compute the statistic and the associated P -value for the following values of n : 15, 25, 50, 75, 100, and 500.

Summarize the results in a table and make a plot. Explain what you observe about the effect of the sample size on statistical significance when the sample proportions \hat{p}_1 and \hat{p}_2 are unchanged.

Now we will do similar calculations for the confidence interval. Here, we suppose that $\tilde{p}_1 = 0.6$ and $\tilde{p}_2 = 0.4$. Compute the margin of error for the 95% confidence interval for the difference in the two proportions for $n = 15, 25, 50, 75, 100$, and 500. Summarize and explain your results.

Prelude

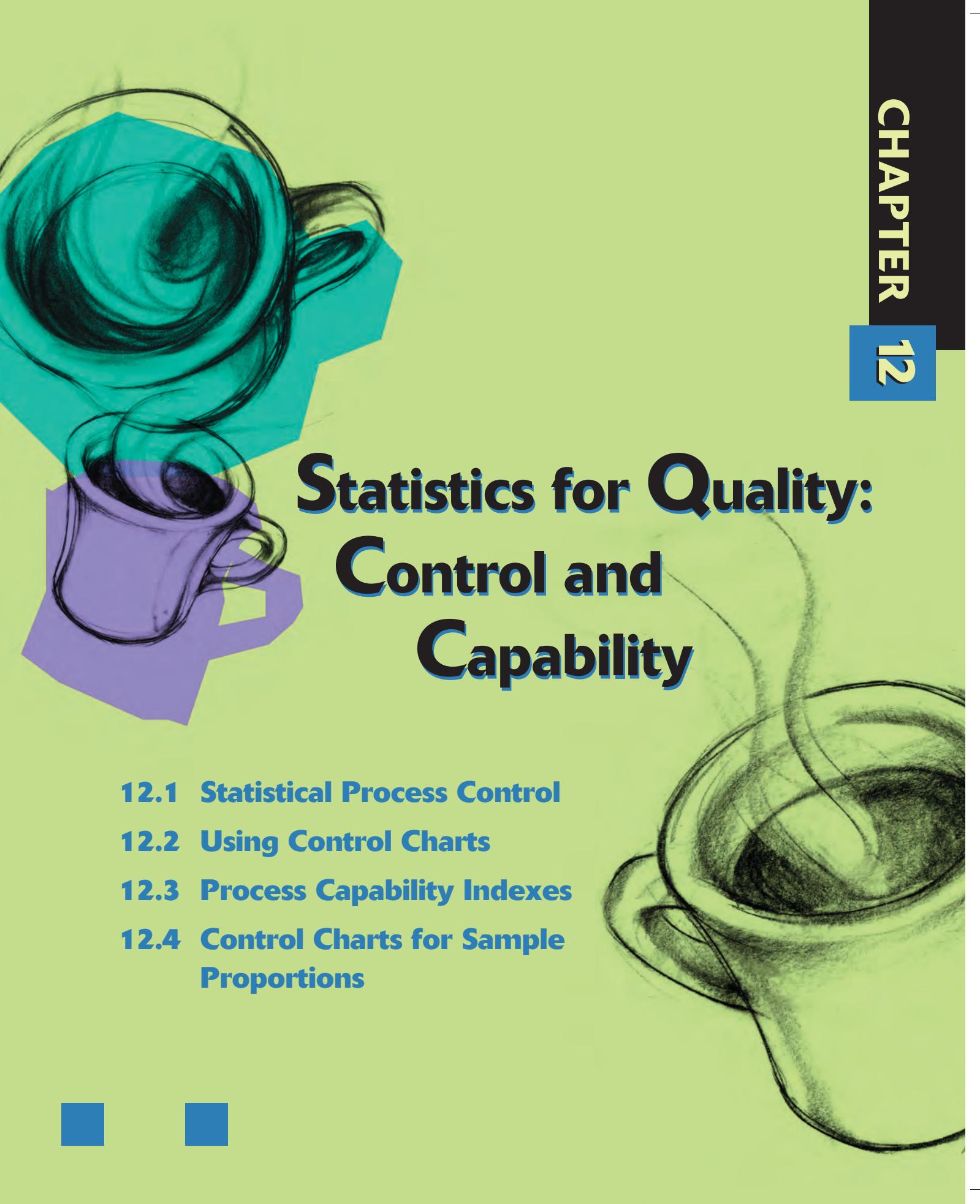
Reducing hospital losses

Both nonprofit and for-profit hospitals are financially pressed by restrictions on reimbursement by insurers and the government. Statistical analysis can help identify the sources of losses in a hospital's complex operations and perhaps suggest ways to reduce these losses.

Analysis at one hospital first looked at losses broken down by diagnosis. The leading source, accounting for 15% of the total dollar losses, was joint replacement surgery, mostly hip and knee replacement. *Control charts* based on hospital records showed that these losses were stable over time. That is, the losses are not due to unusual events and will continue at the same rate unless changes are made. Study of bills for joint replacement showed that charges for the actual surgery, for use of the operating room, and for the implant used in joint replacement made up most of the total.

The next step was to analyze costs broken down by the surgeon in charge. There was much variation among surgeons in two key areas. About one-third of surgeons regularly used more than the standard operating room time, thus incurring extra cost that is not reimbursed. The surgeons choose different brands of implants, some more expensive than others. If implants were standardized, the supplier would offer a lower price.

The analysis shows that much of the loss comes from *variation* in the joint replacement process. If the hospital can *reduce variation* by persuading doctors to agree on standard implants and by helping the slow surgeons improve their technique by learning from their peers, losses will come down with no change in patient satisfaction. This chapter presents statistical tools for monitoring the variation in processes of all kinds.¹



CHAPTER

12

Statistics for Quality: Control and Capability

- 12.1 Statistical Process Control**
- 12.2 Using Control Charts**
- 12.3 Process Capability Indexes**
- 12.4 Control Charts for Sample Proportions**

Introduction

Organizations are (or ought to be) concerned about the quality of the products and services they offer. What you don't know about quality can hurt you: customers often simply leave when they feel they are receiving poor quality, rather than make complaints that an alert organization could use as warnings. A key to maintaining and improving quality is systematic use of *data* in place of intuition or anecdotes.

EXAMPLE 12.1

Using data to improve quality

(a) Sometimes data that are routinely produced make a quality problem obvious. The internal financial statements of a professional society showed that hiring temporary employees to enter membership data was causing expenditures above budgeted levels each year during the several months when memberships were renewed. Investigation led to two actions. Membership renewal dates were staggered across the year to spread the workload more evenly. More important, outdated and inflexible data-entry software was replaced by a modern system that was much easier to use. Result: permanent employees can now process renewals quickly, eliminating the need for temps and also reducing member complaints.

(b) Systematic collection of data helps an organization to move beyond dealing with obvious problems. Motorola places a corporate-wide emphasis on measuring the performance of its services as well as of its manufactured products. Behind the counter of every local service center, for example, is a chart showing the average time from a customer's call until a service person reaches the customer's location, month by month. The trend should be steadily downward as ways are found to speed response.

Because using data is a key to improving quality, statistical methods have much to contribute. Simple tools are often the most effective. Motorola's service centers calculate mean response times each month and make a time plot (Chapter 1). A scatterplot and perhaps a regression line (Chapter 2) can show how the time to answer telephone calls to a corporate call center influences the percent of callers who hang up before their calls are answered. The design of a new product as simple as a multivitamin tablet may involve interviewing samples of consumers to learn what vitamins and minerals they want included and using randomized comparative experiments in designing the manufacturing process (Chapter 3). An experiment might discover, for example, what combination of moisture level in the raw vitamin powder and pressure in the tablet-forming press produces the right tablet hardness.

Quality is a broad concept. You may feel, for example, that a restaurant serving filet mignon is a higher-quality establishment than a fast-food outlet that serves hamburgers. For statistical purposes we need a narrower concept: consistently meeting standards appropriate for a specific product or service. The fast-food outlet, for example, may serve high-quality hamburgers while the expensive restaurant serves low-quality filet mignon. The hamburgers are freshly grilled, are served at the right temperature, and are the same

every time you visit. Statistically minded management can assess quality by sampling hamburgers and measuring the time from order to being served, the temperature of the burgers, and their tenderness.

This chapter focuses on just one aspect of statistics for improving quality: *statistical process control*. The techniques are simple and are based on sampling distributions (Chapter 4), but the underlying ideas are important and a bit subtle.

Processes

In thinking about statistical inference, we distinguish between the *sample* data we have in hand and the wider *population* that the data represent. We hope to use the sample to draw conclusions about the population. In thinking about quality improvement, it is often more natural to speak of *processes* rather than populations. This is because work is organized in processes. Some examples are:

- Processing an application for admission to a university and deciding whether or not to admit the student.
- Reviewing an employee's expense report for a business trip and issuing a reimbursement check.
- Hot forging to shape a billet of titanium into a blank that, after machining, will become part of a medical implant for hip, knee, or shoulder replacement.

Each of these processes is made up of several successive operations that eventually produce the output—an admission decision, reimbursement check, or metal component.

PROCESS

A process is a chain of activities that turns inputs into outputs.

We can accommodate processes in our sample-versus-population framework: think of the population as containing all the outputs that would be produced by the process if it ran forever in its present state. The outputs produced today or this week are a sample from this population. Because the population doesn't actually exist now, it is simpler to speak of a process and of recent output as a sample from the process in its present state.

Describing processes

The first step in improving a process is to understand it. If the process is at all complex, even the people involved with it may not have a full picture of how their work interacts with other steps in ways that influence process quality. A brainstorming session is in order: bring people together to gain an understanding of the process. This understanding is often presented graphically using two simple tools: flowcharts and cause-and-effect diagrams.

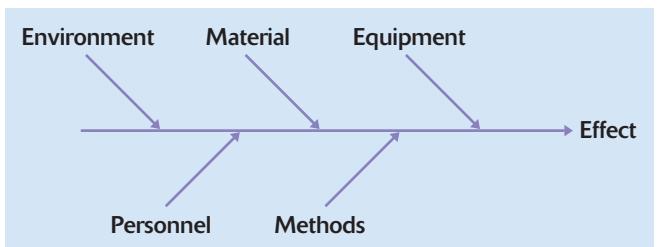


FIGURE 12.1 An outline for a cause-and-effect diagram. Group causes under these main headings in the form of branches.

flowchart

*cause-and-effect
diagram*

A **flowchart** is a picture of the stages of a process. Many organizations have formal standards for making flowcharts. Because flowcharts are not statistical graphics, we will informally illustrate their use in a case study and not insist on a specific format. A **cause-and-effect diagram** organizes the logical relationships between the inputs and stages of a process and an output. Sometimes the output is successful completion of the process task; sometimes it is a quality problem that we hope to solve. A good starting outline for a cause-and-effect diagram appears in Figure 12.1. The main branches organize the causes and serve as a skeleton for detailed entries. You can see why these are sometimes called “fishbone diagrams.” Once again we will illustrate the diagram by a case study rather than insist on a specific format.²



CASE 12.1 HOT FORGING

Hot forging involves heating metal to a plastic state and then shaping it by applying thousands of pounds of pressure to force the metal into a die (a kind of mold). Figure 12.2 is a flowchart of a typical hot-forging process.³

A process improvement team, after making and discussing this flowchart, came to several conclusions:

- Inspecting the billets of metal received from the supplier adds no value. We should insist that the supplier be responsible for the quality of the material. The supplier should put in place good statistical process control (see Section 12.1). We can then eliminate the inspection step.
- Can we buy the metal billets already cut to rough length and deburred by the supplier, thus eliminating the cost of preparing the raw material ourselves?
- Heating the metal billet and forging (pressing the hot metal into the die) are the heart of the process. We should concentrate our attention here.

The team then prepared a cause-and-effect diagram (Figure 12.3) for the heating and forging part of the process. The team members shared their specialist knowledge of the causes in their area, resulting in a more complete picture than any one person could produce. Figure 12.3 is a simplified

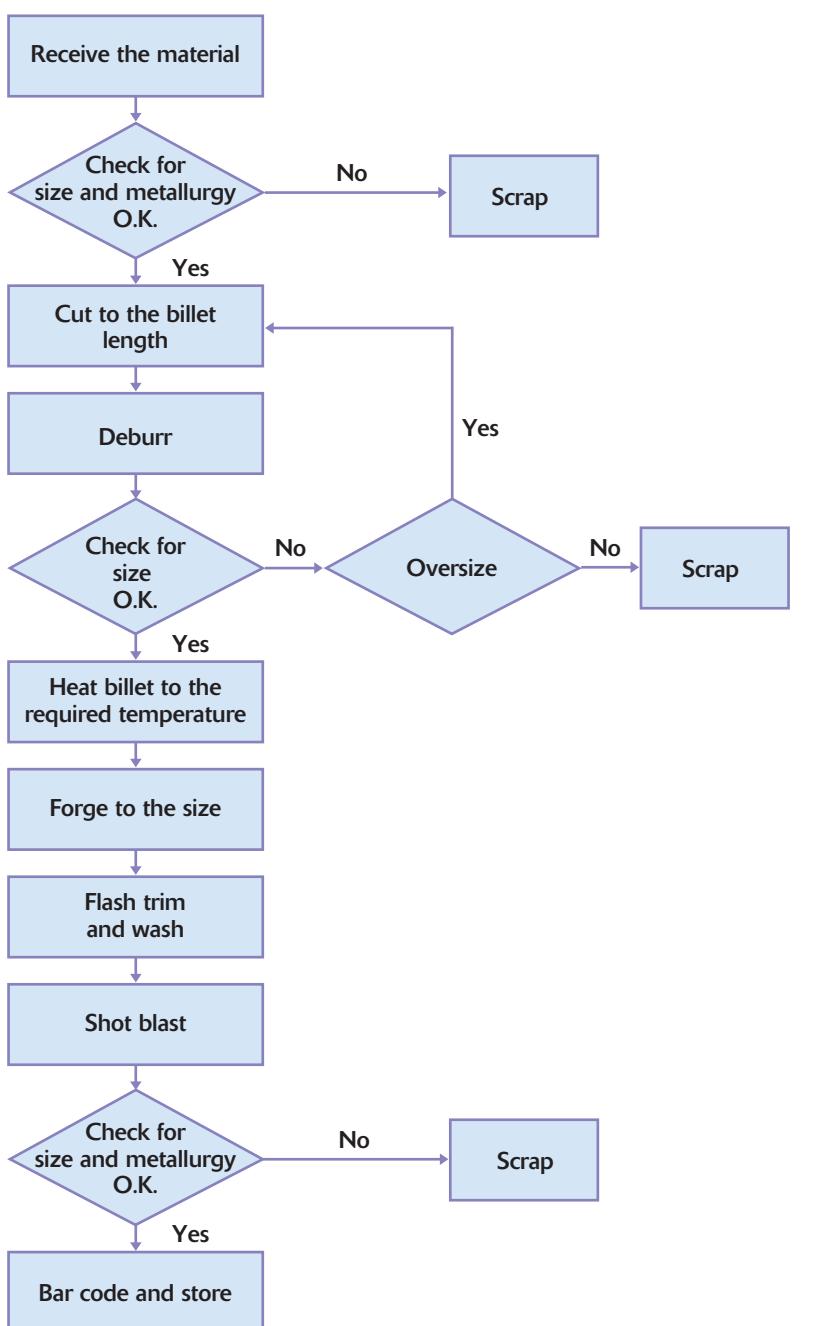


FIGURE 12.2 Flowchart of the hot-forging process in Case 12.1. Use this as a model for flowcharts: decision points appear as diamonds, and other steps in the process appear as rectangles. Arrows represent flow from step to step.

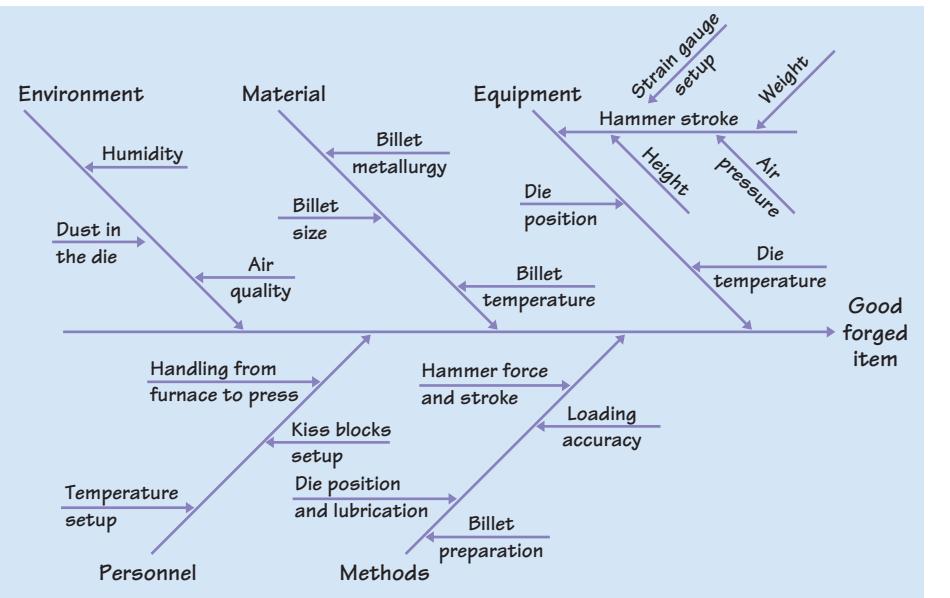


FIGURE 12.3 Simplified cause-and-effect diagram of the hot-forging process in Case 12.1. Good cause-and-effect diagrams require detailed knowledge of the specific process.

version of the actual diagram. We have given some added detail for the “hammer stroke” branch under “equipment” to illustrate the next level of branches. Even this requires some knowledge of hot forging to understand. Based on detailed discussion of the diagram, the team decided what variables to measure and at what stages of the process to measure them. Producing well-chosen data is the key to improving the process.



We will apply statistical methods to a series of measurements made on a process. Deciding what specific variables to measure is an important step in quality improvement. Often we use a “performance measure” that describes an output of a process. A company’s financial office might record the percent of errors that outside auditors find in expense account reports or the number of data-entry errors per week. The personnel department may measure the time to process employee insurance claims or the percent of job offers that are accepted. In the case of complex processes, it is wise to measure key steps within the process rather than just final outputs. The process team in Case 12.1 might recommend that the temperature of the die and of the billet be measured just before forging.

APPLY YOUR KNOWLEDGE

- 12.1 **Describe a process.** Choose a process that you know well. If you lack experience with actual business processes, choose a personal process such as cooking scrambled eggs or balancing your checkbook. Make a flowchart of the process. Make a cause-and-effect diagram that presents the factors that lead to successful completion of the process.

- 12.2 **Process measurement.** Based on your description of the process in Exercise 12.1, suggest specific variables that you might measure in order to
- Assess the overall quality of the process.
 - Gather information on a key step within the process.
- 12.3 **Pareto charts.** Pareto charts are bar graphs with the bars ordered by height. (See page 7 in Chapter 1.) They are often used to isolate the “vital few” categories on which we should focus our attention. Here is an example. The initial study of losses at the hospital described in the chapter Prelude looked at losses broken down by diagnosis. Government standards place cases into Diagnostic Related Groups (DRGs). For example, major joint replacements are DRG 209. Here is what the hospital found:

DRG	Percent of losses
104	5.2
107	10.1
109	7.7
116	13.7
148	6.8
209	15.2
403	5.6
430	6.8
462	9.4

What percent of total losses do these 9 DRGs account for? Make a Pareto chart of losses by DRG. Which DRGs should the hospital study first when attempting to reduce its losses?

12.1 Statistical Process Control

The goal of statistical process control is to make a process stable over time and then keep it stable unless planned changes are made. You might want, for example, to keep your weight constant over time. A manufacturer of machine parts wants the critical dimensions to be the same for all parts. “Constant over time” and “the same for all” are not realistic requirements. They ignore the fact that *all processes have variation*. Your weight fluctuates from day to day; the critical dimension of a machined part varies a bit from item to item; the time to process a college admission application is not the same for all applications. Variation occurs in even the most precisely made product due to small changes in the raw material, the adjustment of the machine, the behavior of the operator, and even the temperature in the plant. Because variation is always present, we can’t expect to hold a variable exactly constant over time. The statistical description of stability over time requires that the *pattern of variation* remain stable, not that there be no variation in the variable measured.

STATISTICAL CONTROL

A variable that continues to be described by the same distribution when observed over time is said to be in statistical control, or simply **in control**.

Control charts are statistical tools that monitor a process and alert us when the process has been disturbed so that it is now **out of control**. This is a signal to find and correct the cause of the disturbance.

common cause has only **common cause** variation. Common cause variation is the inherent variability of the system, due to many small causes that are always present. When the normal functioning of the process is disturbed by some unpredictable event, **special cause** variation is added to the common cause variation. We hope to be able to discover what lies behind special cause variation and eliminate that cause to restore the stable functioning of the process.

EXAMPLE 12.2

Common cause, special cause

Imagine yourself doing the same task repeatedly, say folding an advertising flyer, stuffing it into an envelope, and sealing the envelope. The time to complete the task will vary a bit, and it is hard to point to any one reason for the variation. Your completion time shows only common cause variation.

Now the telephone rings. You answer, and though you continue folding and stuffing while talking, your completion time rises beyond the level expected from common causes alone. Answering the telephone adds special cause variation to the common cause variation that is always present. The process has been disturbed and is no longer in its normal and stable state.

If you are paying temporary employees to fold and stuff advertising flyers, you avoid this special cause by not having telephones present and by asking the employees to turn off their cell phones while they are working.

Control charts work by distinguishing the always-present common cause variation in a process from the additional variation that suggests that the process has been disturbed by a special cause. A control chart sounds an alarm when it sees too much variation. The most common application of control charts is to monitor the performance of industrial and business processes. The same methods, however, can be used to check the stability of quantities as varied as the ratings of a television show, the level of ozone in the atmosphere, and the gas mileage of your car. Control charts combine graphical and numerical descriptions of data with use of sampling distributions.*

*Control charts were invented in the 1920s by Walter Shewhart at the Bell Telephone Laboratories. Shewhart's classic book, *Economic Control of Manufactured Product* (Van Nostrand, 1931), organized the application of statistics to improving quality.

APPLY YOUR KNOWLEDGE

- 12.4 **Special causes.** Jeannine participates in bicycle road races. She regularly rides 25 kilometers over the same course in training. Her time varies a bit from day to day but is generally stable. Give several examples of special causes that might raise Jeannine's time on a particular day.
- 12.5 **Common causes and special causes.** In Exercise 12.1, you described a process that you know well. What are some sources of common cause variation in this process? What are some special causes that might at times drive the process out of control?

 \bar{x} charts for process monitoring

When you first apply control charts to a process, the process may not be in control. Even if it is in control, you don't yet understand its behavior. You will have to collect data from the process, establish control by uncovering and removing special causes, and then set up control charts to maintain control. We call this the **chart setup** stage. Later, when the process has been operating in control for some time, you understand its usual behavior and have a long run of data from the process. You keep control charts to monitor the process because a special cause could erupt at any time. We will call this **process monitoring**.⁴

Although in practice chart setup precedes process monitoring, the big ideas of control charts are more easily understood in the process-monitoring setting. We will start there, then discuss the more complex chart setup setting.

Choose a quantitative variable x that is an important measure of quality. The variable might be the diameter of a part, the number of envelopes stuffed in an hour, or the time to respond to a customer call. Here are the conditions for process monitoring.

PROCESS-MONITORING CONDITIONS

Measure a quantitative variable x that has a **Normal distribution**. The process has been operating in control for a long period, so that we know the **process mean μ** and the **process standard deviation σ** that describe the distribution of x as long as the process remains in control.

In practice, we must of course estimate the process mean and standard deviation from past data on the process. Under the process-monitoring conditions, we have very many observations and the process has remained in control. The law of large numbers tells us that estimates from past data will be very close to the truth about the process. That is, at the process-monitoring stage we can act as if we know the true values of μ and σ . Note carefully that μ and σ describe the center and spread of our variable x *only as long as the process remains in control*. A special cause may at any time disturb the process and change the mean, the standard deviation, or both.

To make control charts, begin by taking small samples from the process at regular intervals. For example, we might measure 4 or 5 consecutive parts or time the responses to 4 or 5 consecutive customer calls.

*chart setup**process monitoring*

There is an important idea here: *the observations in a sample are so close together that we can assume that the process is stable during this short period of time.* Variation within the same sample gives us a benchmark for the common cause variation in the process. *The process standard deviation σ refers to the standard deviation within the time period spanned by one sample.* If the process remains in control, the same σ describes the standard deviation of observations across any time period. Control charts help us decide whether this is the case.

\bar{x} chart

We start with the \bar{x} chart based on plotting the means of the successive samples. Here is the outline:

1. Take samples of size n from the process at regular intervals. Plot the means \bar{x} of these samples against the order in which the samples were taken.
2. We know that the sampling distribution of \bar{x} under the process monitoring conditions is Normal with mean μ and standard deviation σ/\sqrt{n} (see page 291). Draw a solid center line on the chart at height μ .
3. The 99.7 part of the 68–95–99.7 rule for Normal distributions (page 57) says that, as long as the process remains in control, 99.7% of the values of \bar{x} will fall between $\mu - 3\sigma/\sqrt{n}$ and $\mu + 3\sigma/\sqrt{n}$. Draw dashed control limits on the chart at these heights. The control limits mark off the range of variation in sample means that we expect to see when the process remains in control.

If the process remains in control and the process mean and standard deviation do not change, we will rarely observe an \bar{x} outside the control limits. Such an \bar{x} is therefore a signal that the process has been disturbed.



CASE 12.2 MANUFACTURING COMPUTER MONITORS

A manufacturer of computer monitors must control the tension on the mesh of fine vertical wires that lies behind the surface of the viewing screen. Too much tension will tear the mesh, and too little will allow wrinkles. Tension is measured by an electrical device with output readings in millivolts (mV). The manufacturing process has been stable with mean tension $\mu = 275$ mV and process standard deviation $\sigma = 43$ mV.

The mean 275 mV and the common cause variation measured by the standard deviation 43 mV describe the stable state of the process. If these values are not satisfactory—for example, if there is too much variation among the monitors—the manufacturer must make some fundamental change in the process. This might involve buying new equipment or changing the alloy used in the wires of the mesh. One advantage of statistical control is that we know how the process will behave and can decide whether its behavior is satisfactory. In fact, the common cause variation in mesh tension does not affect the performance of the monitors. We want to watch the process and maintain its current condition.

TABLE 12.1 Twenty control chart samples of mesh tension

Sample	Tension measurements				Sample mean	Standard deviation
1	234.5	272.3	234.5	272.3	253.4	21.8
2	311.1	305.8	238.5	286.2	285.4	33.0
3	247.1	205.3	252.6	316.1	255.3	45.7
4	215.4	296.8	274.2	256.8	260.8	34.4
5	327.9	247.2	283.3	232.6	272.7	42.5
6	304.3	236.3	201.8	238.5	245.2	42.8
7	268.9	276.2	275.6	240.2	265.2	17.0
8	282.1	247.7	259.8	272.8	265.6	15.0
9	260.8	259.9	247.9	345.3	278.5	44.9
10	329.3	231.8	307.2	273.4	285.4	42.5
11	266.4	249.7	231.5	265.2	253.2	16.3
12	168.8	330.9	333.6	318.3	287.9	79.7
13	349.9	334.2	292.3	301.5	319.5	27.1
14	235.2	283.1	245.9	263.1	256.8	21.0
15	257.3	218.4	296.2	275.2	261.8	33.0
16	235.1	252.7	300.6	297.6	271.5	32.7
17	286.3	293.8	236.2	275.3	272.9	25.6
18	328.1	272.6	329.7	260.1	297.6	36.5
19	316.4	287.4	373.0	286.0	315.7	40.7
20	296.8	350.5	280.6	259.8	296.9	38.8

The operator measures the tension on a sample of 4 monitors each hour. Table 12.1 gives the last 20 samples. The table also gives the mean \bar{x} and the standard deviation s for each sample. The operator did not have to calculate these—modern measuring equipment often comes equipped with software that automatically records \bar{x} and s and even produces control charts.



Figure 12.4 is an \bar{x} control chart for the 20 mesh tension samples in Table 12.1. We have plotted each sample mean from the table against its sample number. For example, the mean of the first sample is 253.4 mV, and this is the value plotted for sample 1. The center line is at $\mu = 275$ mV. The upper and lower control limits are

$$\mu + 3\frac{\sigma}{\sqrt{n}} = 275 + 3\frac{43}{\sqrt{4}} = 275 + 64.5 = 339.5 \text{ mV} \quad (\text{UCL})$$

$$\mu - 3\frac{\sigma}{\sqrt{n}} = 275 - 3\frac{43}{\sqrt{4}} = 275 - 64.5 = 210.5 \text{ mV} \quad (\text{LCL})$$

As is common, we have labeled the control limits UCL for upper control limit and LCL for lower control limit.

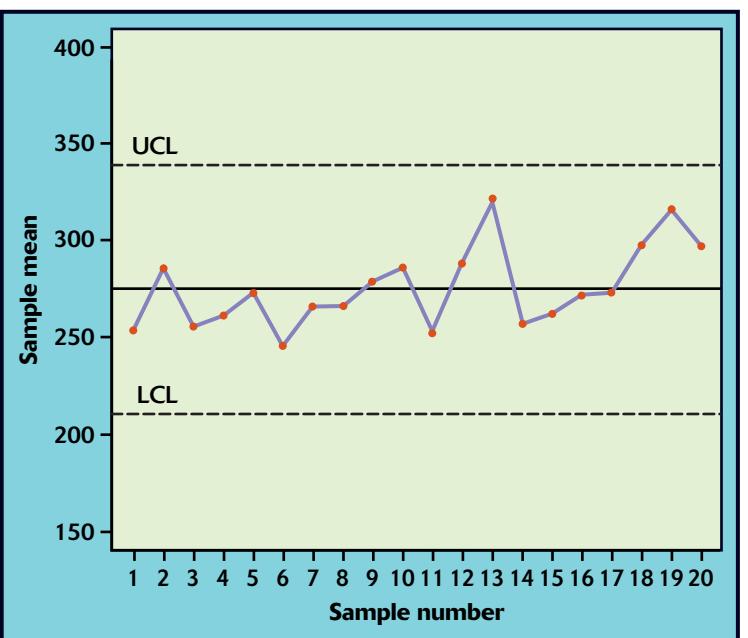


FIGURE 12.4 \bar{x} chart for the mesh tension data of Table 12.1. No points lie outside the control limits.

EXAMPLE 12.3 Interpreting \bar{x} charts



Figure 12.4 is a typical \bar{x} chart for a process in control. The means of the 20 samples do vary, but all lie within the range of variation marked out by the control limits. We are seeing the common cause variation of a stable process.

Figures 12.5 and 12.6 illustrate two ways in which the process can go out of control. In Figure 12.5, the process was disturbed by a special cause sometime between sample 12 and sample 13. As a result, the mean tension for sample 13 falls above the upper control limit. It is common practice to mark all out-of-control points with an “x” to call attention to them. A search for the cause begins as soon as we see a point out of control. Investigation finds that the mounting of the tension-measuring device has slipped, resulting in readings that are too high. When the problem is corrected, samples 14 to 20 are again in control.

Figure 12.6 shows the effect of a steady upward drift in the process center, starting at sample 11. You see that some time elapses before the \bar{x} for sample 18 is out of control. The one-point-out signal works better for detecting sudden large disturbances than for detecting slow drifts in a process.

-
- 12.6 **Auto thermostats.** A maker of auto air conditioners checks a sample of 4 thermostatic controls from each hour's production. The thermostats are set at 75° F and then placed in a chamber where the temperature is raised gradually. The temperature at which the thermostat turns on the air conditioner is recorded. The process mean should be $\mu = 75^\circ$. Past experience indicates that the response temperature of properly adjusted thermostats varies with $\sigma = 0.5^\circ$. The mean response temperature \bar{x} for

APPLY YOUR KNOWLEDGE

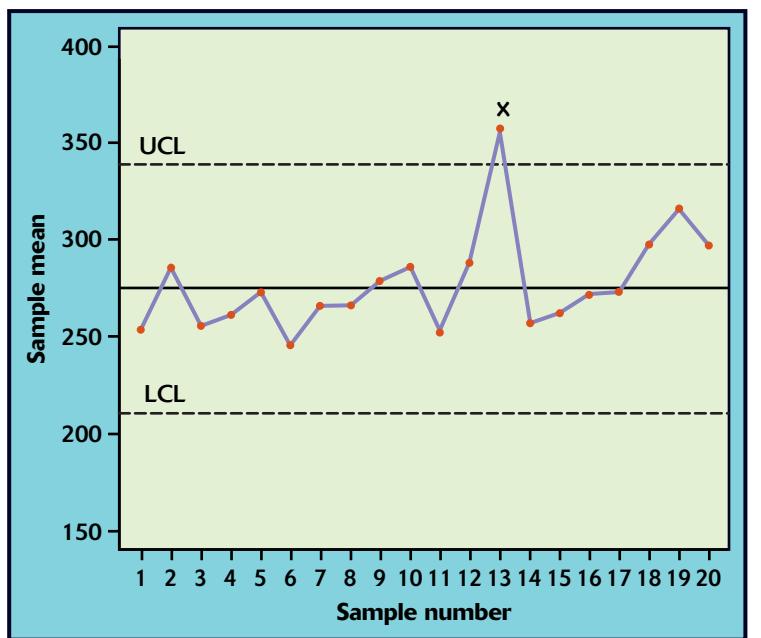


FIGURE 12.5 This \bar{x} chart is identical to that in Figure 12.4, except that a special cause has driven \bar{x} for sample 13 above the upper control limit. The out-of-control point is marked with an x.

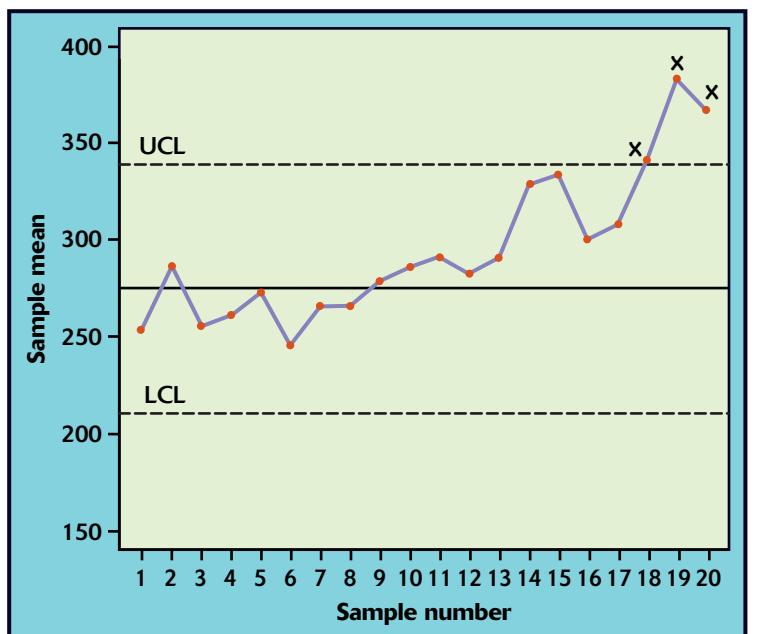


FIGURE 12.6 The first 10 points on this \bar{x} chart are as in Figure 12.4. The process mean drifts upward after sample 10, and the sample means \bar{x} reflect this drift. The points for samples 18, 19, and 20 are out of control.

each hour's sample is plotted on an \bar{x} control chart. Calculate the center line and control limits for this chart.

- 12.7 **Tablet hardness.** A pharmaceutical manufacturer forms tablets by compressing a granular material that contains the active ingredient and various fillers. The hardness of a sample from each lot of tablets is measured in order to control the compression process. The process has been operating in control with mean at the target value $\mu = 11.5$ and estimated standard deviation $\sigma = 0.2$. Table 12.2 gives three sets of data, each representing \bar{x} for 20 successive samples of $n = 4$ tablets. One set remains in control at the target value. In a second set, the process mean μ shifts suddenly to a new value. In a third, the process mean drifts gradually.
- What are the center line and control limits for an \bar{x} chart for this process?
 - Draw a separate \bar{x} chart for each of the three data sets. Mark any points that are beyond the control limits.
 - Based on your work in (b) and the appearance of the control charts, which set of data comes from a process that is in control? In which case does the process mean shift suddenly and at about which sample do you think that the mean changed? Finally, in which case does the mean drift gradually?

Sample	Data set A	Data set B	Data set C
1	11.602	11.627	11.495
2	11.547	11.613	11.475
3	11.312	11.493	11.465
4	11.449	11.602	11.497
5	11.401	11.360	11.573
6	11.608	11.374	11.563
7	11.471	11.592	11.321
8	11.453	11.458	11.533
9	11.446	11.552	11.486
10	11.522	11.463	11.502
11	11.664	11.383	11.534
12	11.823	11.715	11.624
13	11.629	11.485	11.629
14	11.602	11.509	11.575
15	11.756	11.429	11.730
16	11.707	11.477	11.680
17	11.612	11.570	11.729
18	11.628	11.623	11.704
19	11.603	11.472	12.052
20	11.816	11.531	11.905

s charts for process monitoring

The \bar{x} charts in Figures 12.4, 12.5, and 12.6 were easy to interpret because the process standard deviation remained fixed at 43 mV. The effects of moving the process mean away from its in-control value (275 mV) are then clear to see. We know that even the simplest description of a distribution should give both a measure of center and a measure of spread. So it is with control charts. We must monitor both the process center, using an \bar{x} chart, and the process spread, using a control chart for the sample standard deviation s .

The standard deviation s does not have a Normal distribution, even approximately. Under the process monitoring conditions, the sampling distribution of s is skewed to the right. Nonetheless, control charts for any statistic are based on the “plus or minus three standard deviations” idea motivated by the 68–95–99.7 rule for Normal distributions. Control charts are intended to be practical tools that are easy to use. Standard practice in process control therefore ignores such details as the effect of non-Normal sampling distributions. Here is the general control chart setup for a sample statistic Q (short for “quality characteristic”).

THREE-SIGMA CONTROL CHARTS

To make a **three-sigma (3σ) control chart** for any statistic Q :

1. Take samples from the process at regular intervals and plot the values of the statistic Q against the order in which the samples were taken.
2. Draw a **center line** on the chart at height μ_Q , the mean of the statistic when the process is in control.
3. Draw upper and lower **control limits** on the chart three standard deviations of Q above and below the mean. That is,

$$\text{UCL} = \mu_Q + 3\sigma_Q$$

$$\text{LCL} = \mu_Q - 3\sigma_Q$$

Here σ_Q is the standard deviation of the sampling distribution of the statistic Q when the process is in control.

4. The chart produces an **out-of-control signal** when a plotted point lies outside the control limits.

We have applied this general idea to \bar{x} charts. If μ and σ are the process mean and standard deviation, the statistic \bar{x} has mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. The center line and control limits for \bar{x} charts follow from these facts.

What are the corresponding facts for the sample standard deviation s ? Study of the sampling distribution of s for samples from a Normally distributed process characteristic gives these facts:

1. The *mean* of s is a constant times the process standard deviation σ , $\mu_s = c_4\sigma$.
2. The *standard deviation* of s is also a constant times the process standard deviation, $\sigma_s = c_5\sigma$.

The constants are called c_4 and c_5 for historical reasons. Their values depend on the size of the samples. For large samples, c_4 is close to 1. That is, the sample standard deviation s has little bias as an estimator of the process standard deviation σ . Because statistical process control often uses small samples, we pay attention to the value of c_4 . Following the general pattern for three-sigma control charts:

1. The *center line* of an s chart is at $c_4\sigma$.
2. The *control limits* for an s chart are at

$$\begin{aligned} \text{UCL} &= \mu_s + 3\sigma_s = c_4\sigma + 3c_5\sigma = (c_4 + 3c_5)\sigma \\ \text{LCL} &= \mu_s - 3\sigma_s = c_4\sigma - 3c_5\sigma = (c_4 - 3c_5)\sigma \end{aligned}$$

That is, the control limits UCL and LCL are also constants times the process standard deviation. These constants are called (again for historical reasons) B_6 and B_5 . We don't need to remember that $B_6 = c_4 + 3c_5$ and $B_5 = c_4 - 3c_5$, because tables give us the numerical values of B_6 and B_5 .

\bar{x} AND s CONTROL CHARTS FOR PROCESS MONITORING⁵

Take regular samples of size n from a process that has been in control with process mean μ and process standard deviation σ . The center line and control limits for an \bar{x} chart are

$$\begin{aligned} \text{UCL} &= \mu + 3\frac{\sigma}{\sqrt{n}} \\ \text{CL} &= \mu \\ \text{LCL} &= \mu - 3\frac{\sigma}{\sqrt{n}} \end{aligned}$$

The center line and control limits for an s chart are

$$\begin{aligned} \text{UCL} &= B_6\sigma \\ \text{CL} &= c_4\sigma \\ \text{LCL} &= B_5\sigma \end{aligned}$$

The control chart constants c_4 , B_5 , and B_6 depend on the sample size n .

TABLE 12.3 Control chart constants

Sample size n	c_4	c_5	B_5	B_6
2	0.7979	0.6028		2.606
3	0.8862	0.4633		2.276
4	0.9213	0.3889		2.088
5	0.9400	0.3412		1.964
6	0.9515	0.3076	0.029	1.874
7	0.9594	0.2820	0.113	1.806
8	0.9650	0.2622	0.179	1.751
9	0.9693	0.2459	0.232	1.707
10	0.9727	0.2321	0.276	1.669

Table 12.3 gives the values of the control chart constants c_4 , c_5 , B_5 , and B_6 for samples of sizes 2 to 10. This table makes it easy to draw s charts. The table has no B_5 entries for samples of size smaller than $n = 6$. The lower control limit for an s chart is zero for samples of sizes 2 to 5. This is a consequence of the fact that s has a right-skewed distribution and takes only values greater than zero. Three standard deviations above the mean (UCL) lies on the long right side of the distribution. Three standard deviations below the mean (LCL) on the short left side is below zero, so we say that $LCL = 0$.

EXAMPLE 12.4 **\bar{x} and s charts for mesh tension**

Figure 12.7 is the s chart for the computer monitor mesh tension data in Table 12.1. The samples are of size $n = 4$ and the process standard deviation in control is $\sigma = 43$ mV. The center line is therefore

$$CL = c_4\sigma = (0.9213)(43) = 39.6 \text{ mV}$$

The control limits are

$$UCL = B_6\sigma = (2.088)(43) = 89.8$$

$$LCL = B_5\sigma = (0)(43) = 0$$

Figures 12.4 and 12.7 go together: they are \bar{x} and s charts for monitoring the mesh-tensioning process. Both charts are in control, showing only common cause variation within the bounds set by the control limits.

Figures 12.8 and 12.9 are \bar{x} and s charts for the mesh-tensioning process when a new and poorly trained operator takes over between samples 10 and 11. The new operator introduces added variation into the process, increasing the process standard deviation from its in-control value of 43 mV to 60 mV. The \bar{x} chart in Figure 12.8 shows one point out of control. Only on closer inspection do we see that the spread of the \bar{x} 's increases after sample 10. In fact, the process mean has remained unchanged at 275 mV. The apparent lack of control in the \bar{x} chart is

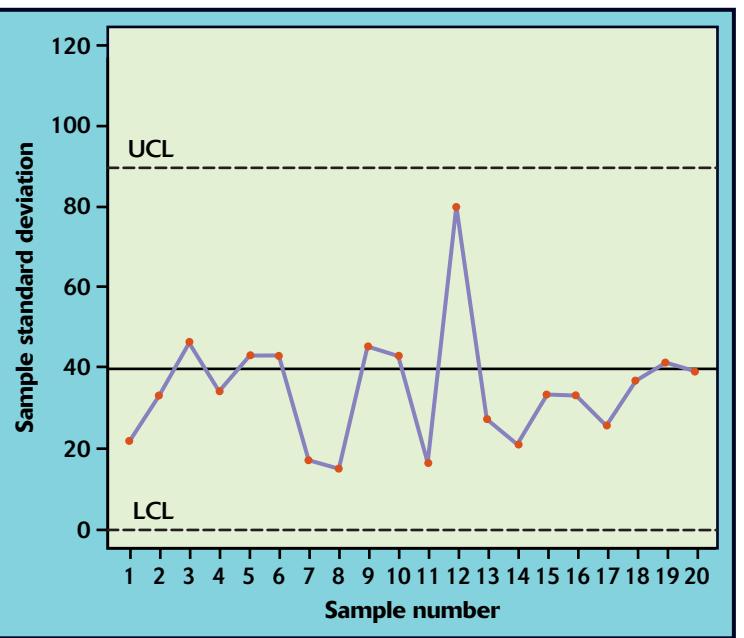


FIGURE 12.7 s chart for the mesh tension data of Table 12.1. Both the s chart and the \bar{x} chart (Figure 12.4) are in control.

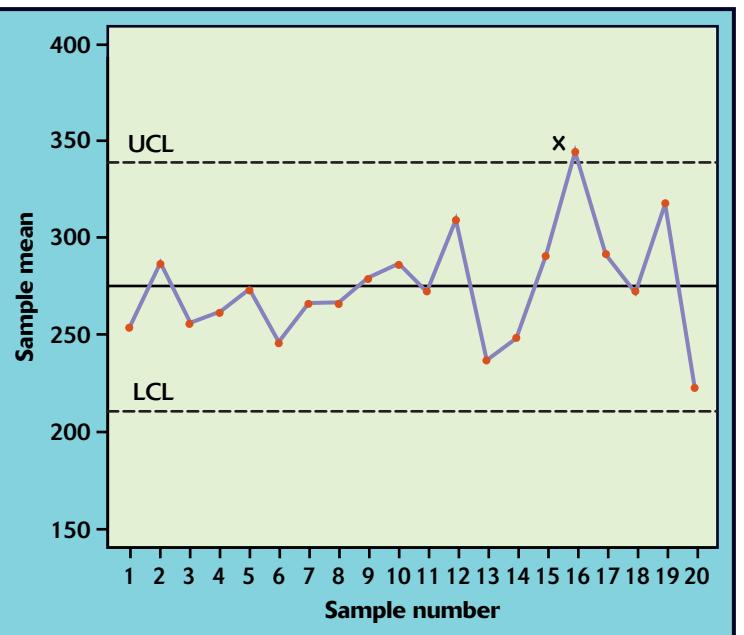


FIGURE 12.8 \bar{x} chart for mesh tension when the process variability increases after sample 10. The \bar{x} chart does show the increased variability, but the s chart is clearer and should be read first.

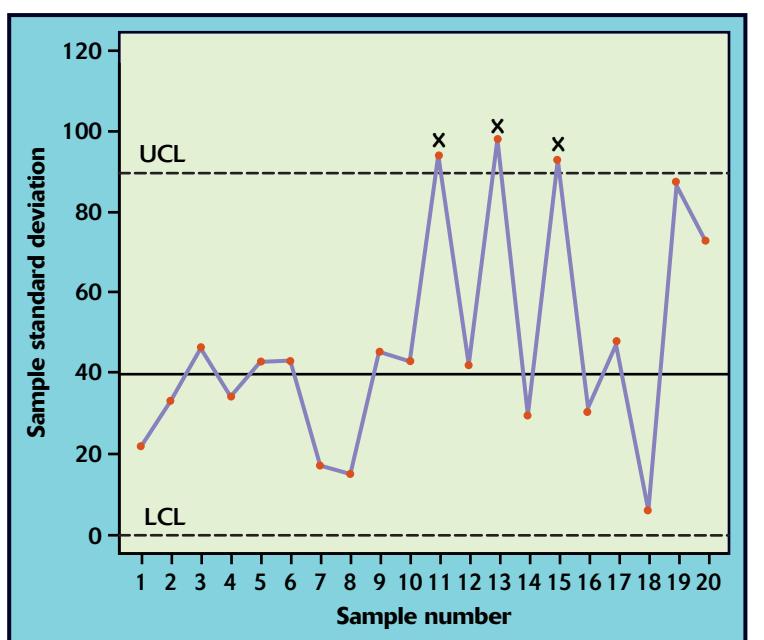


FIGURE 12.9 *s* chart for mesh tension when the process variability increases after sample 10. Increased within-sample variability is clearly visible. Find and remove the *s*-type special cause before reading the \bar{x} chart.

entirely due to the larger process variation. There is a lesson here: *it is difficult to interpret an \bar{x} chart unless *s* is in control*. When you look at \bar{x} and *s* charts, always start with the *s* chart.

The *s* chart in Figure 12.9 shows lack of control starting at sample 11. As usual, we mark the out-of-control points by an “x.” The points for samples 13 and 15 also lie above the UCL, and the overall spread of the sample points is much greater than for the first 10 samples. In practice, the *s* chart would call for action after sample 11. We would ignore the \bar{x} chart until the special cause (the new operator) for the lack of control in the *s* chart has been found and removed by training the operator.

Example 12.4 suggests a strategy for using \bar{x} and *s* charts in practice. First examine the *s* chart. Lack of control on an *s* chart is due to special causes that affect the observations *within a sample* differently. New and nonuniform raw material, a new and poorly trained operator, and mixing results from several machines or several operators are typical “*s*-type” special causes.

Once the *s* chart is in control, the stable value of the process standard deviation σ means that the variation within samples serves as a benchmark for detecting variation in the level of the process over the longer time periods between samples. The \bar{x} chart, with control limits that depend on σ , does this. The \bar{x} chart, as we saw in Example 12.4, responds to *s*-type causes as well as to longer-range changes in the process, so it is important to eliminate *s*-type special causes first. Then the \bar{x} chart will alert us to, for example, a

change in process level caused by new raw material that differs from that used in the past or a gradual drift in the process level caused by wear in a cutting tool.

EXAMPLE 12.5***s*-type and \bar{x} -type special causes**

A large health maintenance organization (HMO) uses control charts to monitor the process of directing patient calls to the proper department or doctor's receptionist. Each day at a random time, 5 consecutive calls are recorded electronically. The first call today is handled quickly by an experienced operator, but the next goes to a newly hired operator who must ask a supervisor for help. The sample has a large s , and lack of control signals the need to train new hires more thoroughly.

The same HMO monitors the time required to receive orders from its main supplier of pharmaceutical products. After a long period in control, the \bar{x} chart shows a systematic shift downward in the mean time because the supplier has changed to a more efficient delivery service. This is a desirable special cause, but it is nonetheless a systematic change in the process. The HMO will have to establish new control limits that describe the new state of the process, with smaller process mean μ .

The second setting in Example 12.5 reminds us that a major change in the process returns us to the chart setup stage. In the absence of deliberate changes in the process, process monitoring uses the same values of μ and σ for long periods of time. One exception is common: careful monitoring and removal of special causes as they occur can permanently reduce the process σ . If the points on the s chart remain near the center line for a long period, it is wise to update the value of σ to the new, smaller value.

APPLY YOUR KNOWLEDGE

- 12.8 **Responding to applicants.** The personnel department of a large company records a number of performance measures. Among them is the time required to respond to an application for employment, measured from the time the application arrives. Suggest some plausible examples of each of the following.
 - (a) Reasons for common cause variation in response time.
 - (b) s -type special causes.
 - (c) \bar{x} -type special causes.
- 12.9 **Auto thermostats.** In Exercise 12.6 you gave the center line and control limits for an \bar{x} chart. What are the center line and control limits for an s chart for this process?
- 12.10 **Tablet hardness.** Exercise 12.7 concerns process control data on the hardness of tablets for a pharmaceutical product. Table 12.4 gives data for 20 new samples of size 4, with the \bar{x} and s for each sample. The process has been in control with mean at the target value $\mu = 11.5$ and standard deviation $\sigma = 0.2$.
 - (a) Make both \bar{x} and s charts for these data based on the information given about the process.

TABLE 12.4 Twenty samples of size 4, with \bar{x} and s

Sample	Hardness				\bar{x}	s
1	11.432	11.350	11.582	11.184	11.387	0.1660
2	11.791	11.323	11.734	11.512	11.590	0.2149
3	11.373	11.807	11.651	11.651	11.620	0.1806
4	11.787	11.585	11.386	11.245	11.501	0.2364
5	11.633	11.212	11.568	11.469	11.470	0.1851
6	11.648	11.653	11.618	11.314	11.558	0.1636
7	11.456	11.270	11.817	11.402	11.486	0.2339
8	11.394	11.754	11.867	11.003	11.504	0.3905
9	11.349	11.764	11.402	12.085	11.650	0.3437
10	11.478	11.761	11.907	12.091	11.809	0.2588
11	11.657	12.524	11.468	10.946	11.649	0.6564
12	11.820	11.872	11.829	11.344	11.716	0.2492
13	12.187	11.647	11.751	12.026	11.903	0.2479
14	11.478	11.222	11.609	11.271	11.395	0.1807
15	11.750	11.520	11.389	11.803	11.616	0.1947
16	12.137	12.056	11.255	11.497	11.736	0.4288
17	12.055	11.730	11.856	11.357	11.750	0.2939
18	12.107	11.624	11.727	12.207	11.916	0.2841
19	11.933	10.658	11.708	11.278	11.394	0.5610
20	12.512	12.315	11.671	11.296	11.948	0.5641

- (b) At some point, the within-sample process variation increased from $\sigma = 0.2$ to $\sigma = 0.4$. About where in the 20 samples did this happen? What is the effect on the s chart? On the \bar{x} chart?
- (c) At that same point, the process mean changed from $\mu = 11.5$ to $\mu = 11.7$. What is the effect of this change on the s chart? On the \bar{x} chart?

SECTION 12.1 SUMMARY

- Work is organized in processes, chains of activities that lead to some result. We use flowcharts and cause-and-effect diagrams to describe processes.
- All processes have variation. If the pattern of variation is stable over time, the process is **in statistical control**. Control charts are statistical plots intended to warn when a process is **out of control**.
- Standard 3σ control charts plot the values of some statistic Q for regular samples from the process against the time order of the samples. The **center line** is at the mean of Q . The **control limits** lie three standard deviations of Q above and below the center line. A point outside the control limits is an **out-of-control signal**. For process monitoring of a

process that has been in control, the mean and standard deviation are based on past data from the process and are updated regularly.

- When we measure some quantitative characteristic of the process, we use \bar{x} and s charts for process control. The s chart monitors variation within individual samples. If the s chart is in control, the \bar{x} chart monitors variation from sample to sample. To interpret the charts, always look first at the s chart.

SECTION 12.1 EXERCISES

- Describe a process.** Each weekday morning, you must get to work or to your first class on time. Make a flowchart of your daily process for doing this, starting when you wake. Be sure to include the time at which you plan to start each step.
- Common cause, special cause.** Each weekday morning, you must get to work or to your first class on time. The time at which you reach work or class varies from day to day, and your planning must allow for this variation. List several common causes of variation in your arrival time. Then list several special causes that might result in unusual variation, such as being late to work or class.
- Pareto charts.** Continue the study of the process of getting to work or class on time. If you kept good records, you could make a Pareto chart of the reasons (special causes) for late arrivals at work or class. Make a Pareto chart that you think roughly describes your own reasons for lateness. That is, list the reasons from your experience and chart your estimates of the percent of late arrivals each reason explains.
- Pareto charts.** Painting new auto bodies is a multistep process. There is an “electrocoat” that resists corrosion, a primer, a color coat, and a gloss coat. A quality study for one paint shop produced this breakdown of the primary problem type for those autos whose paint did not meet the manufacturer’s standards:

Problem	Percent
Electrocoat uneven—redone	4
Poor adherence of color to primer	5
Lack of clarity in color	2
“Orange peel” texture in color	32
“Orange peel” texture in gloss	1
Ripples in color coat	28
Ripples in gloss coat	4
Uneven color thickness	19
Uneven gloss thickness	5
Total	100

Make a Pareto chart. Which stage of the painting process should we look at first?

- 12.15 Milling.** The width of a slot cut by a milling machine is important to the proper functioning of a hydraulic system for large tractors. The manufacturer checks the control of the milling process by measuring a sample of 5 consecutive items during each hour's production. The target width for the slot is $\mu = 0.8750$ inch. The process has been operating in control with center close to the target and $\sigma = 0.0012$ inch. What center line and control limits should be drawn on the s chart? On the \bar{x} chart?
- 12.16 Dyeing yarn.** The unique colors of the cashmere sweaters your firm makes result from heating undyed yarn in a kettle with a dye liquor. The pH (acidity) of the liquor is critical for regulating dye uptake and hence the final color. There are 5 kettles, all of which receive dye liquor from a common source. Twice each day, the pH of the liquor in each kettle is measured, giving a sample of size 5. The process has been operating in control with $\mu = 4.22$ and $\sigma = 0.127$.
- Give the center line and control limits for the s chart.
 - Give the center line and control limits for the \bar{x} chart.
- 12.17 Mounting-hole distances.** Figure 12.10 reproduces a data sheet from the floor of a factory that makes electrical meters.⁶ The sheet shows measurements on the distance between two mounting holes for 18 samples of size 5. The heading informs us that the measurements are in multiples of 0.0001 inch above 0.6000 inch. That is, the first measurement, 44, stands for 0.6044 inch. All the measurements end in 4. Although we don't know why this is true, it is clear that in effect the measurements were made to the nearest 0.001 inch, not to the nearest 0.0001 inch.
- Calculate \bar{x} and s for the first two samples. The data file *ex12-17.dat* contains \bar{x} and s for all 18 samples. Based on long experience with this process, you are keeping control charts based on $\mu = 43$ and $\sigma = 12.74$. Make s and \bar{x} charts for the data in Figure 12.10 and describe the state of the process.

VARIABLES CONTROL CHART (\bar{X} & R)																		Part No.	Chart No.
Part name (project) Metal frame				Operation (process) Distance between mounting holes										Specification limits $0.6054'' \pm 0.0010''$					
Operator		Machine R-5		Gage		Unit of measure 0.0001"				Zero equals 0.6000"									
Date	3/7	3/8				3/9													
Time	8:30	10:30	11:45	1:30	8:15	10:15	11:45	2:00	3:00	4:00	8:30	10:00	11:45	1:30	2:30	3:30	4:30	5:30	
Sample measurements	1	44	64	34	44	34	34	54	64	24	34	34	54	44	24	54	54	54	
	2	44	44	44	54	14	64	64	34	54	44	44	44	24	24	24	34	34	24
	3	44	34	54	54	84	34	34	54	44	44	34	24	34	54	54	24	74	64
	4	44	34	44	34	54	44	44	44	34	34	64	54	34	44	44	44	44	34
	5	64	54	54	44	44	34	44	34	34	34	24	44	44	44	54	54	44	
Average, \bar{X}																			
Range, R	20	30	20	20	70	30	30	30	30	10	30	30	20	30	40	30	40	40	

FIGURE 12.10 A process control record sheet kept by operators. This is typical of records kept by hand when measurements are not automated. We will see in the next section why such records mention \bar{X} and R control charts rather than \bar{x} and s charts.

- 12.18 **Dyeing yarn: special causes.** The process described in Exercise 12.16 goes out of control. Investigation finds that a new type of yarn was recently introduced. The pH in the kettles is influenced by both the dye liquor and the yarn. Moreover, on a few occasions a faulty valve on one of the kettles had allowed water to enter that kettle; as a result, the yarn in that kettle had to be discarded.
- Which of these special causes appears on the s chart and which on the \bar{x} chart? Explain your answer.
- 12.19 **Probability out?** An \bar{x} chart plots the means of samples of size 4 against center line $CL = 700$ and control limits $LCL = 687$ and $UCL = 713$. The process has been in control. Now the process is disrupted in a way that changes the mean to $\mu = 693$ and the standard deviation to $\sigma = 12$. What is the probability that the first sample after the disruption gives a point beyond the control limits of the \bar{x} chart?
- 12.20 **Alternative control limits.** American and Japanese practice uses 3σ control charts. That is, the control limits are three standard deviations on either side of the mean. When the statistic being plotted has a Normal distribution, the probability of a point outside the limits is about 0.003 (or about 0.0015 in each direction) by the 68–95–99.7 rule. European practice uses control limits placed so that the probability of a point out is 0.001 in each direction. For a Normally distributed statistic, how many standard deviations on either side of the mean do these alternative control limits lie?
- 12.21 **2σ control charts.** Some special situations call for 2σ control charts. That is, the control limits for a statistic Q will be $\mu_Q \pm 2\sigma_Q$. Suppose that you know the process mean μ and standard deviation σ and will plot \bar{x} and s from samples of size n .
- What are the 2σ control limits for an \bar{x} chart?
 - Find expressions for the upper and lower 2σ control limits for an s chart in terms of the control chart constants c_4 and c_5 introduced on pages 12–18.

12.2 Using Control Charts

We are now familiar with the ideas that undergird all control charts and also with the details of making \bar{x} and s charts. This section discusses a variety of topics related to using control charts in practice.

\bar{x} and R charts

We have seen that it is essential to monitor both the center and the spread of a process. Control charts were originally intended to be used by factory workers with limited knowledge of statistics in the era before even calculators, let alone software, were common. In that environment, the standard deviation is too difficult to calculate. The \bar{x} chart for center was therefore combined with a control chart for spread based on the **sample range** rather than the sample standard deviation. The range R of a sample is just the difference between the largest and smallest observations. It is easy to find R without a calculator. Using R rather than s to measure the spread

sample range

R chart

of samples replaces the s chart with an R chart. It also changes the \bar{x} chart because the control limits for \bar{x} use the estimated process spread. So \bar{x} and R charts differ in the details of both charts from \bar{x} and s charts.

Because the range R uses only the largest and smallest observations in a sample, it is less informative than the standard deviation s calculated from all the observations. For this reason, \bar{x} and s charts are now preferred to \bar{x} and R charts. R charts remain common because tradition dies hard and also because it is easier for workers to understand R than s . In this short introduction, we concentrate on the principles of control charts, so we won't give the details of constructing \bar{x} and R charts. These details appear in any text on quality control.⁷ If you meet a set of \bar{x} and R charts, remember that the interpretation of these charts is just like the interpretation of \bar{x} and s charts.

EXAMPLE 12.6 **\bar{x} and R charts for pen diameter**

Figure 12.11 is a display produced by custom process control software attached to a laser micrometer. In this demonstration prepared by the software maker, the micrometer is measuring the diameter in millimeters of samples of pens shipped by an office supply company. The software controls the laser, records measurements, makes the control charts, and sounds an alarm when a point is out of control. This is typical of process control technology in modern manufacturing settings.

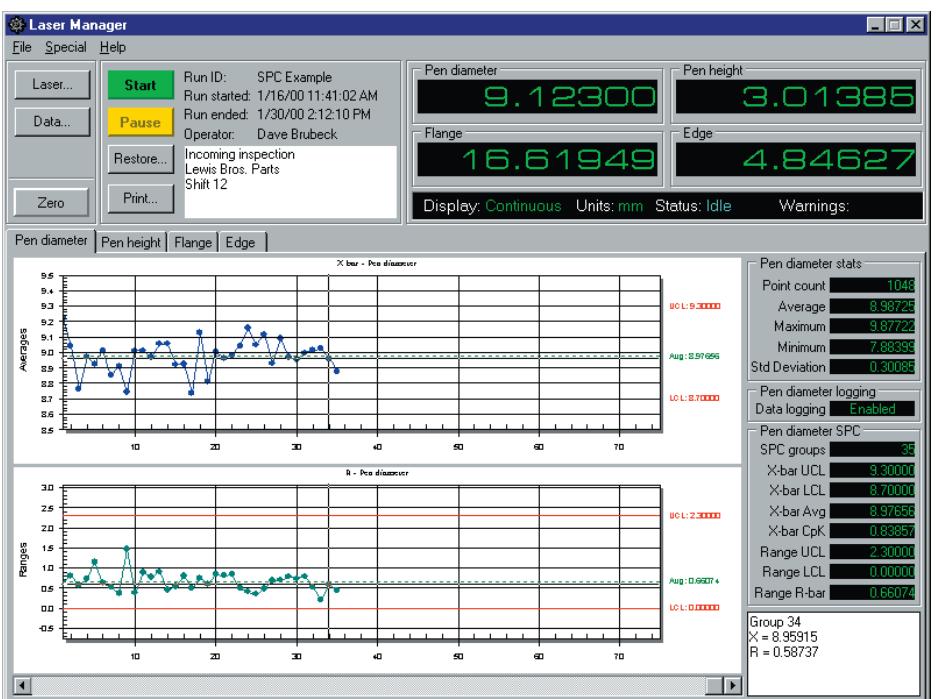


FIGURE 12.11 Output for operators from the Laser Manager software by System Dynamics, Inc. The software prepares control charts directly from measurements made by a laser micrometer. Compare the hand record sheet in Figure 12.10. (Image provided by Gordon A. Feingold, System Dynamics, Inc. Used by permission.)

Despite the advanced technology involved, the software presents \bar{x} and R charts rather than \bar{x} and s charts, no doubt because R is easier to explain. The R chart monitors within-sample variation (just like an s chart), so we look at it first. We see that the process spread is stable and well within the control limits. Just as in the case of s , the LCL for R is 0 for the samples of size $n = 5$ used here. The \bar{x} chart is also in control, so process monitoring will continue. The software will sound an alarm if either chart goes out of control.



Additional out-of-control signals

So far, we have used only the basic “one point beyond the control limits” criterion to signal that a process may have gone out of control. We would like a quick signal when the process moves out of control, but we also want to avoid “false alarms,” signals that occur just by chance when the process is really in control. The standard 3σ control limits are chosen to prevent too many false alarms, because an out-of-control signal calls for an effort to find and remove a special cause. As a result, \bar{x} charts are often slow to respond to a gradual drift in the process center that continues for some time before finally forcing a reading outside the control limits. We can speed the response of a control chart to lack of control—at the cost of also enduring more false alarms—by adding patterns other than “one-point-out” as signals. The most common step in this direction is to add a *runs signal* to the \bar{x} chart.

OUT-OF-CONTROL SIGNALS

- \bar{x} and s or \bar{x} and R control charts produce an out-of-control signal if:
- One-point-out:** A single point lies outside the 3σ control limits of either chart.
 - Run:** The \bar{x} chart shows 9 consecutive points above the center line or 9 consecutive points below the center line. The signal occurs when we see the 9th point of the run.

EXAMPLE 12.7

Using the runs signal

Figure 12.12 reproduces the \bar{x} chart from Figure 12.6. The process center began a gradual upward drift at sample 11. The chart shows the effect of the drift—the sample means plotted on the chart move gradually upward, with some random variation. The one-point-out signal does not call for action until sample 18 finally produces an \bar{x} above the UCL. The runs signal reacts more quickly: sample 17 is the 9th consecutive point above the center line.



It is a mathematical fact that the runs signal responds to a gradual drift more quickly (on the average) than the one-point-out signal does. The motivation for a runs signal is that when a process is in control, half the points on an \bar{x} chart should lie above the center line and half below. That's true on the average in the long term. In the short term, we will see runs of points above or below, just as we see runs of heads or tails in tossing a coin.

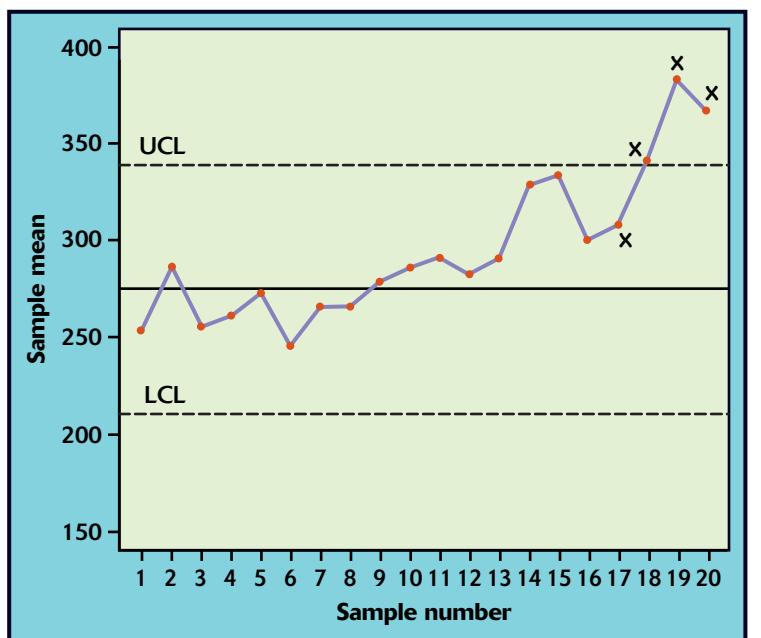


FIGURE 12.12 \bar{x} chart for mesh tension data when the process center drifts upward. The “run of 9” signal gives an out-of-control warning at sample 17.

How long a run suggests that the process center has moved, so that points are now more likely to fall on one side of the center line? We must once more worry about the cost of false alarms, so we choose a signal for which these are about as frequent as for the familiar one-point-out signal. The 99.7 part of the 68–95–99.7 rule says that we will get a point outside the 3σ control limits about 3 times for every 1000 points plotted when the process is in control. The chance of 9 straight points above the center line when the process is in control is $(1/2)^9 = 1/512$, or about 2 per 1000. The chance for a run of 9 below the center line is the same. That’s about 4 false alarms per 1000 plotted points overall, close to the false-alarm rate for one-point-out.

There are many other signals that can be added to the rules for responding to \bar{x} and s or \bar{x} and R charts. In our enthusiasm to detect various special kinds of loss of control, it is easy to forget that adding signals always increases the frequency of false alarms. Frequent false alarms are so annoying that the people responsible for responding soon begin to ignore out-of-control signals. It is better to use only a few signals and to reserve signals other than one-point-out and runs for processes that are known to be prone to specific special causes for which there is a tailor-made signal.⁸

APPLY YOUR KNOWLEDGE

- 12.22 Special causes. Is each of the following examples of a special cause most likely to first result in (i) one-point-out on the s or R chart, (ii) one-point-out on the \bar{x} chart, or (iii) a run on the \bar{x} chart? In each case, briefly explain your reasoning.
- An etching solution deteriorates as more items are etched.
 - Buildup of dirt reduces the precision with which parts are placed for machining.

- (c) A new customer service representative for a Spanish-language help line is not a native speaker and has difficulty understanding customers.
 - (d) A data-entry employee grows less attentive as her shift continues.
- 12.23 **Mixtures.** Here is an artificial situation that illustrates an unusual control chart pattern. Invoices are processed and paid by two clerks, one very experienced and the other newly hired. The experienced clerk processes invoices quickly. The new hire must often refer to a handbook and is much slower. Both are quite consistent, so that their times vary little from invoice to invoice. It happens that each sample of invoices comes from one of the clerks, so that some samples are from one and some from the other clerk. Sketch the \bar{x} chart pattern that will result.

Setting up control charts

When you first approach a process that has not been carefully studied, it is quite likely that the process is not in control. Your first goal is to discover and remove special causes and so bring the process into control. Control charts are an important tool. Control charts for *process monitoring* follow the process forward in time to keep it in control. Control charts at the *chart setup* stage, on the other hand, look back in an attempt to discover the present state of the process. A case study will illustrate the method.



VISCOSITY OF AN ELASTOMER

The viscosity of a material is its resistance to flow when under stress. Viscosity is a critical characteristic of rubber and rubber-like compounds called elastomers, which have many uses in consumer products. Viscosity is measured by placing specimens of the material above and below a slowly rotating roller, squeezing the assembly, and recording the drag on the roller. Measurements are in “Mooney units,” named after the inventor of the instrument.

A specialty chemical company is beginning production of an elastomer that is supposed to have viscosity 45 ± 5 Mooneys. Each lot of the elastomer is produced by “cooking” raw material with catalysts in a reactor vessel. Table 12.5 records \bar{x} and s from samples of size $n = 4$ lots from the first 24 shifts as production begins.⁹ An s chart therefore monitors variation among lots produced during the same shift. If the s chart is in control, an \bar{x} chart looks for shift-to-shift variation.



Estimating μ . We do not know the process mean μ and standard deviation σ . What shall we do? Sometimes we can easily adjust the center of a process by setting some control, such as the depth of a cutting tool in a machining operation or the temperature of a reactor vessel in a pharmaceutical plant. In such cases it is common to simply take the process mean μ to be the target value, the depth or temperature that the design of the process specifies as correct. The \bar{x} chart then helps us keep the process mean at this target value.

TABLE 12.5 \bar{x} and s for 24 samples of elastomer viscosity

Sample	\bar{x}	s	Sample	\bar{x}	s
1	49.750	2.684	13	47.875	1.118
2	49.375	0.895	14	48.250	0.895
3	50.250	0.895	15	47.625	0.671
4	49.875	1.118	16	47.375	0.671
5	47.250	0.671	17	50.250	1.566
6	45.000	2.684	18	47.000	0.895
7	48.375	0.671	19	47.000	0.447
8	48.500	0.447	20	49.625	1.118
9	48.500	0.447	21	49.875	0.447
10	46.250	1.566	22	47.625	1.118
11	49.000	0.895	23	49.750	0.671
12	48.125	0.671	24	48.625	0.895

There is less likely to be a “correct value” for the process mean μ if we are monitoring response times to customer calls or data entry errors. In Case 12.3, we have the target value 45 Mooneys, but there is no simple way to set viscosity at the desired level. In such cases, we want the μ we use in our \bar{x} chart to describe the center of the process as it has actually been operating. To do this, just take the mean of all the individual measurements in the past samples. Because the samples are all the same size, this is just the mean of the sample \bar{x} 's. The overall “mean of the sample means” is therefore usually called $\bar{\bar{x}}$. For the 24 samples in Table 12.5,

$$\begin{aligned}\bar{\bar{x}} &= \frac{1}{24}(49.750 + 49.375 + \cdots + 48.625) \\ &= \frac{1161.125}{24} = 48.380\end{aligned}$$

Estimating σ . It is almost never safe to use a “target value” for the process standard deviation σ because it is almost never possible to directly adjust process variation. We must estimate σ from past data. We want to combine the sample standard deviations s from past samples rather than use the standard deviation of all the individual observations in those samples. That is, in Case 12.3, we want to combine the 24 sample standard deviations in Table 12.5 rather than calculate the standard deviation of the 96 observations in these samples. The reason is that it is the *within-sample* variation that is the benchmark against which we compare the longer-term process variation. Even if the process has been in control, we want only the variation over the short time period of a single sample to influence our value for σ .

There are several ways to estimate σ from the sample standard deviations. In practice, software may use a somewhat sophisticated method and then calculate the control limits for you. We use a simple method that is traditional in quality control because it goes back to the era before software. If we are

basing chart setup on k past samples, we have k sample standard deviations s_1, s_2, \dots, s_k . Just average these to get

$$\bar{s} = \frac{1}{k}(s_1 + s_2 + \dots + s_k)$$

For the viscosity case, we average the s -values for the 24 samples in Table 12.5:

$$\begin{aligned}\bar{s} &= \frac{1}{24}(2.684 + 0.895 + \dots + 0.895) \\ &= \frac{24.156}{24} = 1.0065\end{aligned}$$

Combining the sample s -values to estimate σ introduces a complication: the samples used in process control are often small (size $n = 4$ in the viscosity case), so s has some bias as an estimator of σ . Recall that $\mu_s = c_4\sigma$. The mean \bar{s} inherits this bias: its mean is also not σ but $c_4\sigma$. The proper estimate of σ corrects this bias. It is

$$\hat{\sigma} = \frac{\bar{s}}{c_4}$$

We get control limits from past data by using the estimates \bar{x} and $\hat{\sigma}$ in place of the μ and σ used in charts at the process-monitoring stage. Here are the results.¹⁰

\bar{x} AND s CONTROL CHARTS USING PAST DATA

Take regular samples of size n from a process. Estimate the process mean μ and the process standard deviation σ from past samples by

$$\begin{aligned}\hat{\mu} &= \bar{x} && \text{(or use a target value)} \\ \hat{\sigma} &= \frac{\bar{s}}{c_4}\end{aligned}$$

The center line and control limits for an \bar{x} chart are

$$\text{UCL} = \hat{\mu} + 3\frac{\hat{\sigma}}{\sqrt{n}}$$

$$\text{CL} = \hat{\mu}$$

$$\text{LCL} = \hat{\mu} - 3\frac{\hat{\sigma}}{\sqrt{n}}$$

The center line and control limits for an s chart are

$$\text{UCL} = B_6\hat{\sigma}$$

$$\text{CL} = c_4\hat{\sigma} = \bar{s}$$

$$\text{LCL} = B_5\hat{\sigma}$$

If the process was not in control when the samples were taken, these should be regarded as trial control limits.

We are now ready to outline the chart setup procedure for elastomer viscosity.

Step 1. As usual, we look first at an s chart. For chart setup, control limits are based on the same past data that we will plot on the chart. Calculate from Table 12.5 that

$$\bar{s} = 1.0065$$

$$\hat{\sigma} = \frac{\bar{s}}{c_4} = \frac{1.0065}{0.9213} = 1.0925$$

The center line and control limits for an s chart based on past data are

$$UCL = B_6 \hat{\sigma} = (2.088)(1.0925) = 2.281$$

$$CL = \bar{s} = 1.0065$$

$$LCL = B_5 \hat{\sigma} = (0)(1.0925) = 0$$

Figure 12.13 is the s chart. The points for shifts 1 and 6 lie above the UCL. Both are near the beginning of production. Investigation finds that the reactor operator made an error on one lot in each of these samples. The error changed the viscosity of that lot and increased s for that one sample. The error will not be repeated now that the operators have gained experience. That is, this special cause has already been removed.

Step 2. Remove the two values of s that were out of control. This is proper because the special cause responsible for these readings is no

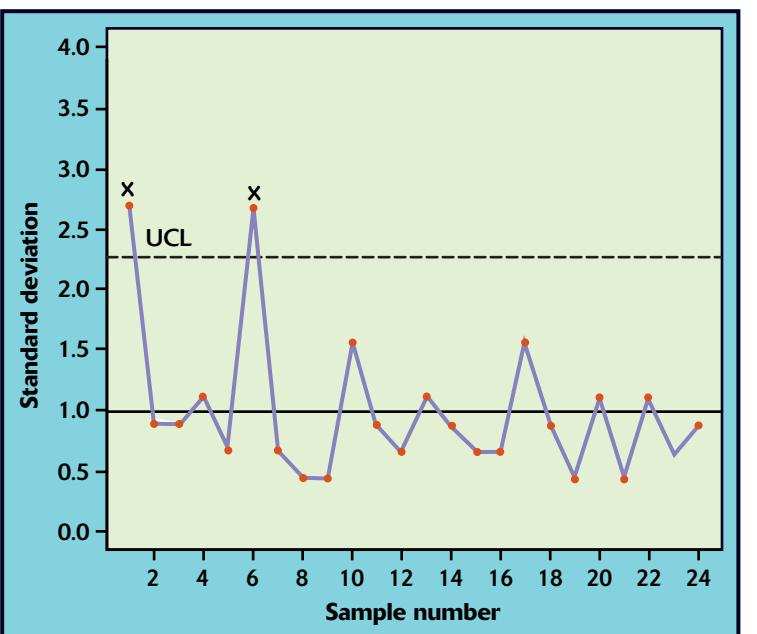


FIGURE 12.13 s chart based on past data for the viscosity data of Table 12.5. The control limits are based on the same s -values that are plotted on the chart. Points 1 and 6 are out of control.

longer present. Recalculate from the remaining 22 shifts that $\bar{s} = 0.854$ and $\hat{\sigma} = 0.854/0.9213 = 0.927$. Make a new s chart with

$$\text{UCL} = B_6 \hat{\sigma} = (2.088)(0.927) = 1.936$$

$$\text{CL} = \bar{s} = 0.854$$

$$\text{LCL} = B_5 \hat{\sigma} = (0)(0.927) = 0$$

We don't show the chart, but you can see from Table 12.5 that none of the remaining s -values lies above the new, lower, UCL; the largest remaining s is 1.566. If additional points were now out of control, we would repeat the process of finding and eliminating s -type causes until the s chart for the remaining shifts is in control. In practice, of course, this is often a challenging task.

Step 3. Once s -type causes have been eliminated, make an \bar{x} chart *using only the samples that remain* after dropping those that had out-of-control s -values. For the 22 remaining samples, we know that $\hat{\sigma} = 0.927$ and we calculate that $\bar{\bar{x}} = 48.4716$. The center line and control limits for the \bar{x} chart are

$$\text{UCL} = \bar{\bar{x}} + 3 \frac{\hat{\sigma}}{\sqrt{n}} = 48.4716 + 3 \frac{0.927}{\sqrt{4}} = 49.862$$

$$\text{CL} = \bar{\bar{x}} = 48.4716$$

$$\text{LCL} = \bar{\bar{x}} - 3 \frac{\hat{\sigma}}{\sqrt{n}} = 48.4716 - 3 \frac{0.927}{\sqrt{4}} = 47.081$$

Figure 12.14 is the \bar{x} chart. Shifts 1 and 6 have been dropped. Seven of the 22 points are beyond the 3σ limits, four high and three low. Although within-shift variation is now stable, there is excessive variation from shift to shift. To find the cause, we must understand the details of the process, but knowing that the special cause or causes operate between shifts is a big help. If the reactor is set up anew at the beginning of each shift, that's one place to look more closely.

Step 4. Once the \bar{x} and s charts are both in control (looking backward), use the estimates $\hat{\mu}$ and $\hat{\sigma}$ from the points in control to set tentative control limits to monitor the process going forward. If it remains in control, we can update the charts and move to the process-monitoring stage.

APPLY YOUR KNOWLEDGE

- 12.24 **From setup to monitoring.** Suppose that when the chart setup project of Case 12.3 is complete, the points remaining after removing special causes have $\bar{\bar{x}} = 48.7$ and $\bar{s} = 0.92$. What are the center line and control limits for the \bar{x} and s charts you would use to monitor the process going forward? 
- 12.25 **Estimating process parameters.** The \bar{x} and s control charts for the mesh-tensioning example (Figures 12.4 and 12.7) were based on $\mu = 275$ mV and $\sigma = 43$ mV. Table 12.1 gives the 20 most recent samples from this process.
- Estimate the process μ and σ based on these 20 samples.
 - Your calculations suggest that the process σ may now be less than 43 mV. Explain why the s chart in Figure 12.7 (page 12-20) suggests the same conclusion. (If this pattern continues, we would eventually update the value of σ used for control limits.)

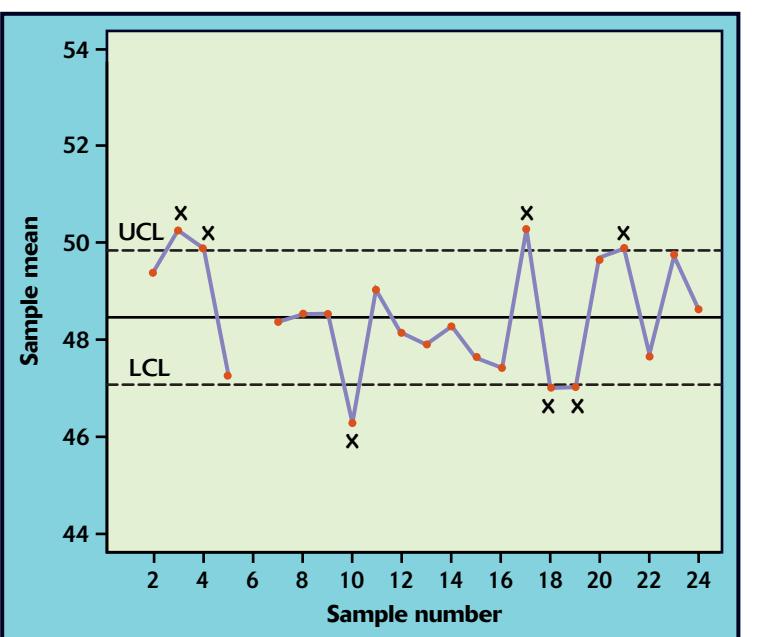


FIGURE 12.14 \bar{x} chart based on past data for the viscosity data of Table 12.5. The samples for shifts 1 and 6 have been removed because s -type special causes active in those samples are no longer active. The \bar{x} chart shows poor control.

- 12.26 Hospital losses. Table 12.6 gives data on the losses (in dollars) incurred by a hospital in treating DRG 209 (major joint replacement) patients.¹¹ The hospital has taken from its records a random sample of 8 such patients each month for 15 months.
- Make an s control chart using center lines and limits calculated from these past data. There are no points out of control.
 - Because the s chart is in control, base the \bar{x} chart on all 15 samples. Make this chart. Is it also in control?

Comments on statistical control

Having seen how \bar{x} and s (or \bar{x} and R) charts work, we can turn to some important comments and cautions about statistical control in practice.

Focus on the process rather than on the products. This is perhaps the fundamental idea in statistical process control. We might attempt to attain high quality by careful inspection of the finished product, measuring every completed forging and reviewing every outgoing invoice and expense account payment. Inspection of finished products can ensure good quality, but it is expensive. Perhaps more important, final inspection comes too late: when something goes wrong early in a process, much bad product may be produced before final inspection discovers the problem. This adds to the expense, because the bad product must then be scrapped or reworked.

The small samples that are the basis of control charts are intended to monitor the process at key points, not to ensure the quality of the particular

TABLE 12.6 Hospital losses for 15 samples of DRG 209 patients

Sample	Loss (dollars)									Sample mean	Standard deviation
1	6835	5843	6019	6731	6362	5696	7193	6206	6360.6	521.7	
2	6452	6764	7083	7352	5239	6911	7479	5549	6603.6	817.1	
3	7205	6374	6198	6170	6482	4763	7125	6241	6319.8	749.1	
4	6021	6347	7210	6384	6807	5711	7952	6023	6556.9	736.5	
5	7000	6495	6893	6127	7417	7044	6159	6091	6653.2	503.7	
6	7783	6224	5051	7288	6584	7521	6146	5129	6465.8	1034.3	
7	8794	6279	6877	5807	6076	6392	7429	5220	6609.2	1104.0	
8	4727	8117	6586	6225	6150	7386	5674	6740	6450.6	1033.0	
9	5408	7452	6686	6428	6425	7380	5789	6264	6479.0	704.7	
10	5598	7489	6186	5837	6769	5471	5658	6393	6175.1	690.5	
11	6559	5855	4928	5897	7532	5663	4746	7879	6132.4	1128.6	
12	6824	7320	5331	6204	6027	5987	6033	6177	6237.9	596.6	
13	6503	8213	5417	6360	6711	6907	6625	7888	6828.0	879.8	
14	5622	6321	6325	6634	5075	6209	4832	6386	5925.5	667.8	
15	6269	6756	7653	6065	5835	7337	6615	8181	6838.9	819.5	

items in the samples. If the process is kept in control, we know what to expect in the finished product. We want to do it right the first time, not inspect and fix finished product. Choosing the “key points” at which we will measure and monitor the process is important. The choice requires that you understand the process well enough to know where problems are likely to arise. Flowcharts and cause-and-effect diagrams can help. It should be clear that control charts that monitor only the final output are often *not* the best choice.

Rational subgroups. The interpretation of control charts depends on the distinction between \bar{x} -type special causes and s -type special causes. This distinction in turn depends on how we choose the samples from which we calculate s (or R). We want the variation *within* a sample to reflect only the item-to-item chance variation that (when in control) results from many small common causes. Walter Shewhart, the founder of statistical process control, used the term **rational subgroup** to emphasize that we should think about the process when deciding how to choose samples.

rational subgroup

EXAMPLE 12.8 Random sampling versus rational subgroups

A pharmaceutical manufacturer forms tablets by compressing a granular material that contains the active ingredient and various fillers. To monitor the compression process, we will measure the hardness of a sample from each 10 minutes’ production of tablets. Should we choose a random sample of tablets from the several thousand produced in a 10-minute period?

A random sample would contain tablets spread across the entire 10 minutes. It fairly represents the 10-minute period, but that isn’t what we want for process

control. If the setting of the press drifts or a new lot of filler arrives during the 10 minutes, the spread of the sample will be increased. That is, a random sample contains both the short-term variation among tablets produced in quick succession and the longer-term variation among tablets produced minutes apart. We prefer to measure a rational subgroup of 5 consecutive tablets every 10 minutes. We expect the process to be stable during this very short time period, so that variation within the subgroups is a benchmark against which we can see special cause variation.



Samples of consecutive items are rational subgroups when we are monitoring the output of a single activity that does the same thing over and over again. Several consecutive items is the most common type of sample for process control. When the stream of product contains output from several machines or several people, however, the choice of samples is more complicated. Do you want to include variation due to different machines or different people within your samples? If you decide that this variation is common cause variation that is acceptable for this process, be sure that the sample items are spread across machines or people. If all the items in each sample have a common origin, \bar{s} will be small and the control limits for the \bar{x} chart will be narrow. Points on the \bar{x} chart from samples representing different machines or different people will often be out of control, some high and some low.

There is no formula for choosing samples that are rational subgroups. You must think about causes of variation in your process and decide which you are willing to think of as common causes that you will not try to eliminate. Rational subgroups are samples chosen to express variation due to these causes and no others. Because the choice requires detailed process knowledge, we will usually accept samples of consecutive items as being rational subgroups. Just remember that real processes are messier than textbooks suggest.

Why statistical control is desirable. To repeat, if the process is kept in control, we know what to expect in the finished product. The process mean μ and standard deviation σ remain stable over time, so (assuming Normal variation) the 99.7 part of the 68–95–99.7 rule tells us that almost all measurements on individual products will lie in the range $\mu \pm 3\sigma$. These are sometimes called the **natural tolerances** for the product. Be careful to distinguish $\mu \pm 3\sigma$, the range we expect for *individual measurements*, from the \bar{x} chart control limits $\mu \pm 3\sigma / \sqrt{n}$, which mark off the expected range of *sample means*.

natural tolerances

EXAMPLE 12.9

Natural tolerances for mesh tension

The process of setting the mesh tension on computer monitors has been operating in control. The \bar{x} and s charts were based on $\mu = 275$ mV and $\sigma = 43$ mV. The s chart in Figure 12.7 and your calculation in Exercise 12.25 suggest that the process σ is now less than 43 mV. We may prefer to calculate the natural tolerances from



CASE 12.2

the recent data on 20 samples (80 monitors) in Table 12.1 (page 12-13). The estimate of the mean is $\bar{\bar{x}} = 275.065$, very close to the target value.

Now a subtle point arises. The estimate $\hat{\sigma} = \bar{s}/c_4$ used for past-data control charts is based entirely on variation *within the samples*. That's what we want for control charts, because within-sample variation is likely to be "pure common cause" variation. Even when the process is in control, there is some additional variation from sample to sample, just by chance. So the variation in the process output will be greater than the variation within samples. *To estimate the natural tolerances, we should estimate σ from all 80 individual monitors rather than by averaging the 20 within-sample standard deviations.* The standard deviation for all 80 mesh tensions is

$$s = 38.38$$

(For a sample of size 80, c_4 is very close to 1, so we can ignore it.)

We are therefore confident that almost all individual monitors will have mesh tension between

$$\bar{\bar{x}} \pm 3s = 275.065 \pm (3)(38.38) \doteq 275 \pm 115$$

We expect mesh tension measurements to vary between 160 mV and 390 mV. You see that the spread of individual measurements is wider than the spread of sample means used for the control limits of the \bar{x} chart.



The natural tolerances in Example 12.9 depend on the fact that the mesh tensions of individual monitors follow a Normal distribution. We know that the process was in control when the 80 measurements in Table 12.1 were made, so we can use them to assess Normality. Figure 12.15 is a Normal quantile plot of these measurements. There are no strong deviations from Normality. All 80 observations, including the one point that may appear suspiciously low in Figure 12.15, lie within the natural tolerances. Examining the data strengthens our confidence in the natural tolerances.

Because we can predict the performance of the mesh-tensioning process, we can tell the computer makers who buy our monitors what to expect. These customers, in fact, require us to maintain statistical control of our processes so that they need not inspect the monitors we ship to them. What is more, if a process is in control, we can see the effect of any changes we make. A process operating out of control is erratic. We can't do reliable statistical studies on such a process, and if we make a change in the process we can't clearly see the results of the change—they are hidden by erratic special cause variation. If we want to improve a process, we must first bring it into control so that we have a stable starting point from which to improve.

- 12.27 No incoming inspection.** The computer makers who buy monitors require that the monitor manufacturer practice statistical process control and submit control charts for verification. This allows the computer makers to eliminate inspection of monitors as they arrive, a considerable cost saving. Explain carefully why incoming inspection can safely be eliminated.



APPLY YOUR KNOWLEDGE

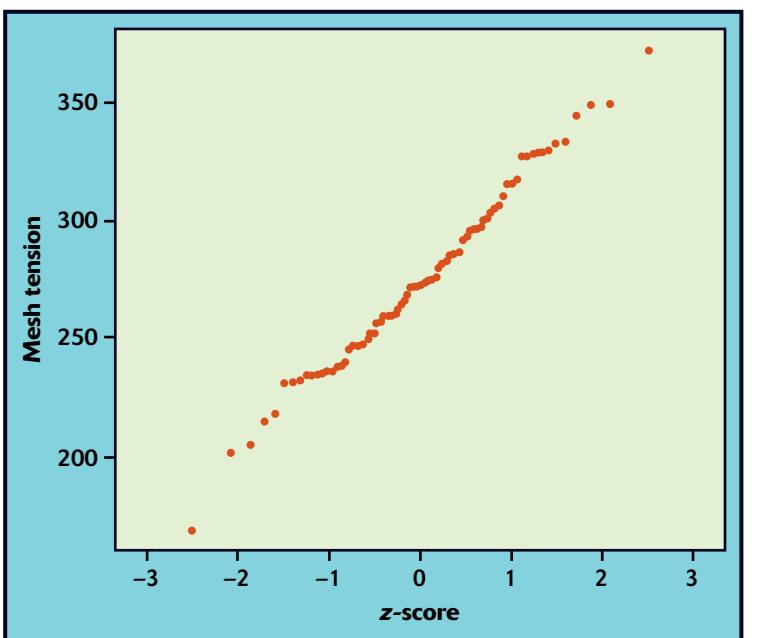


FIGURE 12.15 Normal quantile plot for the 80 mesh tension measurements of Table 12.1. Calculations about individual measurements, such as natural tolerances, depend on approximate Normality.

- 12.28 **Natural tolerances.** Table 12.6 (page 12-36) gives data on hospital losses for samples of DRG 209 patients. The distribution of losses has been stable over time. What are the natural tolerances within which you expect losses on nearly all such patients to fall?
- 12.29 **Normality?** Do the losses on the 120 individual patients in Table 12.6 appear to come from a single Normal distribution? Make a Normal quantile plot and discuss what it shows. Are the natural tolerances you found in the previous exercise trustworthy?

Don't confuse control with capability!

A process in control is stable over time. We know how much variation the finished product will show. Control charts are, so to speak, the voice of the process telling us what state it is in. *There is no guarantee that a process in control produces products of satisfactory quality.* “Satisfactory quality” is measured by comparing the product to some standard outside the process, set by technical specifications, customer expectations, or the goals of the organization. These external standards are unrelated to the internal state of the process, which is all that statistical control pays attention to.

CAPABILITY

Capability refers to the ability of a process to meet or exceed the requirements placed on it.

Capability has nothing to do with control—except for the very important point that if a process is not in control, it is hard to tell if it is capable or not.

EXAMPLE 12.10
Capability


The primary customer for our monitors is a large maker of computers. The customer informed us that adequate image quality requires that the mesh tension lie between 100 and 400 mV. Because the mesh-tensioning process is in control, we know (Example 12.9) that almost all monitors will have mesh tension between 160 and 390 mV. The process is capable of meeting the customer's requirement.

Figure 12.16 compares the distribution of mesh tension for individual monitors with the customer's specifications. The distribution of tension is approximately Normal, and we estimate its mean to be very close to 275 mV and the standard deviation to be about 38.4 mV. The distribution is safely within the specifications.

Times change, however. As computer buyers demand better screen quality, the computer maker restudies the effect of mesh tension and decides to require that tension lie between 150 and 350 mV. These new specification limits also appear in Figure 12.16. The process is not capable of meeting the new requirements. The process remains in control. The change in its capability is entirely due to a change in external requirements.



Because the mesh-tensioning process is in control, we know that it is not capable of meeting the new specifications. That's an advantage of control, but the fact remains that control does not guarantee capability. We will discuss numerical measures of capability in Section 12.3.

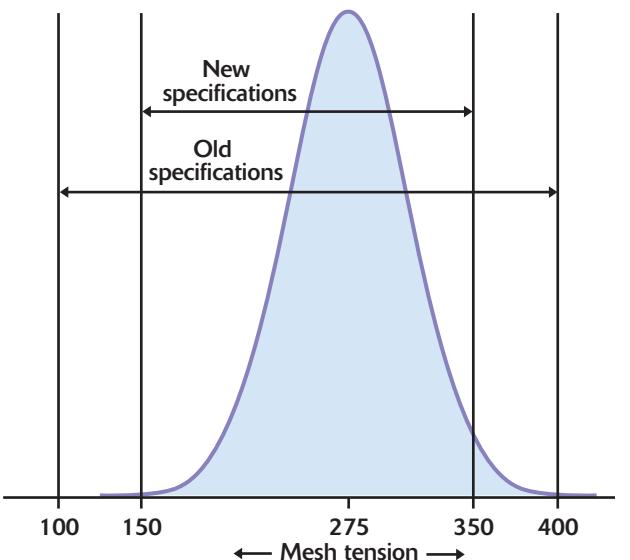


FIGURE 12.16 Comparison of the distribution of mesh tension (Normal curve) with original and tightened specifications. The process in its current state is not capable of meeting the new specifications.

Managers must understand that *if a process that is in control does not have adequate capability, fundamental changes in the process are needed.* The process is doing as well as it can and displays only the chance variation that is natural to its present state. Slogans to encourage the workers or disciplining the workers for poor performance will not change the state of the process. Better training for workers is a change in the process that may improve capability. New equipment or more uniform material may also help, depending on the findings of a careful investigation.

APPLY YOUR KNOWLEDGE

- 12.30 **Describing capability.** If the mesh tension of individual monitors follows a Normal distribution, we can describe capability by giving the percent of monitors that meet specifications. The old specifications for mesh tension are 100 to 400 mV. The new specifications are 150 to 350 mV. Because the process is in control, we can estimate that tension has mean 275 mV and standard deviation 38.4 mV.
- (a) What percent of monitors meet the old specifications?
(b) What percent meet the new specifications?
- 12.31 **Improving capability.** The center of the specifications for mesh tension is 250 mV, but the center of our process is 275 mV. We can improve capability by adjusting the process to have center 250 mV. This is an easy adjustment that does not change the process variation. What percent of monitors now meet the new specifications?

**SECTION 12.2 SUMMARY**

- An **R chart** based on the range of observations in a sample is often used in place of an **s chart**. Interpret \bar{x} and R charts exactly as you would interpret \bar{x} and s charts.
- It is common to use **out-of-control signals** in addition to “one point outside the control limits.” In particular, a **runs signal** for the \bar{x} chart allows the chart to respond more quickly to a gradual drift in the process center.
- **Control charts based on past data** are used at the **chart setup** stage for a process that may not be in control. Start with control limits calculated from the same past data that you are plotting. Beginning with the **s chart**, narrow the limits as you find special causes, and remove the points influenced by these causes. When the remaining points are in control, use the resulting limits to monitor the process.
- Statistical process control maintains quality more economically than inspecting the final output of a process. Samples that are **rational subgroups** are important to effective control charts. A process in control is stable, so that we can predict its behavior. If individual measurements have a Normal distribution, we can give the **natural tolerances**.
- A process is **capable** if it can meet the requirements placed on it. Control (stability over time) does not in itself improve capability. Remember that control describes the internal state of the process, whereas capability relates the state of the process to external specifications.

SECTION 12.2 EXERCISES

12.32 Measuring bone density. Loss of bone density is a serious health problem for many people, especially older women. Conventional X-rays often fail to detect loss of bone density until the loss reaches 25% or more. New equipment such as the Lunar bone densitometer is much more sensitive. A health clinic installs one of these machines. The manufacturer supplies a “phantom,” an aluminum piece of known density that can be used to keep the machine calibrated. Each morning, the clinic makes two measurements on the phantom before measuring the first patient. Control charts based on these measurements alert the operators if the machine has lost calibration. Table 12.7 contains data for the first 30 days of operation.¹² The units are

TABLE 12.7 Daily calibration samples for a Lunar bone densitometer

Day	Measurements		\bar{x}	s
1	1.261	1.260	1.2605	0.000707
2	1.261	1.268	1.2645	0.004950
3	1.258	1.261	1.2595	0.002121
4	1.261	1.262	1.2615	0.000707
5	1.259	1.262	1.2605	0.002121
6	1.269	1.260	1.2645	0.006364
7	1.262	1.263	1.2625	0.000707
8	1.264	1.268	1.2660	0.002828
9	1.258	1.260	1.2590	0.001414
10	1.264	1.265	1.2645	0.000707
11	1.264	1.259	1.2615	0.003536
12	1.260	1.266	1.2630	0.004243
13	1.267	1.266	1.2665	0.000707
14	1.264	1.260	1.2620	0.002828
15	1.266	1.259	1.2625	0.004950
16	1.257	1.266	1.2615	0.006364
17	1.257	1.266	1.2615	0.006364
18	1.260	1.265	1.2625	0.003536
19	1.262	1.266	1.2640	0.002828
20	1.265	1.266	1.2655	0.000707
21	1.264	1.257	1.2605	0.004950
22	1.260	1.257	1.2585	0.002121
23	1.255	1.260	1.2575	0.003536
24	1.257	1.259	1.2580	0.001414
25	1.265	1.260	1.2625	0.003536
26	1.261	1.264	1.2625	0.002121
27	1.261	1.264	1.2625	0.002121
28	1.260	1.262	1.2610	0.001414
29	1.260	1.256	1.2580	0.002828
30	1.260	1.262	1.2610	0.001414

- grams per square centimeter (for technical reasons, area rather than volume is measured).
- (a) Calculate \bar{x} and s for the first 2 days to verify the table entries for those quantities.
 - (b) What kind of variation does the s chart monitor in this setting? Make an s chart and comment on control. If any points are out of control, remove them and recompute the chart limits until all remaining points are in control. (That is, assume that special causes are found and removed.)
 - (c) Make an \bar{x} chart using the samples that remain after you have completed part (b). What kind of variation will be visible on this chart? Comment on the stability of the machine over these 30 days based on both charts.
- 12.33 **Mounting-hole distances.** Figure 12.10 (page 12-25) displays a record sheet for 18 samples of distances between mounting holes in an electrical meter. The data file *ex12-17.dat* adds \bar{x} and s for each sample. In Exercise 12.17, you found that sample 5 was out of control on the process monitoring s chart. The special cause responsible was found and removed. Based on the 17 samples that were in control, what are the natural tolerances for the distance between the holes?
- 12.34 **Measuring bone density, continued.** Remove any samples in Table 12.7 that your work in Exercise 12.32 showed to be out of control on either chart. Estimate the mean and standard deviation of individual measurements on the phantom. What are the natural tolerances for these measurements?
- 12.35 **Mounting-hole distances, continued.** The record sheet in Figure 12.10 gives the specifications as 0.6054 ± 0.0010 inch. That's 54 ± 10 as the data are coded on the record. Assuming that the distance varies Normally from meter to meter, about what percent of meters meet the specifications?
- 12.36 **Are bone density measurements Normal?** Are the 60 individual measurements in Table 12.7 at least approximately Normal, so that the natural tolerances you calculated in Exercise 12.34 can be trusted? Make a Normal quantile plot (or another graph if your software is limited) and discuss what you see.
- 12.37 **Are mounting-hole distances Normal?** Make a Normal quantile plot of the 85 distances in data file *ex12-17.dat* that remain after removing sample 5. How does the plot reflect the limited precision of the measurements (all of which end in 4)? Is there any departure from Normality that would lead you to discard your conclusions from the previous exercise? (If your software will not make Normal quantile plots, use a histogram to assess Normality.)
- 12.38 **A cutting operation.** A machine tool in your plant is cutting an outside diameter. A sample of 4 pieces is taken near the end of each hour of production. Table 12.8 gives \bar{x} and s for the first 21 samples, coded in units of 0.0001 inch from the center of the specifications. The specifications allow a range of ± 0.0002 inch about the center (a range of -2 to $+2$ as coded).
- (a) Make an s chart based on past data and comment on control of short-term process variation.
 - (b) Because the data are coded about the center of the specs, we have a given target $\mu = 0$ (as coded) for the process mean. Make an \bar{x} chart

TABLE 12.8 \bar{x} and s for samples of outside diameter

Sample	\bar{x}	s	Sample	\bar{x}	s
1	-0.14	0.48	12	0.55	0.10
2	0.09	0.26	13	0.50	0.25
3	0.17	0.24	14	0.37	0.45
4	0.08	0.38	15	0.69	0.21
5	-0.17	0.50	16	0.47	0.34
6	0.36	0.26	17	0.56	0.42
7	0.30	0.39	18	0.78	0.08
8	0.19	0.31	19	0.75	0.32
9	0.48	0.13	20	0.49	0.23
10	0.29	0.13	21	0.79	0.12
11	0.48	0.25			

and comment on control of long-term process variation. What special \bar{x} -type cause probably explains the lack of control of \bar{x} ?

- 12.39 **Special causes.** Is each of the following examples of a special cause most likely to first result in (i) a sudden change in level on the s or R chart, (ii) a sudden change in level on the \bar{x} chart, or (iii) a gradual drift up or down on the \bar{x} chart? In each case, briefly explain your reasoning.

- (a) An airline pilots' union puts pressure on management during labor negotiations by asking its members to "work to rule" in doing the detailed checks required before a plane can leave the gate.
- (b) Measurements of part dimensions that were formerly made by hand are now made by a very accurate laser system. (The process producing the parts does not change—measurement methods can also affect control charts.)
- (c) Inadequate air conditioning on a hot day allows the temperature to rise during the afternoon in an office that prepares a company's invoices.

- 12.40 **Deming speaks.** The quality guru W. Edwards Deming (1900–1993) taught (among much else) that¹³

- (a) "People work in the system. Management creates the system."
- (b) "Putting out fires is not improvement. Finding a point out of control, finding the special cause and removing it, is only putting the process back to where it was in the first place. It is not improvement of the process."
- (c) "Eliminate slogans, exhortations and targets for the workforce asking for zero defects and new levels of productivity."

Choose one of these sayings. Explain carefully what facts about improving quality the saying attempts to summarize.

- 12.41 **The Boston Marathon.** The Boston Marathon has been run each year since 1897. Winning times were highly variable in the early years, but control improved as the best runners became more professional. A clear downward trend continued until the 1980s. Rick plans to make a control chart for the

winning times from 1950 to the present. The first few times are 153, 148, 152, 139, 141, and 138. Calculation from the winning times from 1950 to 2002 gives

$$\bar{x} = 134.906 \text{ minutes} \quad \text{and} \quad s = 6.523 \text{ minutes}$$

Rick draws a center line at \bar{x} and control limits at $\bar{x} \pm 3s$ for a plot of individual winning times. Explain carefully why these control limits are too wide to effectively signal unusually fast or slow times.

- 12.42 Is Joe's weight stable?** Joe has recorded his weight, measured at the gym after a workout, for several years. The mean is 162 pounds and the standard deviation is 1.5 pounds, with no signs of lack of control. An injury keeps Joe away from the gym for several months. The data below give his weight, measured once each week for the first 16 weeks after he returns from the injury:

Week	1	2	3	4	5	6	7	8
Weight	168.7	167.6	165.8	167.5	165.3	163.4	163.0	165.5
Week	9	10	11	12	13	14	15	16
Weight	162.6	160.8	162.3	162.7	160.9	161.3	162.1	161.0

Joe wants to plot these individual measurements on a control chart. When each “sample” is just one measurement, short-term variation is estimated by advanced techniques.¹⁴ The short-term variation in Joe’s weight is estimated to be about $\sigma = 1.3$ pounds. Joe has a target of $\mu = 162$ pounds for his weight. Make a control chart for his measurements, using control limits $\mu \pm 2\sigma$. It is common to use these narrower limits on an “individuals chart.” Comment on individual points out of control and on runs. Is Joe’s weight stable or does it change systematically over this period?

12.3 Process Capability Indexes*

Capability describes the quality of the output of a process relative to the needs or requirements of the users of that output. To be more precise, capability relates the *actual performance* of a process in control, after special causes have been removed, to the *desired* performance.

Suppose, to take a simple but common setting, that there are *specifications* set for some characteristic of the process output. The viscosity of the elastomer in Case 12.3 (page 12–30) is supposed to be 45 ± 5 Mooneyes. The speed with which calls are answered at a corporate customer service call center is supposed to be no more than 30 seconds. We might measure capability by the *percent of output that meets the specifications*. When the variable we measure has a Normal distribution, we can estimate this percent using the mean and standard deviation estimated from past control chart samples. When the variable is not Normal, we can use the actual percent of the measurements in the samples that meet the specifications.

*This material is important in practice but is not needed to understand statistical process control.

EXAMPLE 12.11**Percent meeting specifications**

(a) At the conclusion of the chart setup study in Case 12.3, we have found and fixed special causes and eliminated from our data the samples on which those causes operated. The remaining viscosity measurements have $\bar{x} = 48.7$ and $s = 0.85$. Note once again that to draw conclusions about viscosity for individual lots we estimate the standard deviation σ from all individual lots, not from the average \bar{s} of sample standard deviations.

The specifications call for the viscosity of the elastomer to lie in the range 45 ± 5 . A Normal quantile plot shows the viscosities to be quite Normal. Figure 12.17(a) shows the Normal distribution of lot viscosities with the specification limits 45 ± 5 . These are marked LSL for *lower specification limit* and USL for *upper specification limit*. The percent of lots that meet the specifications is about

$$\begin{aligned} P(40 \leq \text{viscosity} \leq 50) &= P\left(\frac{40 - 48.7}{0.85} \leq Z \leq \frac{50 - 48.7}{0.85}\right) \\ &= P(-10.2 \leq Z \leq 1.53) = 0.937 \end{aligned}$$

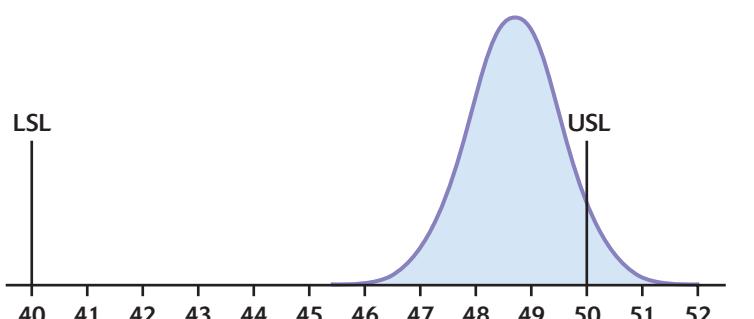
About 94% of the lots meet the specifications. If we can adjust the process center to the center of the specifications, $\mu = 45$, it is clear from Figure 12.17(a) that essentially 100% of lots will meet the specifications.

(b) Times to answer calls to a corporate customer service center are usually right-skewed. Figure 12.17(b) is a histogram of the times for 300 calls to the call center of a large bank.¹⁵ The specification limit of 30 seconds is marked USL. The median is 20 seconds, but the mean is 32 seconds. Of the 300 calls, 203 were answered in no more than 30 seconds. That is, $203/300 = 68\%$ of the times meet the specifications.

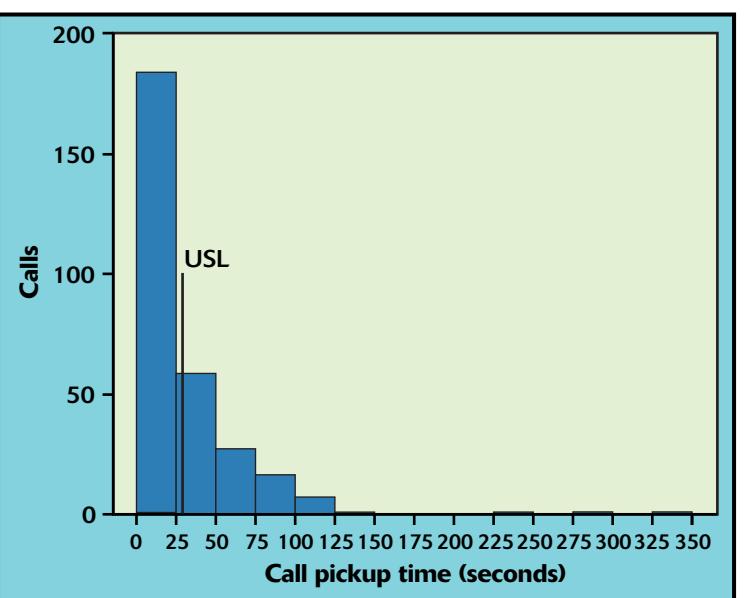


In fact, *percent meeting specifications* is a poor measure of capability. Figure 12.18 shows why. This figure compares the distributions of the diameter of the same part manufactured by two processes. The target diameter and the specification limits are marked. All the parts produced by Process A meet the specifications, but about 1.5% of those from Process B fail to do so. Nonetheless, Process B is superior to Process A because it is less variable: much more of Process B's output is close to the target. Process A produces many parts close to LSL and USL. These parts meet the specifications, but they will fit and perform more poorly than parts with diameters close to the center of the specifications. A distribution like that for Process A might result from inspecting all the parts and discarding those whose diameters fall outside the specifications. That's not an efficient way to achieve quality.

We need a way to measure process capability that pays attention to the variability of the process (smaller is better). The standard deviation does that, but it doesn't measure capability because it takes no account of the specifications that the output must meet. *Capability indexes* start with the idea of comparing process variation with the specifications. Process B will beat Process A by such a measure. Capability indexes also allow us to



(a)



(b)

FIGURE 12.17 Comparing distributions of individual measurements with specifications. (a) Viscosity has a Normal distribution. The capability is poor but will be good if we can properly center the process. (b) Response times to customer calls have a right-skewed distribution and only an upper specification limit. Capability is again poor.

measure process improvement—we can continue to drive down variation, and so improve the process, long after 100% of the output meets specifications. Continual improvement of processes is our goal, not just reaching “satisfactory” performance. The real importance of capability indexes is that they give us numerical measures to describe ever-better process quality.

- 12.43 Specification limits versus control limits. The manager you report to is confused by LSL and USL versus LCL and UCL. The notations look similar. Carefully explain the conceptual difference between specification limits for individual measurements and control limits for \bar{x} .

APPLY YOUR KNOWLEDGE

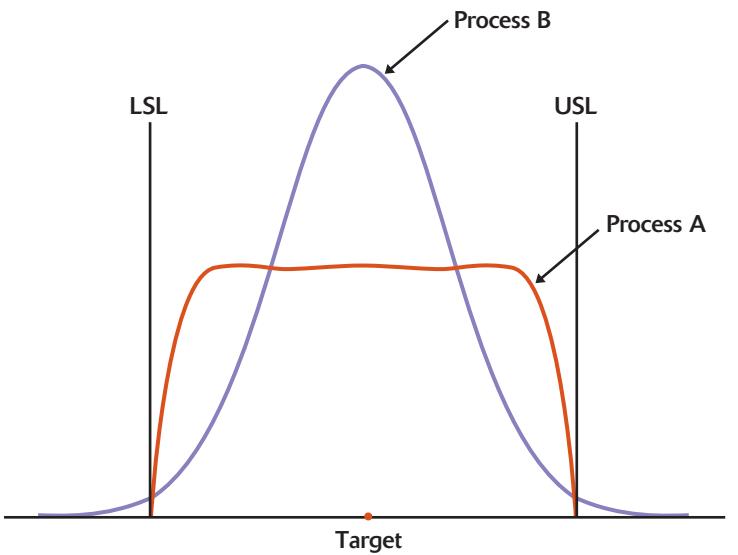


FIGURE 12.18 Two distributions for part diameters. All of the parts from Process A meet the specifications, but a higher proportion of parts from Process B have diameters close to the target.

The capability indexes C_p and C_{pk}

Capability indexes are numerical measures of process capability that, unlike percent meeting specifications, have no upper limit such as 100%. We can use capability indexes to measure continuing improvement of a process. Reporting just one number has limitations, of course. What is more, the usual indexes are based on thinking about Normal distributions. They are not meaningful for distinctly non-Normal output distributions like the call center response times in Figure 12.17(b).

CAPABILITY INDEXES

Consider a process with specification limits LSL and USL for some measured characteristic of its output. The process mean for this characteristic is μ and the standard deviation is σ . The capability index C_p is

$$C_p = \frac{USL - LSL}{6\sigma}$$

The capability index C_{pk} is

$$C_{pk} = \frac{|\mu - \text{nearer spec limit}|}{3\sigma}$$

Set $C_{pk} = 0$ if the process mean μ lies outside the specification limits. Large values of C_p or C_{pk} indicate more capable processes.

Capability indexes start from the fact that *Normal distributions are in practice about 6 standard deviations wide*. That's the 99.7 part of the

68–95–99.7 rule. Conceptually, C_p is the specification width as a multiple of the process width 6σ . When $C_p = 1$, the process output will just fit within the specifications if the center is midway between LSL and USL. Larger values of C_p are better—the process output can fit within the specs with room to spare. But a process with high C_p can produce poor-quality product if it is not correctly centered. C_{pk} remedies this deficiency by considering both the center μ and the variability σ of the measurements. The denominator 3σ in C_{pk} is half the process width. It is the space needed on either side of the mean if essentially all the output is to lie between LSL and USL. When $C_{pk} = 1$, the process has just this much space between the mean and the nearer of LSL and USL. Again, higher values are better. C_{pk} is the most common capability index, but starting with C_p helps us see how the indexes work.

EXAMPLE 12.12**Interpreting capability indexes**

Consider the series of pictures in Figure 12.19. We might think of a process that machines a metal part. Measure a dimension of the part that has LSL and USL

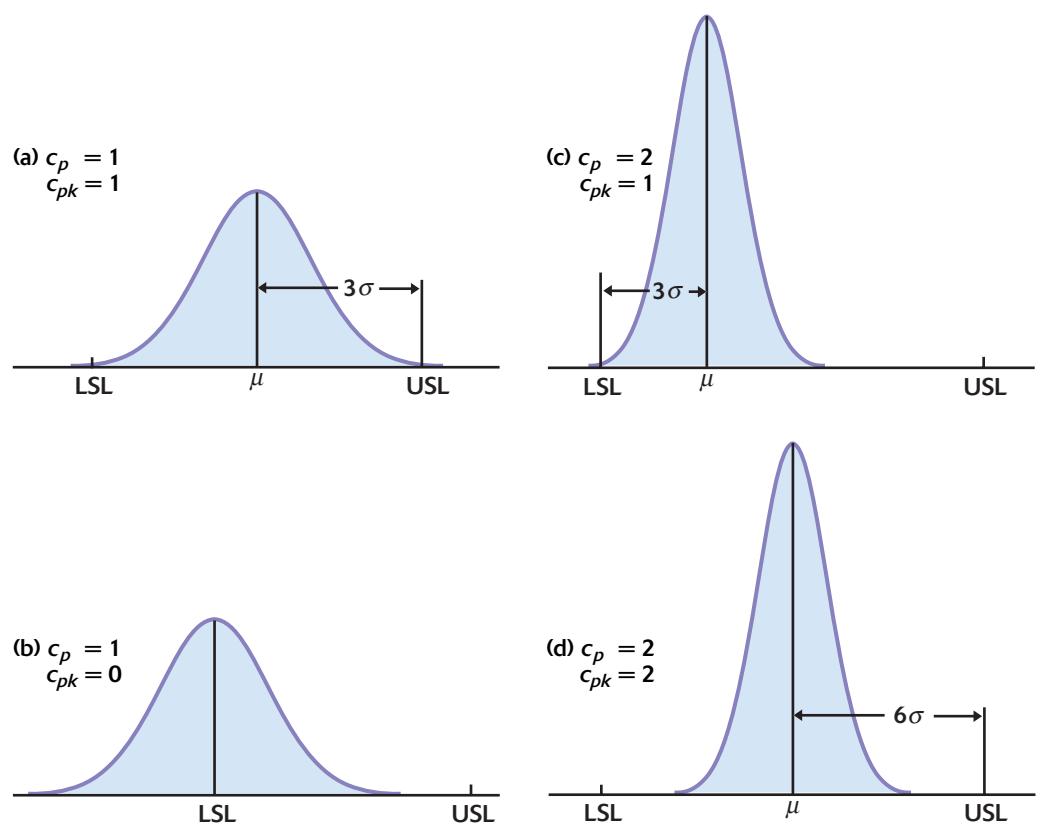


FIGURE 12.19 How capability indexes work. (a) Process centered, process width equal to specification width. (b) Process off-center, process width equal to specification width. (c) Process off-center, process width equal to half the specification width. (d) Process centered, process width equal to half the specification width.

as its specification limits. There is of course variation from part to part. The dimensions vary Normally with mean μ and standard deviation σ .

Figure 12.19(a) shows process width equal to the specification width. That is, $C_p = 1$. Almost all the parts will meet specifications if, as in this figure, the process mean μ is at the center of the specs. Because the mean is centered, it is 3σ from both LSL and USL, so $C_{pk} = 1$ also. In Figure 12.19(b), the mean has moved down to LSL. Only half the parts will meet the specifications. C_p is unchanged because the process width has not changed. But C_{pk} sees that the center μ is right on the edge of the specifications, $C_{pk} = 0$. The value remains 0 if μ moves outside the specifications.

In Figures 12.19(c) and (d), the process σ has been reduced to half the value it had in (a) and (b). The process width 6σ is now half the specification width, so $C_p = 2$. In Figure 12.19(c) the center is just 3 of the new σ 's above LSL, so that $C_{pk} = 1$. Figure 12.19(d) shows the same smaller σ accompanied by mean μ correctly centered between LSL and USL. C_{pk} rewards the process for moving the center from 3σ to 6σ away from the nearer limit by increasing from 1 to 2. You see that C_p and C_{pk} are equal if the process is properly centered. If not, C_{pk} is smaller than C_p .



EXAMPLE 12.13



Elastomer viscosity process capability

Figure 12.17(a) compares the distribution of the viscosities of lots of elastomers with the specifications $LSL = 40$ and $USL = 50$. The distribution here, as is always true in practice, is *estimated* from past observations on the process. The estimates are

$$\hat{\mu} = \bar{x} = 48.7$$

$$\hat{\sigma} = s = 0.85$$

Because capability describes the distribution of individual measurements, we once more estimate σ from individual measurements rather than using the estimate \bar{s}/c_4 that we employ for control charts.

These estimates may be quite accurate if we have data on many past lots. Estimates based on only a small amount of data may, however, be inaccurate because statistics from small samples can have large sampling variability. This important point is often not appreciated when capability indexes are used in practice. To emphasize that we can only estimate the indexes, we write \hat{C}_p and \hat{C}_{pk} for values calculated from sample data. They are

$$\begin{aligned}\hat{C}_p &= \frac{USL - LSL}{6\hat{\sigma}} \\ &= \frac{50 - 40}{(6)(0.85)} = \frac{10}{5.1} = 1.96\end{aligned}$$

$$\begin{aligned}\hat{C}_{pk} &= \frac{|\hat{\mu} - \text{nearer limit}|}{3\hat{\sigma}} \\ &= \frac{50 - 48.7}{(3)(0.85)} = \frac{1.3}{2.55} = 0.51\end{aligned}$$

$\hat{C}_p = 1.96$ is quite satisfactory because the process width is only about half the specification width. The small value of \hat{C}_{pk} reflects the fact that the process center is not close to the center of the specs. If we can move the center μ to 45, then \hat{C}_{pk} will also be 1.96.


APPLY YOUR KNOWLEDGE

- 12.44 C_p versus C_{pk} . Sketch Normal curves that represent measurements on products from a process with
- $C_p = 3$ and $C_{pk} = 1$.
 - $C_p = 3$ and $C_{pk} = 2$.
 - $C_p = 3$ and $C_{pk} = 3$.
- 12.45 Mesh tension process capability. Table 12.1 (page 12-13) gives 20 process control samples of the mesh tension of computer monitors. In Example 12.9, we estimated from these samples that **CASE 12.2** $\hat{\mu} = \bar{x} = 275.065$ mV and $\hat{\sigma} = s = 38.38$ mV.
- The original specifications for mesh tension were LSL = 100 mV and USL = 400 mV. Estimate C_p and C_{pk} for this process.
 - A major customer tightened the specifications to LSL = 150 mV and USL = 350 mV. Now what are \hat{C}_p and \hat{C}_{pk} ?
- 12.46 Mesh tension process improvement. We could improve the performance of the mesh-tensioning process discussed in the previous exercise by making an adjustment that moves the center of the process to $\mu = 250$ mV, the center of the specifications. We should do this even if the original specifications remain in force, because screens with tension closer to 250 perform better. Suppose that we succeed in moving μ to 250 with no change in the process variability σ , estimated by $s = 38.38$.
- What are \hat{C}_p and \hat{C}_{pk} with the original specifications? Compare the values with those from part (a) of the previous exercise.
 - What are \hat{C}_p and \hat{C}_{pk} with the tightened specifications? Again compare with the previous results.



Cautions about capability indexes

Capability indexes are widely used, especially in manufacturing. Some large manufacturers even set standards, such as $C_{pk} \geq 1.33$, that their suppliers must meet. That is, suppliers must show that their processes are in control (through control charts) and also that they are capable of high quality (as measured by C_{pk}). There are good reasons for requiring C_{pk} : it is a better description of process quality than “100% of output meets specs,” and it can document continual improvement. Nonetheless, it is easy to trust C_{pk} too much. We will point to three possible pitfalls.

How to cheat on C_{pk} . Estimating C_{pk} requires estimates of the process mean μ and standard deviation σ . The estimates are usually based on samples measured in order to keep control charts. There is only one reasonable estimate of μ . This is the mean \bar{x} of all measurements in recent samples, which is the same as the mean $\bar{\bar{x}}$ of the sample means. There are two different ways of estimating σ , however. The standard deviation s of all measurements in recent samples will usually be larger than the control chart estimate \bar{s}/c_4

based on averaging the sample standard deviations. The proper estimate is s because we want to describe all the variation in the process output. Larger C_{pk} 's are better, and a supplier wanting to satisfy a customer can make C_{pk} a bit larger simply by using the smaller estimate \bar{s}/c_4 for σ . That's cheating.

Non-Normal distributions. Many business processes, and some manufacturing processes as well, give measurements that are clearly right-skewed rather than approximately Normal. Measuring the times required to deal with customer calls or prepare invoices typically gives a right-skewed distribution—there are many routine cases and a few unusual or difficult situations that take much more time. Other processes have “heavy tails,” with more measurements far from the mean than in a Normal distribution. Process capability concerns the behavior of individual outputs, so the central limit theorem effect that improves the Normality of \bar{x} does not help us. Capability indexes are therefore more strongly affected by non-Normality than are control charts. It is hard to interpret C_{pk} when the measurements are strongly non-Normal. Until you gain experience, it is best to apply capability indexes only when Normal quantile plots show that the distribution is at least roughly Normal.

Sampling variation. We know that all statistics are subject to sampling variation. If we draw another sample from the same process at the same time, we get slightly different \bar{x} and s due to the luck of the draw in choosing samples. In process control language, the samples differ due to the common cause variation that is always present. C_p and C_{pk} are in practice calculated from process data because we don't know the true process mean and standard deviation. That is, these capability indexes are statistics subject to sampling variation. A supplier under pressure from a large customer to measure C_{pk} often may base calculations on small samples from the process. The resulting estimate \hat{C}_{pk} can differ from the true process C_{pk} in either direction.

EXAMPLE 12.14



Simulation: mesh tension capability

Suppose that the process of setting mesh tension in computer monitors is in control at its original level. Tension measurements are Normally distributed with mean $\mu = 275$ mV and standard deviation $\sigma = 43$ mV. The tightened specification limits are LSL = 150 and USL = 350, so the true capability is

$$C_{pk} = \frac{350 - 275}{(3)(43)} = 0.58$$

Suppose also that the manufacturer measures 4 monitors each hour and then calculates \hat{C}_{pk} at the end of an 8-hour shift. That is, \hat{C}_{pk} uses measurements from 32 monitors.

Figure 12.20 is a histogram of 24 computer-simulated \hat{C}_{pk} 's from this setting. They vary from 0.44 to 0.84, almost a two-to-one spread. It is clear that 32 measurements are not enough to reliably estimate C_{pk} .



As a very rough rule of thumb, don't trust \hat{C}_{pk} unless it is based on at least 100 measurements.

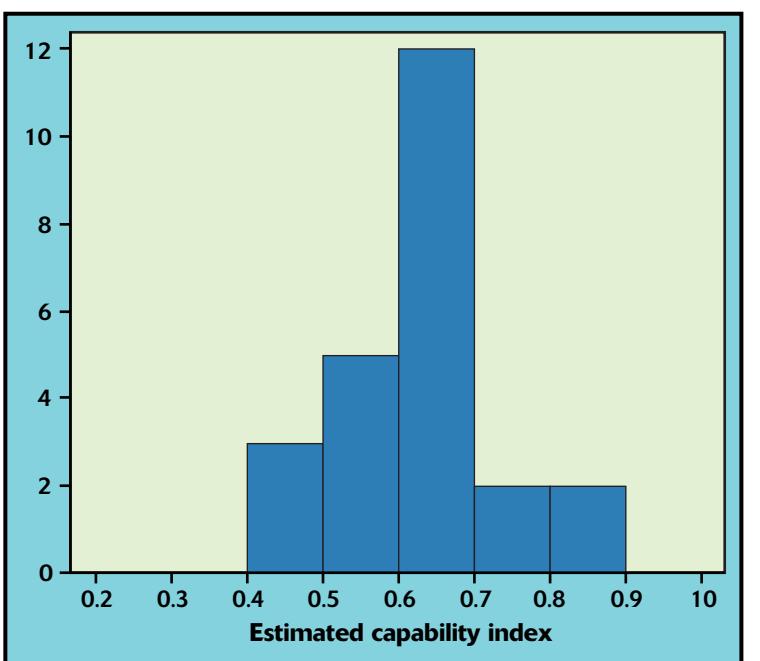


FIGURE 12.20 Capability indexes estimated from samples will vary from sample to sample. The histogram shows the variation in \hat{C}_{pk} in 24 samples, each of size 32. The process capability is in fact $C_{pk} = 0.58$.

APPLY YOUR KNOWLEDGE

- 12.47 A non-Normal distribution. Suppose that a quality characteristic has the Uniform distribution on 0 to 1. Figure 12.21 shows the density curve. You can see that the process mean (the balance point of the density curve) is $\mu = 1/2$. The standard deviation turns out to be $\sigma = 1/12$. Suppose also that LSL = 1/4 and USL = 3/4.
- Mark LSL and USL on a sketch of the density curve. What is C_{pk} ? What percent of the output meets the specifications?
 - For comparison, consider a process with Normally distributed output having mean $\mu = 1/2$ and standard deviation $\sigma = 1/12$. This process has the same C_{pk} that you found in (a). What percent of its output meets the specifications?
 - What general fact do your calculations illustrate?

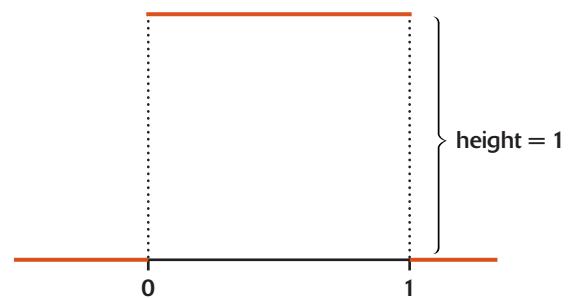


FIGURE 12.21 Density curve for the uniform distribution on 0 to 1.

- 12.48 **Same data, higher capability.** In Exercise 12.46(b) you found \hat{C}_{pk} for specifications $LSL = 150$ and $USL = 350$ using the standard deviation $s = 38.38$ for all 80 individual monitors in Table 12.1. Repeat the calculation using the control chart estimate $\hat{\sigma} = \bar{s}/c_4$. The second \hat{C}_{pk} is slightly larger.

SECTION 12.3 SUMMARY

- Capability indexes measure process variability (C_p) or process center and variability (C_{pk}) against the standard provided by external specifications for the output of the process. Larger values indicate higher capability.
- Interpretation of C_p and C_{pk} requires that measurements on the process output have a roughly Normal distribution. These indexes are not meaningful unless the process is in control so that its center and variability are stable.
- Estimates of C_p and C_{pk} can be quite inaccurate when based on small numbers of observations, due to sampling variability. You should mistrust estimates not based on at least 100 measurements.

SECTION 12.3 EXERCISES

- 12.49 **Mounting-hole distances.** Figure 12.10 (page 12-25) displays a record sheet on which operators have recorded 18 samples of measurements on the distance between two mounting holes on an electrical meter. Sample 5 was out of control on an s chart. We remove it from the data after the special cause has been fixed. The data with all sample \bar{x} 's and s 's are in the file *ex12-17.dat*. In Exercise 12.37, you saw that the measurements are reasonably Normal.
- Based on the remaining 17 samples, estimate the mean and standard deviation of the distance between holes for the population of all meters produced by this process. Make a sketch comparing the Normal distribution with this mean and standard deviation with the specification limits 54 ± 10 .
 - What are \hat{C}_p and \hat{C}_{pk} based on the data? How would you characterize the capability of the process (mention both center and variability)?
- 12.50 **DRG 209 again.** Table 12.6 (page 12-36) gives data on a hospital's losses for 120 DRG 209 patients, collected as 15 monthly samples of 8 patients each. The process has been in control and losses have a roughly Normal distribution. The hospital decides that suitable specification limits for its loss in treating one such patient are $LSL = \$4000$ and $USL = \$8000$.
- Estimate the percent of losses that meet the specifications.
 - Estimate C_p .
 - Estimate C_{pk} .
- 12.51 **Measuring capability.** You are in charge of a process that makes metal clips. The critical dimension is the opening of a clip, which has specifications 15 ± 0.5 millimeters (mm). The process is monitored by \bar{x} and s charts based on samples of 5 consecutive clips each hour. Control has recently

been excellent. The 200 individual measurements from the past week's 40 samples have

$$\bar{x} = 14.99 \text{ mm} \quad s = 0.2239 \text{ mm}$$

A Normal quantile plot shows no important deviations from Normality.

- (a) What percent of clip openings will meet specifications if the process remains in its current state?
 - (b) Estimate the capability index C_{pk} .
- 12.52 **Improving capability.** Based on the results of the previous exercise, you conclude that the capability of the clip-making process is inadequate. Here are some suggestions for improving the capability of this process. Comment on the usefulness of each action suggested.
- (a) An award program for operators producing the fewest nonconforming clips.
 - (b) A capital investment program to install new fabricating machinery.
 - (c) Additional training of operators to ensure correct operating procedures.
 - (d) Narrowing the control limits so that the process is adjusted more often.
 - (e) Purchasing more uniform (and more expensive) metal stock from which to form the clips.
- 12.53 **Six-sigma quality.** A process with $C_p \geq 2$ is sometimes said to have "six sigma quality." Sketch the specification limits and a Normal distribution of individual measurements for such a process when it is properly centered. Explain from your sketch why this is called six-sigma quality.
- 12.54 **More on six-sigma quality.** The originators of the "six-sigma quality" standard reasoned as follows. Short-term process variation is described by σ . In the long term, the process mean μ will also vary. Studies show that in most manufacturing processes, $\pm 1.5\sigma$ is adequate to allow for changes in μ . The six-sigma standard is intended to allow the mean μ to be as much as 1.5σ away from the center of the specifications and still meet high standards for percent of output lying outside the specifications.
- (a) Sketch the specification limits and a Normal distribution for process output when $C_p = 2$ and the mean is 1.5σ away from the center of the specifications.
 - (b) What is C_{pk} in this case? Is six-sigma quality as strong a requirement as $C_{pk} \geq 2$?
 - (c) Because most people don't understand standard deviations, six-sigma quality is usually described as guaranteeing a certain level of parts per million of output that fails to meet specifications. Based on your sketch in (a), what is the probability of an outcome outside the specification limits when the mean is 1.5σ away from the center? How many parts per million is this? (You will need software or a calculator for Normal probability calculations, because the value you want is beyond the limits of the standard Normal table.)

Table 12.9 gives the process control samples that lie behind the histogram of call center response times in Figure 12.17(b) on page 12-47. A sample of 6 calls is recorded each shift for quality improvement purposes. The time from the first ring until a representative answers the call is recorded.

TABLE 12.9 Fifty control chart samples of call center response times

Sample	Time (seconds)						Sample mean	Standard deviation
1	59	13	2	24	11	18	21.2	19.93
2	38	12	46	17	77	12	33.7	25.56
3	46	44	4	74	41	22	38.5	23.73
4	25	7	10	46	78	14	30.0	27.46
5	6	9	122	8	16	15	29.3	45.57
6	17	17	9	15	24	70	25.3	22.40
7	9	9	10	32	9	68	22.8	23.93
8	8	10	41	13	17	50	23.2	17.79
9	12	82	97	33	76	56	59.3	32.11
10	42	19	14	21	12	44	25.3	14.08
11	63	5	21	11	47	8	25.8	23.77
12	12	4	111	37	12	24	33.3	39.76
13	43	37	27	65	32	3	34.5	20.32
14	9	26	5	10	30	27	17.8	10.98
15	21	14	19	44	49	10	26.2	16.29
16	24	11	10	22	43	70	30.0	22.93
17	27	10	32	96	11	29	34.2	31.71
18	7	28	22	17	9	24	17.8	8.42
19	15	14	34	5	38	29	22.5	13.03
20	16	65	6	5	58	17	27.8	26.63
21	7	44	14	16	4	46	21.8	18.49
22	32	52	75	11	11	17	33.0	25.88
23	31	8	36	25	14	85	33.2	27.45
24	4	46	23	58	5	54	31.7	24.29
25	28	6	46	4	28	11	20.5	16.34
26	111	6	3	83	27	6	39.3	46.34
27	83	27	2	56	26	21	35.8	28.88
28	276	14	30	8	7	12	57.8	107.20
29	4	29	21	23	4	14	15.8	10.34
30	23	22	19	66	51	60	40.2	21.22
31	14	111	20	7	7	87	41.0	45.82
32	22	11	53	20	14	41	26.8	16.56
33	30	7	10	11	9	9	12.7	8.59
34	101	55	18	20	77	14	47.5	36.16
35	13	11	22	15	2	14	12.8	6.49
36	20	83	25	10	34	23	32.5	25.93
37	21	5	14	22	10	68	23.3	22.82

(continued)

TABLE 12.9 Fifty control chart samples of call center response times
(continued)

Sample	Time (seconds)						Sample mean	Standard deviation
38	8	70	56	8	26	7	29.2	27.51
39	15	7	9	144	11	109	49.2	60.97
40	20	4	16	20	124	16	33.3	44.80
41	16	47	97	27	61	35	47.2	28.99
42	18	22	244	19	10	6	53.2	93.68
43	43	20	77	22	7	33	33.7	24.49
44	67	20	4	28	5	7	21.8	24.09
45	118	18	1	35	78	35	47.5	43.00
46	71	85	24	333	50	11	95.7	119.53
47	12	11	13	19	16	91	27.0	31.49
48	4	63	14	22	43	25	28.5	21.29
49	18	55	13	11	6	13	19.3	17.90
50	4	3	17	11	6	17	9.7	6.31

Table 12.9 gives data for 50 shifts, 300 calls total. Exercises 12.55 to 12.57 make use of this setting.

- 12.55 **Rational subgroups?** The 6 calls each shift are chosen at random from all calls received during the shift. Discuss the reasons behind this choice and those behind a choice to time 6 consecutive calls.
- 12.56 **Chart setup.** Table 12.9 also gives \bar{x} and s for each of the 50 samples.
 - (a) Make an s chart and check for points out of control.
 - (b) If the s -type cause responsible is found and removed, what would be the new control limits for the s chart? One more s is now above the upper control limit. Suppose that this cause also is eliminated. Find the new control limits and verify that no points s are now out of control.
 - (c) Use the remaining 46 samples to find the center line and control limits for an \bar{x} chart. Comment on the control (or lack of control) of \bar{x} . (Because the distribution of response times is strongly skewed, \bar{s} is large and the control limits for \bar{x} are wide. Control charts based on Normal distributions often work poorly when measurements are strongly skewed.)
- 12.57 **Using process knowledge.** Each of the 3 out-of-control values of s in part (a) of the previous exercise is explained by a single outlier, a very long response time to one call in the sample. You can see these outliers in Figure 12.17(b). What are the values of these outliers, and what are the s -values for the 3 samples when the outliers are omitted? (The interpretation of the data is, unfortunately, now clear. Few customers will wait 5 minutes for a call to be answered, as the customer whose call took 333 seconds to answer did. We

suspect that other customers hung up before their calls were answered. If so, response time data for the calls that were answered don't adequately picture the quality of service. We should now look at data on calls lost before being answered to see a fuller picture.)

12.4 Control Charts for Sample Proportions

We have considered control charts for just one kind of data: measurements of a quantitative variable in some meaningful scale of units. We describe the distribution of measurements by its center and spread and use \bar{x} and s or \bar{x} and R charts for process control. There are control charts for other statistics that are appropriate for other kinds of data. The most common of these is the p chart for use when the data are proportions.

***p* CHART**

A p chart is a control chart based on plotting sample proportions \hat{p} from regular samples from a process against the order in which the samples were taken.

EXAMPLE 12.15

***p* chart settings**

Here are two examples of the usefulness of p charts:

Manufacturing. Measure two dimensions of a part and also grade its surface finish by eye. The part conforms if both dimensions lie within their specifications and the finish is judged acceptable. Otherwise, it is nonconforming. Plot the proportion of nonconforming parts in samples of parts from each shift.

School absenteeism. An urban school system records the percent of its eighth-grade students who are absent three or more days each month. Because students with high absenteeism in eighth grade often fail to complete high school, the school system has launched programs to reduce absenteeism. These programs include calls to parents of absent students, public-service messages to change community expectations, and measures to ensure that the schools are safe and attractive. A p chart will show if the programs are having an effect.

The manufacturing example illustrates an advantage of p charts: they can combine several specifications in a single chart. Nonetheless, p charts have been rendered outdated in many manufacturing applications by improvements in typical levels of quality. For example, Delphi, the largest North American auto electronics manufacturer, says that it reduced its proportion of problem parts from 200 per million in 1997 to 20 per million in 2001.¹⁶ At either of these levels, even large samples of parts will rarely contain any bad parts. The sample proportions will almost all be 0, so that plotting them is uninformative. It is better to choose important measured

characteristics—voltage at a critical circuit point, for example—and keep \bar{x} and s charts. Even if the voltage is satisfactory, quality can be improved by moving it yet closer to the exact voltage specified in the design of the part.

The school absenteeism example is a management application of p charts. More than 20% of all American eighth graders miss three or more days of school per month, and this proportion is higher in large cities. A p chart will be useful. Proportions of “things going wrong” are often higher in business processes than in manufacturing, so that p charts are an important tool in business.

Control limits for p charts

We studied the sampling distribution of a sample proportion \hat{p} in Chapter 8. The center line and control limits for a 3σ control chart follow directly from the facts stated there, in the box on page 505. We ought to call such charts “ \hat{p} charts” because they plot sample proportions. Unfortunately, they have always been called p charts in quality control circles. We will keep the traditional name but also keep our usual notation: p is a *process* proportion and \hat{p} is a *sample* proportion.

p CHART USING PAST DATA

Take regular samples from a process that has been in control. The samples need not all have the same size. Estimate the process proportion p of “successes” by

$$\bar{p} = \frac{\text{total number of successes in past samples}}{\text{total number of opportunities in these samples}}$$

The center line and control limits for a p chart for future samples of size n are

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

$$\text{CL} = \bar{p}$$

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

Common **out-of control signals** are one sample proportion \hat{p} outside the control limits or a run of 9 sample proportions on the same side of the center line.

If we have k past samples of the *same* size n , then \bar{p} is just the average of the k sample proportions. In some settings, you may meet samples of

unequal size—differing numbers of students enrolled in a month or differing numbers of parts inspected in a shift. The average \bar{p} estimates the process proportion p even when the sample sizes vary. Note that the control limits use the actual size n of a sample.



CASE 12.4 REDUCING ABSENTEEISM

Unscheduled absences by clerical and production workers are an important cost in many companies. Reducing the rate of absenteeism is therefore an important goal for a company's human relations department. A rate of absenteeism above 5% is a serious concern. Many companies set 3% absent as a desirable target. You have been asked to improve absenteeism in a production facility where 12% of the workers are now absent on a typical day.

You first do some background study—in greater depth than this very brief summary. Companies try to avoid hiring workers who are likely to miss work often, such as substance abusers. They may have policies that reward good attendance or penalize frequent absences by individual workers. Changing those policies in this facility will have to wait until the union contract is renegotiated. What might you do with the current workers under current policies? Studies of absenteeism by clerical and production workers who do repetitive, routine work under close supervision point to unpleasant work environment and harsh or unfair treatment by supervisors as factors that increase absenteeism. It's now up to you to apply this general knowledge to your specific problem.



First, collect data. Daily absenteeism data are already available. You carry out a sample survey that asks workers about their absences and the reasons for them (responses are anonymous, of course). Workers who are more often absent complain about their supervisors and about the lighting at their workstations. Female workers complain that the rest rooms are dirty and unpleasant. You do more data analysis:

- A Pareto chart of average absenteeism rate for the past month broken down by supervisor (Figure 12.22) shows important differences among supervisors. Only supervisors B, E, and H meet your goal of 5% or less absenteeism. Workers supervised by I and D have particularly high rates.
- Another Pareto chart (not shown) by type of workstation shows that a few types of work have high rates.

Now you take action. You retrain all the supervisors in human relations skills, using B, E, and H as discussion leaders. In addition, a trainer works individually with supervisors I and D. You ask supervisors to talk with any absent worker when he or she returns to work. Working with the engineering department, you study the workstations with high absenteeism rates and

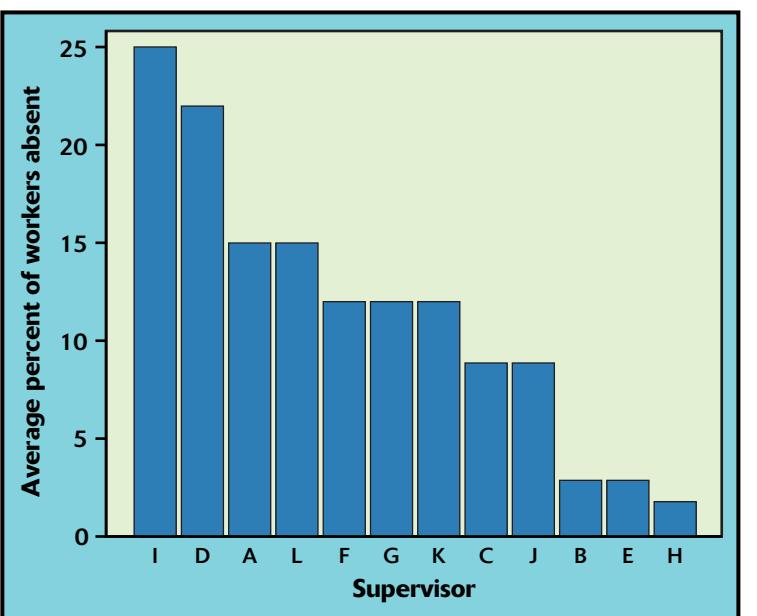


FIGURE 12.22 Pareto chart of the average absenteeism rate for workers reporting to each of 12 supervisors.

make changes such as better lighting. You refurbish the rest rooms (for both genders even though only women complained) and schedule more frequent cleaning.

EXAMPLE 12.16



Absenteeism rate p chart

Are your actions effective? You hope to see a reduction in absenteeism. To view progress (or lack of progress), you will keep a p chart of the proportion of absentees. The plant has 987 production workers. For simplicity, you just record the number who are absent from work each day. Only unscheduled absences count, not planned time off such as vacations. Each day you will plot

$$\hat{p} = \frac{\text{number of workers absent}}{987}$$

You first look back at data for the past three months. There were 64 work days in these months. The total work days available for the workers was

$$(64)(987) = 63,168 \text{ person-days}$$

Absences among all workers totaled 7580 person-days. The average daily proportion absent was therefore

$$\begin{aligned}\bar{p} &= \frac{\text{total days absent}}{\text{total days available for work}} \\ &= \frac{7580}{63,168} = 0.120\end{aligned}$$

The daily rate has been in control at this level.

TABLE 12.10 Proportions of workers absent during four weeks

Day	M	T	W	Th	F	M	T	W	Th	F
Workers absent	129	121	117	109	122	119	103	103	89	105
Proportion \hat{p}	0.131	0.123	0.119	0.110	0.124	0.121	0.104	0.104	0.090	0.106
Day	M	T	W	Th	F	M	T	W	Th	F
Workers absent	99	92	83	92	92	115	101	106	83	98
Proportion \hat{p}	0.100	0.093	0.084	0.093	0.093	0.117	0.102	0.107	0.084	0.099

These past data allow you to set up a p chart to monitor future proportions absent:

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.120 + 3\sqrt{\frac{(0.120)(0.880)}{987}}$$

$$= 0.120 + 0.031 = 0.151$$

$$CL = \bar{p} = 0.120$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.120 - 3\sqrt{\frac{(0.120)(0.880)}{987}}$$

$$= 0.120 - 0.031 = 0.089$$

Table 12.10 gives the data for the next four weeks. Figure 12.23 is the p chart.

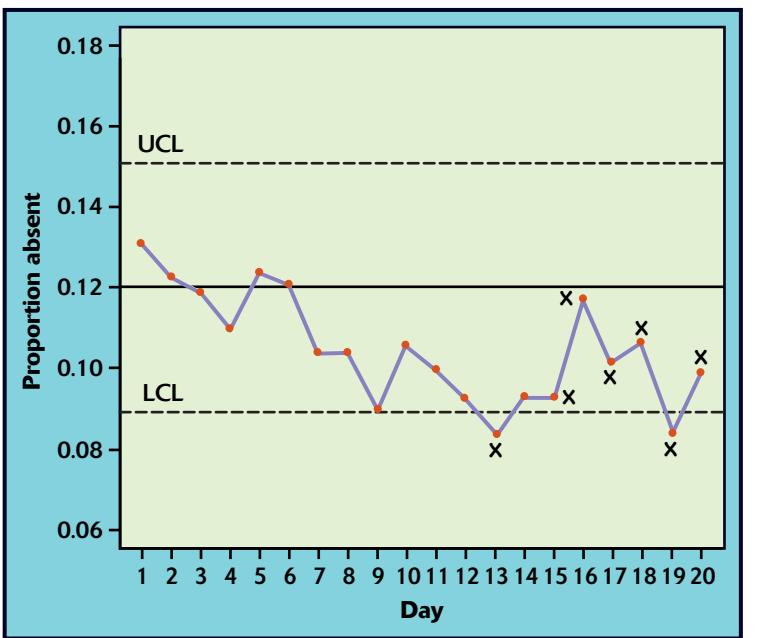


FIGURE 12.23 p chart for daily proportion of workers absent over a four-week period. The lack of control shows an improvement (decrease) in absenteeism. Update the chart to continue monitoring the process.

Figure 12.23 shows a clear downward trend in the daily proportion of workers who are absent. Days 13 and 19 lie below LCL, and a run of 9 days below the center line is achieved at Day 15 and continues. The points marked “x” are therefore all out of control. It appears that a special cause (the various actions you took) has reduced the absenteeism rate from around 12% to around 10%. The last two weeks’ data suggest that the rate has stabilized at this level. You will update the chart based on the new data. If the rate does not decline further (or even rises again as the effect of your actions wears off), you will consider further changes.

Example 12.16 is a bit oversimplified. The number of workers available did not remain fixed at 987 each day. Hirings, resignations, and planned vacations change the number a bit from day to day. The control limits for a day’s \hat{p} depend on n , the number of workers that day. If n varies, the control limits will move in and out from day to day. Software will do the extra arithmetic needed for a different n each day, but as long as the count of workers remains close to 987 the greater detail will not change your conclusion.

A single p chart for all workers is not the only, or even the best, choice in this setting. Because of the important role of supervisors in absenteeism, it would be wise to also keep separate p charts for the workers under each supervisor. These charts may show that you must reassign some supervisors.

SECTION 12.4 SUMMARY

- There are control charts for several different types of process measurements. One important type is the p chart for sample proportions \hat{p} .
- The interpretation of p charts is very similar to that of \bar{x} charts. The out-of-control signals used are also the same.

SECTION 12.4 EXERCISES

- 12.58 **Setting up a p chart.** After inspecting Figure 12.23, you decide to monitor the next four weeks’ absenteeism rates using a center line and control limits calculated from the second two weeks’ data recorded in Table 12.10. Find \bar{p} for these 10 days and give the new values of CL, LCL, and UCL. (Until you have more data, these are trial control limits. As long as you are taking steps to improve absenteeism, you have not reached the process-monitoring stage.)



- 12.59 **Unpaid invoices.** The controller’s office of a corporation is concerned that invoices that remain unpaid after 30 days are damaging relations with vendors. To assess the magnitude of the problem, a manager searches payment records for invoices that arrived in the past 10 months. The average number of invoices is 2875 per month, with relatively little month-to-month variation. Of all these invoices, 960 remained unpaid after 30 days.

- (a) What is the total number of opportunities for unpaid invoices? What is \bar{p} ?

- (b) Give the center line and control limits for a p chart on which to plot the future monthly proportions of unpaid invoices.
- 12.60 **Lost baggage.** The Department of Transportation reports that about 1 of every 200 passengers on domestic flights of the 10 largest U.S. airlines files a report of mishandled baggage. Starting with this information, you plan to sample records for 1000 passengers per day at a large airport to monitor the effects of efforts to reduce mishandled baggage. What are the initial center line and control limits for a chart of the daily proportion of mishandled baggage reports? (You will find that $LCL < 0$. Because proportions \hat{p} are always 0 or positive, take $LCL = 0$.)
- 12.61 **Aircraft rivets.** After completion of an aircraft wing assembly, inspectors count the number of missing or deformed rivets. There are hundreds of rivets in each wing, but the total number varies depending on the aircraft type. Recent data for wings with a total of 34,700 rivets show 208 missing or deformed. The next wing contains 1070 rivets. What are the appropriate center line and control limits for plotting the \hat{p} from this wing on a p chart?
- 12.62 **Doctor's prescriptions.** A regional chain of retail pharmacies finds that about 1% of prescriptions it receives from doctors are incorrect or illegible. The chain puts in place a secure online system that doctors' offices can use to enter prescriptions directly. It hopes that fewer prescriptions entered online will be incorrect or illegible. A p chart will monitor progress. Use information about past prescriptions to set initial center line and control limits for the proportion of incorrect or illegible prescriptions on a day when the chain fills 75,000 online prescriptions. What are the center line and control limits for a day when only 50,000 online prescriptions are filled?
- 12.63 **School absenteeism.** Here are data from an urban school district on the number of eighth-grade students with three or more unexcused absences from school during each month of a school year. Because the total number of eighth graders changes a bit from month to month, these totals are also given for each month.

Month	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
Students	911	947	939	942	918	920	931	925	902	883
Absent	291	349	364	335	301	322	344	324	303	344

- (a) Find \bar{p} . Because the number of students varies from month to month, also find \bar{n} , the average per month.
- (b) Make a p chart using control limits based on \bar{n} students each month. Comment on control.
- (c) The exact control limits are different each month because the number of students n is different each month. This situation is common in using p charts. What are the exact limits for October and June, the months with the largest and smallest n ? Add these limits to your p chart, using short lines spanning a single month. Do exact limits affect your conclusions?

- 12.64 **p charts are out of date.** A manufacturer of consumer electronic equipment makes full use not only of statistical process control but of automated testing

equipment that efficiently tests all completed products. Data from the testing equipment show that finished products have only 3.5 defects per million opportunities.

- (a) What is \bar{p} for the manufacturing process? If the process turns out 5000 pieces per day, how many defects do you expect to see per day? In a typical month of 24 working days, how many defects do you expect to see?
 - (b) What are the center line and control limits for a p chart for plotting daily defect proportions?
 - (c) Explain why a p chart is of no use at such high levels of quality.
- 12.65 Manufacturing isn't everything.** Because the manufacturing quality in the previous exercise is so high, the process of writing up orders is the major source of quality problems: the defect rate there is 8000 per million opportunities. The manufacturer processes about 500 orders per month.
- (a) What is \bar{p} for the order-writing process? How many defective orders do you expect to see in a month?
 - (b) What are the center line and control limits for a p chart for plotting monthly proportions of defective orders? What is the smallest number of bad orders in a month that will result in a point above the upper control limit?

STATISTICS IN SUMMARY

Statistical process control uses a combination of graphical and numerical description to decide when a process has become unstable and requires intervention. Control charts combine time plots with limits based on the sampling distributions of the statistic being charted. This chapter presents only the most common control charts, \bar{x} and s charts and p charts. Other types are also in common use, but in all cases the details are simple and the ideas and interpretation are similar. This chapter emphasizes some aspects of practical use of control charts and places them in context by also discussing process capability. Here are the specific skills you should now possess.

A. PROCESSES

1. Describe the process leading to some desired output using flowcharts and cause-and-effect diagrams.
2. Choose promising targets for process improvement, combining the process description with data collection and tools such as Pareto charts.
3. Demonstrate understanding of statistical control, common causes, and special causes by applying these ideas to specific processes.
4. Choose rational subgroups for control charting based on an understanding of the process.

B. CONTROL CHARTS

1. Make \bar{x} and s charts using given values of the process μ and σ (usually from large amounts of past data) for monitoring a process that has been in control.

2. Demonstrate understanding of the distinction between short-term (within sample) and longer-term (across samples) variation by identifying possible \bar{x} -type and s -type special causes for a specific process.
3. Interpret \bar{x} and s charts, starting with the s chart. Use both one-point-out and runs signals.
4. Estimate the process μ and σ from recent samples.
5. Set up initial control charts using recent process data, removing special causes, and basing an initial chart on the remaining data.
6. Decide when a p chart is appropriate. Make a p chart based on past data.

C. PROCESS CAPABILITY

1. Know the distinction between control and capability and apply this distinction in discussing specific processes.
2. Give the natural tolerances for a process in control, after verifying Normality of individual measurements on the process.
3. (Optional) Calculate the capability indexes C_p and C_{pk} from a process in control. Describe with sketches of Normal curves the meaning of these indexes.
4. (Optional) Detect practical pitfalls in using capability indexes—in particular, lack of attention to sampling variability in small samples.

CHAPTER 12 REVIEW EXERCISES

- 12.66 **Enlighten management.** A manager who knows no statistics asks you, “What does it mean to say that a process is in control? Is being in control a guarantee that the quality of the product is good?” Answer these questions in plain language that the manager can understand.
- 12.67 **Pareto charts.** You manage the customer service operation for a maker of electronic equipment sold to business customers. Traditionally, the most common complaint is that equipment does not operate properly when installed, but attention to manufacturing and installation quality will reduce these complaints. You hire an outside firm to conduct a sample survey of your customers. Here are the percent of customers with each of several kinds of complaints:

Category	Percent
Accuracy of invoices	25
Clarity of operating manual	8
Complete invoice	24
Complete shipment	16
Correct equipment shipped	15
Ease of obtaining invoice adjustments/credits	33
Equipment operates when installed	6
Meeting promised delivery date	11
Sales rep returns calls	4
Technical competence of sales rep	12

- (a) Why do the percents not add to 100%?
(b) Make a Pareto chart. What area would you choose as a target for improvement?
- 12.68 **What type of chart?** What type of control chart or charts would you use as part of efforts to improve each of the following performance measures in a corporate personnel office? Explain your choices.
(a) Time to get security clearance.
(b) Percent of job offers accepted.
(c) Employee participation in voluntary health screening.
- 12.69 **What type of chart?** What type of control chart or charts would you use as part of efforts to improve each of the following performance measures in a corporate information systems department? Explain your choices.
(a) Computer system availability.
(b) Time to respond to requests for help.
(c) Percent of programming changes not properly documented.
- 12.70 **Purchased material.** At the present time, about 5 lots out of every 1000 lots of material arriving at a plant site from outside vendors are rejected because they are incorrect. The plant receives about 300 lots per week. As part of an effort to reduce errors in the system of placing and filling orders, you will monitor the proportion of rejected lots each week. What type of control chart will you use? What are the initial center line and control limits?

You have just installed a new system that uses an interferometer to measure the thickness of polystyrene film. To control the thickness, you plan to measure 3 film specimens every 10 minutes and keep \bar{x} and s charts. To establish control, you measure 22 samples of 3 films each at 10-minute intervals. Table 12.11 gives \bar{x} and s for these samples. The units are $\text{mm} \times 10^{-4}$. Exercises 12.71 to 12.75 are based on this chart setup setting.

TABLE 12.11 \bar{x} and s for samples of film thickness

Sample	\bar{x}	s	Sample	\bar{x}	s
1	848	20.1	12	823	12.6
2	832	1.1	13	835	4.4
3	826	11.0	14	843	3.6
4	833	7.5	15	841	5.9
5	837	12.5	16	840	3.6
6	834	1.8	17	833	4.9
7	834	1.3	18	840	8.0
8	838	7.4	19	826	6.1
9	835	2.1	20	839	10.2
10	852	18.9	21	836	14.8
11	836	3.8	22	829	6.7

- 12.71 **s chart.** Calculate control limits for s , make an s chart, and comment on control of short-term process variation.
- 12.72 **\bar{x} chart.** Interviews with the operators reveal that in samples 1 and 10 mistakes in operating the interferometer resulted in one high outlier thickness reading that was clearly incorrect. Recalculate \bar{x} and s after removing samples 1 and 10. Recalculate UCL for the s chart and add the new UCL to your s chart from the previous exercise. Control for the remaining samples is excellent. Now find the appropriate center line and control limits for an \bar{x} chart, make the \bar{x} chart, and comment on control.
- 12.73 **(Optional) Capability.** The specifications call for film thickness $830 \pm 25 \text{ mm} \times 10^{-4}$.
- What is the estimate $\hat{\sigma}$ of the process standard deviation based on the sample standard deviations (after removing samples 1 and 10)? Estimate the capability ratio C_p and comment on what it says about this process.
 - Because the process mean can easily be adjusted, C_p is more informative than C_{pk} . Explain why this is true.
 - The estimate of C_p from (a) is probably too optimistic as a description of the film produced. Explain why.
- 12.74 **Capability.** Examination of individual measurements shows that they are close to Normal. If the process mean is set to the target value, about what percent of films will meet the specifications?
- 12.75 **Categorizing the output.** Previously, control of the process was based on categorizing the thickness of each film inspected as satisfactory or not. Steady improvement in process quality has occurred, so that just 15 of the last 5000 films inspected were unsatisfactory.
- What type of control chart would be used in this setting, and what would be the control limits for a sample of 100 films?
 - The chart in (a) is of little practical value at current quality levels. Explain why.
- 12.76 **More signals.** There are other out-of-control signals that are sometimes used with \bar{x} charts. One is “15 points in a row within the 1σ level.” That is, 15 consecutive points fall between $\mu - \sigma/\sqrt{n}$ and $\mu + \sigma/\sqrt{n}$. This signal suggests either that the value of σ used for the chart is too large or that careless measurement is producing results that are suspiciously close to the target. Find the probability that the next 15 points will give this signal when the process remains in control with the given μ and σ .

Notes for Chapter 12

- Based on Arvind Salvekar, “Application of six sigma to DRG 209,” found at the Smarter Solutions Web site, www.smartersolutions.com.
- Texts on quality management give more detail about these and other simple graphical methods for quality problems. The classic reference is Kaoru Ishikawa, *Guide to Quality Control*, Asian Productivity Organization, 1986.
- The flowchart and a more elaborate version of the cause-and-effect diagram for Case 12.1 were prepared by S. K. Bhat of the General Motors Technical Center as part of a course assignment at Purdue University.

4. We have adopted the terms “chart setup” and “process monitoring” from Andrew C. Palm’s discussion of William H. Woodall, “Controversies and contradictions in statistical process control,” *Journal of Quality Technology*, 32 (2000), pp. 341–350. Palm’s discussion appears in the same issue, pp. 356–360. We have combined Palm’s stages B (“process improvement”) and C (“process monitoring”) in writing for beginners because the distinction between them is one of degree.
5. It is common to call these “standards given” \bar{x} and s charts. We avoid this term because it easily leads to the common and serious error of confusing control limits (based on the process itself) with standards or specifications imposed from outside.
6. Provided by Charles Hicks, Purdue University.
7. See, for example, Chapter 3 of Stephen B. Vardeman and J. Marcus Jobe, *Statistical Quality Assurance Methods for Engineers*, Wiley, 1999.
8. The classic discussion of out-of-control signals and the types of special causes that may lie behind special control chart patterns is the *AT&T Statistical Quality Control Handbook*, Western Electric, 1956.
9. The data in Table 12.5 are adapted from data on viscosity of rubber samples appearing in Table P3.3 of Irving W. Burr, *Statistical Quality Control Methods*, Marcel Dekker, 1976.
10. The control limits for the s chart based on past data are commonly given as $B_4\bar{s}$ and $B_3\bar{s}$. That is, $B_4 = B_6/c_4$ and $B_3 = B_5/c_4$. This is convenient for users, but we choose to minimize the number of control chart constants students must keep straight and to emphasize that process-monitoring and past-data charts are exactly the same except for the source of μ and σ .
11. Simulated data based on information appearing in the article cited in Note 1.
12. Data provided by Linda McCabe, Purdue University.
13. The first two Deming quotes are from *Public Sector Quality Report*, December 1993, p. 5. They were found online at deming.eng.clemson.edu/pub/den/files/demqtes.txt. The third quote is part of the 10th of Deming’s “14 points of quality management,” from his book *Out of the Crisis*, MIT Press, 1986.
14. Control charts for *individual measurements* cannot use within-sample standard deviations to estimate short-term process variability. The spread between successive observations is the next best thing. Texts such as that cited in Note 7 give the details.
15. The data in Figure 12.17(b) are simulated from a probability model for call pickup times. That pickup times for large financial institutions have median 20 seconds and mean 32 seconds is reported by Jon Anton, “A case study in benchmarking call centers,” Purdue University Center for Customer-Driven Quality, no date.
16. Micheline Maynard, “Building success from parts,” *New York Times*, March 17, 2002.

Chapter 12

12.1 A flowchart for making coffee might look like Measure coffee \oslash Grind coffee \oslash Add coffee and water to coffeemaker \oslash Brew coffee \oslash Pour coffee into mug and add milk and sugar if desired. In making a cause-and-effect diagram, consider what factors might affect the final cup of coffee at each stage of the flowchart.

12.3 The 9 DRGs account for 80.5% of total losses. A Pareto chart indicates that the hospital ought to study DRGs 209 and 116 first in attempting to reduce its losses.

12.5 In Exercise 12.1 we described the process of making a good cup of coffee. Some sources of common cause variation are variation in how long the coffee has been stored and the conditions under which it has been stored, variation in the measured amount of coffee used, variation in how finely ground the coffee is, variation in the amount of water added to the coffeemaker, variation in the length of time the coffee sits between when it has finished brewing and when it is drunk, and variation in the amount of milk and/or sugar added.

Some special causes that might at times drive the process out of control would be a bad batch of coffee beans, a serious mismeasurement of the amount of coffee used or the amount of water used, a malfunction of the coffeemaker or a power outage, interruptions that result in the coffee sitting a long time before it is drunk, and the use of milk that has gone bad.

12.7 (a) CL = 11.5, UCL = 11.8, LCL = 11.2. (b) In the chart for data set A one finds that samples 12 and 20 fall above UCL. In the chart for data set B no samples are outside the control limits. In the chart for data set C, samples 19 and 20 are above UCL. (c) Data set B comes from a process that is in control. Data set A comes from a process in which the mean shifted suddenly. Data set C comes from a process in which the mean drifted gradually upward.

12.9 CL = 0.46065, UCL = 1.044, LCL = 0.

12.11 5:45 a.m., alarm rings \oslash 5:45 a.m., get out of bed \oslash 5:46 a.m., start coffeemaker \oslash 5:47 a.m., shower, shave, and dress \oslash 6:15 a.m., breakfast \oslash 6:30 a.m., brush teeth \oslash 6:35 a.m., leave home \oslash 6:50 a.m., park at school \oslash 6:55 a.m., arrive at office.

12.13 My main reasons for late arrivals at work are “don’t get out of bed when the alarm rings” (responsible for about 40% of late arrivals), “too long eating breakfast” (responsible for about 25% of late arrivals), “too long showering, shaving, and dressing” (responsible for about 20% of late arrivals), and “slow traffic on the way to work” (responsible for about 15% of late arrivals).

12.15 For the s chart, CL = 0.001128, UCL = 0.0023568, LCL = 0. For the \bar{x} chart, CL = 0.8750, UCL = 0.87662, LCL = 0.87338.

12.17 First sample: $\bar{x} = 48, s = 8.94$. Second sample: $\bar{x} = 46, s = 13.03$. For the \bar{x} chart UCL = 60.09, CL = 43, LCL = 25.9. For the s chart we note that UCL = 25.02, CL = 11.98, LCL = 0. The s chart shows a lack of control at sample point 5 (8:15 3/8), but otherwise neither chart shows a lack of control. We would want to find out what happened at sample 5 to cause a lack of control in the s chart.

12.19 We want to compute the probability that \bar{x} will be beyond the control limits, that is, that \bar{x} will be either larger than UCL = 713 or smaller than LCL = 687. The desired probability is 0.1591.

12.21 (a) 2σ control limits are UCL = $\mu + 2(\sigma/\sqrt{n})$, LCL = $\mu - 2(\sigma/\sqrt{n})$. (b) For an s chart, the 2σ control limits are UCL = $(c_4 + 2c_5)\sigma$, LCL = $(c_4 - 2c_5)\sigma$.

12.23 Presumably, a given sample is equally likely to be from the experienced clerk or the inexperienced clerk, and which clerk a sample comes from will be random. Thus, the \bar{x} chart should display two types of points: those that have relatively small values (corresponding to the experienced clerk) and those with relatively large values (corresponding to the inexperienced clerk). Both types should occur about equally often in the chart, and the pattern of large and small values should appear random.

12.25 (a) $\hat{\mu} = 275.065, \hat{\sigma} = 37.5$. (b) In Figure 12.7 we see that most of the points lie below 40 (and more than half of those below 40 lie well below 40). Of the points above 40, all but one (sample 12) are only slightly larger than 40. The s chart suggests that typical values of s are below 40, which is consistent with the estimate of σ in part (a).

12.27 By practicing statistical process control, the manufacturer is, in essence, inspecting samples of the monitors it produces and fixing any problems that arise. The control charts the manufacturer creates are a record of this inspection process. Incoming inspection is thus redundant and no longer necessary.

12.29 The plot shows no serious departures from Normality. For Normal data the natural tolerances are trustworthy, so based on our Normal quantile plot, the natural tolerances we found in the previous exercise are trustworthy.

12.31 99.06% of monitors will meet the new specifications if the process is centered at 250 mV.

12.33 The natural tolerances for the distance between the holes are $\hat{\mu} \pm 3\hat{\sigma} = 43.41 \pm 37.17$.

12.35 About 43.16% of meters meet specifications.

12.37 The large-scale pattern in the plot looks Normal in that the centers of the horizontal bands of points appear to lie along a straight line. These horizontal bands correspond to observations having the same value, resulting from the limited precision (roundoff). With greater precision, these observations would undoubtedly differ and would probably appear to come from a Normal distribution. However, because of the lack of precision, we should probably view these data as only approximately Normal and thus the calculations in Exercise 12.34 should be considered only approximate.

- 12.39 (a) If all pilots suddenly adopt the policy of “work to rule” (and are now more careful and thorough about doing all the checks), one would expect to see a sudden increase in the time to complete the checks but probably not much of a change in the variability of the time. Thus, we would expect to see a sudden change in level on the \bar{x} chart. (b) Presumably samples of measurements made by the laser system will have a much smaller standard deviation than samples of measurements made by hand. We would expect to see this reflected in a sudden change (decrease) in the level of an s or R chart. (c) Presumably the measurement of interest is either the temperature in the office or the number of invoices prepared. If it is the temperature of the office, then as time passes, the temperature will gradually increase as outdoor temperature increases. This would be reflected in a gradual shift up on the \bar{x} chart. If the measurement of interest is the number of invoices prepared, as the temperature in the office gradually increases, the workers gradually become less comfortable. They may take more breaks for water or become tired more quickly and thus accomplish less. We might expect to see a gradual decrease in the number of invoices prepared and hence a gradual shift down in the \bar{x} chart.
- 12.41 The winning time for the Boston Marathon has been gradually decreasing over time. There are thus two sources of variation in the data. One is the process variation that we would observe if times were stable (random variation that we might observe over short periods of time), and the other is variation due to the downward trend in times (the large differences between times from the early 1950s compared to recent times). In using the standard deviation s of all the times between 1950 and 2002, we include both of the sources of variation. However, to determine if times in the next few years are unusual, we should consider only the process variation (variability in recent times, where the process is relatively stable). Using s overestimates the process variation, resulting in control limits that are too wide to effectively signal unusually fast or slow times.
- 12.43 LSL and USL define the acceptable set of values for *individual* measurements on the output of a process. These limits represent the *desired performance* of the process. LCL and UCL for \bar{x} are the lower and upper control limits for the *means of samples* of several individual measurements on the output of a process. They represent the *actual performance* of the process when it is in control.
- 12.45 (a) $\hat{C}_p = 1.303$, $\hat{C}_{pk} = 1.085$. (b) $\hat{C}_p = 0.869$, $\hat{C}_{pk} = 0.651$.
- 12.49 (a) $\hat{\mu} = 43.41$, $\hat{\sigma} = 12.39$. (b) $\hat{C}_p = 0.27$, $\hat{C}_{pk} = 0.02$. The capability is poor (both indexes are small). The reasons are that the process is not centered (we estimate the process mean to be 43.41, but the midpoint of the specification limits is 54), and the process variability is large (we saw in Exercise 12.33 that the natural tolerances for the process are 43.41 ± 37.17 , a much wider range than the specification limits of 54 ± 10).
- 12.51 (a) 17.78% of clip openings will meet specifications if the process remains in its current state. (b) $\hat{C}_{pk} = 0.06$.
- 12.53 $C_p = \text{USL} - \text{LSL}/6\sigma$, so if $C_p \geq 2$ we must have $\text{USL} - \text{LSL} \geq 22\sigma$. If the process is properly centered, then its mean μ will be halfway between USL

and LSL, and hence USL and LSL will be at least 6σ from μ . Thus, this is called six-sigma quality.

12.55 If the same representatives work a given shift, and calls can be expected to arrive at random during a given shift, then it might make sense to choose calls at random from all calls received during the shift. In this case, the process mean and standard deviation should be stable over the entire shift, and random selection should lead to sensible estimates of the process mean and standard deviation.

If different representatives work during a shift, or if the rate of calls varies over a shift, then the process may not be stable over the entire shift. There may be changes in either the process mean or the process standard deviation over the shift. A random sample from all calls received during the shift would overestimate the process variability because it would include variability due to special causes. Stability would more likely be seen only over much shorter time periods, and thus it would be more reasonable to time 6 consecutive calls.

12.57 The three outliers are a response time of 276 seconds (occurring in Sample 28), resulting in a value of $s = 107.20$; a response time of 244 seconds (occurring in Sample 42), resulting in a value of $s = 93.68$; and a response time of 333 seconds (occurring in Sample 46), resulting in a value of $s = 119.53$. When the outliers are omitted, the values of s become $s = 9.28$ (Sample 28), $s = 6.71$ (Sample 42), $s = 31.01$ (Sample 46).

12.59 (a) The total number of opportunities for unpaid invoices = 128750, $\bar{p} = 0.0334$. (b) UCL = 0.0435, CL = 0.0334, LCL = 0.0233.

12.61 UCL = 0.01307, CL = 0.00599, LCL = 0.

12.63 (a) $\bar{p} = 0.356$, $\bar{n} = 921.8$. (b) For the p chart, UCL = 0.4033, CL = 0.356, LCL = 0.3087. The process (proportion of students per month with 3 or more unexcused absences) appears to be in control. That is, there are no months with unusually high or low proportions of absences, and no obvious trends. (c) For October, UCL = 0.4027 and LCL = 0.3093. For June, UCL = 0.4043 and LCL = 0.3077. These exact limits do not affect our conclusions in this case.

12.65 (a) $\bar{p} = 0.008$ and if the manufacturer processes 500 orders per month, we would expect to see 4 defective orders per month. (b) UCL = 0.020, CL = 0.008, LCL = 0. To be above the UCL, the number of defective orders must be greater than 10.

12.67 (a) The percents add up to larger than 100% because customers can have more than one complaint. (b) The category with the largest number of complaints is the ease of obtaining invoice adjustments/credits, so we might target this area for improvement.

12.69 (a) Presumably we are interested in monitoring and controlling the amount of time the system is unavailable. We might measure the time the system is not available in sample time periods and use an \bar{x} and an s chart to monitor how this length of time the system is unavailable varies. (b) To monitor the time to respond to requests for help we might measure the response times in a sample of time periods and use an \bar{x} and an s chart to monitor how the response times vary. (c) We might examine samples of programming changes

and record the proportion in the sample that are not properly documented. Because we are monitoring a proportion, we would use a p chart.

12.71 CL = 7.65, UCL = 19.642, LCL = 0. The first sample is out of control, perhaps reflecting an initial lack of skill or initial problems in using the new system. After the first sample, the process is in control with respect to short-term variation, although sample 10 is very close to the upper control limit.

12.73 (a) $\hat{C}_p = 1.14$. This value tells us that the specification limits will lie just within 3 standard deviations of the process mean if the process mean is in the center of the specification limits. (b) C_{pk} depends on the value of the process mean. If the process mean can easily be adjusted, it is easy to change the value of C_{pk} . A better measure of the process capability is to center the process mean within the specification limits and then compute a capability index. If the process is properly centered, C_{pk} and C_p will be the same, so using C_p is ultimately more informative about the process capability. (c) We used \bar{s}/c_4 to estimate σ , the process variation. A better estimate would be to compute the sample standard deviation s of all $22 \times 3 = 66$ observations in the samples. This gives an estimate of all the variation in the output of the process (including sample-to-sample variation). Using \bar{s}/c_4 is likely to give a slightly too small estimate of the process variation and hence a slightly too large (optimistic) estimate of C_p .

12.75 (a) We would use a p chart. The control limits would be UCL = 0.019, CL = 0.003, and LCL = 0. (b) If the proportion of unsatisfactory films is 0.003, then in a sample of 100 films the expected number of unsatisfactory films is 0.3. Thus, most of our samples will have no unsatisfactory films, and plotting the sample values (most of which will be 0) will not be very informative.

CHAPTER 13

Time Series Forecasting

Introduction

Business decisions often involve data tracked over time. Quarterly sales figures, annual health benefits costs, monthly product demand, daily stock prices, and changes in market share are all examples of *time series* data.

time series

TIME SERIES

Measurements of a variable taken at regular intervals over time form a *time series*.

We've already encountered time series data in earlier chapters. Figure 1.7 (page 21) displayed a *time plot* of monthly orange prices from January 1991 to December 2000. Case 1.1 looked at unemployment rates including Figure 1.8 (page 22) displaying a time plot of national unemployment rates from January 1990 to August 2001. We handle time series data with the same approach used in earlier chapters:

- Plot the data, then add numerical summaries.
- Identify overall patterns and deviations from those patterns.
- When the overall pattern is quite regular, use a compact mathematical model to describe it.

In this chapter, we will focus on the plots and calculations that are most helpful when describing time series data. We will learn to identify patterns common to time series data as well as the models that are commonly used to describe those patterns.

EXAMPLE 13.1 Monthly retail sales

The Census Bureau tracks retail sales using the Monthly Retail Trade Survey.¹ Retail establishments are categorized and tracked by their North American Industry Classification System (NAICS) code. NAICS code 452

corresponds to General Merchandise Stores (like Wal-mart). Comparing sales data for a specific retail company with national sales data lets one see how the company is performing compared to its industry as a whole.

Figure 13.1 plots monthly retail sales of General Merchandise Stores (NAICS 452) beginning January 1992 and ending May 2002 (125 months). The plot reveals interesting characteristics of retail sales for these stores:

- Since January 1992, overall sales have gradually increased.
- A distinct pattern repeats itself approximately every 12 months: January sales are lower than the rest of the year, sales then increase but level off until October when sales dramatically increase to a peak in December.

The gradual increase in sales is an example of a *trend* in a time series. Embedded in the monthly ups and downs of sales figures, we see a persistent pattern describing the general, long-run behavior of sales over the time period from January 1992 to May 2002.

The repeating pattern evident in Figure 13.1 is an example of *seasonal variation* in a time series. Seasonal variation must be taken into account when making a short-run prediction of future sales. It would be unwise to ignore the peaks in December and the troughs in January when predicting sales for specific months.

Trend and seasonal variation were defined in Chapter 1 (page 20). A prediction of a future value of a time series is called a *forecast*. This chapter focuses on exploring time series data with the eventual goal of forecasting the variable of interest.

trend

seasonal variation

forecast

APPLY YOUR KNOWLEDGE

13.1 JCPenney sales. Table 13.1 contains retail sales for JCPenney in millions of dollars. The data are quarterly beginning with the first quarter of 1996 and ending with the fourth quarter of 2001.²

- Before plotting these data, inspect the values in the table. Do you see any interesting features of JCPenney quarterly sales?
- Now, make a time plot of the data. Be sure to connect the points in your plot to highlight patterns.
- Is there an obvious trend in JCPenney quarterly sales? If so, is the trend positive or negative?
- Is there an obvious repeating pattern in the data? If so, clearly describe the repeating pattern.

13.1 Trends and Seasons

In the modern business environment, large and small companies alike depend on long-run forecasts of numerous variables to guide their own long-run business plans. A retail chain uses population growth forecasts to help decide where to locate future stores. A manufacturer of car alarm systems looks at auto theft growth rates in key areas of the country to estimate demand for their products. Some companies provide one-year-ahead forecasts of key quantities in their annual reports, so that a 2001 annual report will contain forecasts for how the company will perform in 2002. The first step toward a reliable forecast is to identify any trend in the time series of interest.

Identifying trends

EXAMPLE 13.2 Monthly retail sales

Consider the monthly retail sales data from Example 13.1. Since January 1992, overall sales have gradually increased. We can use the ideas of simple linear regression (Chapter 10) and software to estimate this upward trend. From the Excel regression output below, we estimate the linear trend to be

$$\widehat{\text{SALES}} = 18,736 + 145.5x$$

where x is the number of months elapsed beginning with the first month of the time series; that is, $x = 1$ corresponds to January 1992, $x = 2$ corresponds to February 1992, etc. Sales are measured in millions of dollars.

	A	B	C	D	E
1	SUMMARY OUTPUT				
2					
<i>Regression Statistics</i>					
4	Multiple R	0.655338351			
5	R Square	0.429468354			
6	Adjusted R Square	0.424829885			
7	Standard Error	6101.522201			
8	Observations	125			
9					
10		Coefficients	Standard Error	t Stat	P-value
11	Intercept	18736.49742	1098.055282	17.0633462	3.1962E-34
12	X	145.5311521	15.12438406	9.6222862	1.11261E-16

Figure 13.2 displays the time series plot of Figure 13.1 with the trend line superimposed. The equation of the line provides a mathematical model for the observed trend in sales. The estimated slope of 145.5 indicates retail sales at General Merchandise Stores increased an average of 145.5 million dollars per month from January 1992 to May 2002.

Notice how the trend line in Figure 13.2 ignores the seasonal variation in the retail sales time series. Because of this, using the equation above to forecast sales for, say, December 2002 will likely result in a gross underestimate since the line underestimates sales for all the Decembers in the data set. We will need to take the month-to-month pattern into account if we wish to accurately forecast sales in a specific month.

In Chapter 11, we used regression techniques to fit a variety of models to data. For example, we used a polynomial model to predict the prices of houses in Example 11.14 (page 672). Trends in time series data may also be best described by a curved model like a polynomial. The techniques of Chapter 11 help us fit such trend curves to time series data. Case 13.1 illustrates another type of curved relationship—*exponential trend*.

CASE 13.1 SELLING DVD PLAYERS

The popularity of the DVD format has exploded in the relatively short period of time since its introduction in March 1997. At the end of June 2002, nearly 33 million DVD players had been sold in the U.S. with over 18,000 titles available in the DVD format. The Consumer Electronics Association tracks monthly sales of DVD players. The data are used by permission of CEA and provided in the data file *ca13_001.dat*.³ Figure 13.3 plots the number of units sold over time. The plot includes sales data from April 1997 to June 2002 (63 months).

Let your eye follow the sales numbers from April to April. DVD player sales increased very little from April 1997 to April 1998. Notice how the April-to-April sales increases have grown in magnitude over the years with April 2001 to April 2002 showing the largest increase. This pattern of increasing growth is an example of an *exponential trend*. Figure 13.4 shows the DVD player sales data with an exponential trend superimposed. Statistical software estimates the exponential trend to be

$$\widehat{\text{UNITS}} = 29,523.65e^{0.069x}$$

where x is the number of months elapsed beginning with the first month of the time series; that is, $x = 1$ corresponds to April 1997, $x = 2$ corresponds to May 1997, etc. (The mathematical constant e was introduced in Chapter 5 on page 336.)

We can use our trend model to forecast the number of DVD players to be sold in a future month. July 2002 corresponds to $x = 64$, and we get

$$\widehat{\text{UNITS}} = 29,523.65e^{0.069 \times 64} = 2,443,512$$

Our forecasted value ignores the monthly ups and downs in the time series. In fact, if you look closely at the data, you will see that July sales have

*exponential
trend
Case 13.1*

always been less than June sales (except in 1998). Our trend model ignores such information about the time series. Our forecast of over 2.4 million DVD players to be sold in July 2002 is greater than the 1.6 million DVD players sold in June 2002.

Remember, a trend equation may be a good description of the long-run behavior of the data, but we need to account for short-run phenomena (like seasonal variation) to improve the accuracy of our forecasts.

APPLY YOUR KNOWLEDGE

13.2 Number of Macs shipped. Table 13.2 displays the time series of number of Macintosh computers shipped in each of eight consecutive fiscal quarters.⁴ We usually want a time series longer than just eight time periods, but this short series will give you a chance to do some calculations by hand. Hand calculations sometimes take the mystery out of what computers do so quickly and efficiently for us.

- (a) Make a time plot of these data.
- (b) Using the following summary information, calculate the least-squares regression line for predicting the number of Macs shipped (in thousands of units). The variable Time simply takes on the values 1, 2, 3, . . . , 8 in time order.

Variable	Mean	Std Dev	Correlation
Time	4.5	2.44949	0.3095
Macs shipped	773.5	62.56197	

- (c) Sketch the least-squares line on your time plot from part (a). Does the linear model appear to fit this data well?

13.3 JCPenney sales. In Exercise 13.1, you took a first look at the data in Table 13.1. Use statistical software to further investigate the JCPenney sales data.

- (a) Find the least-squares line for the sales data. Use 1, 2, 3, . . . as the values for the explanatory variable with $x = 1$ corresponding to the first quarter of 1996, $x = 2$ corresponding to the second quarter of 1996, etc.
- (b) The intercept is a prediction of sales for what quarter?
- (c) Interpret the slope in the context of JCPenney quarterly sales.

Seasonal patterns

Variables of economic interest are often tied to other events that repeat with regular frequency over time. Agriculture-related variables will vary with the growing and harvesting seasons. In Example 1.7 (page 20), we saw that the prices of oranges tend to be lowest in January and February and highest in August and September. Sales data may be linked to events like regular changes in the weather, the start of the

school year, and the celebration of certain holidays. The monthly retail sales data of Example 13.1 (page xxx) exhibit a strong pattern that repeats every twelve months with sales peaking in December. To improve the accuracy of our forecasts we need to account for seasonal variation in our time series.

Using indicator variables

Indicator variables were introduced in Chapter 11 (page 675). We can use indicator variables to add the seasonal pattern in a time series to a trend model. Let's look at the details for the monthly retail sales data of Examples 13.1 and 13.2.

EXAMPLE 13.3 Monthly retail sales

The seasonal pattern in the sales data seems to repeat every 12 months, so we begin by creating $12 - 1 = 11$ indicator variables.

$$X_1 = \begin{cases} 1 & \text{if the month is January} \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if the month is February} \\ 0 & \text{otherwise} \end{cases}$$

⋮

$$X_{11} = \begin{cases} 1 & \text{if the month is November} \\ 0 & \text{otherwise} \end{cases}$$

December data are indicated when all eleven indicator variables are zero.

We can extend the trend model estimated in Example 13.2 with these indicator variables. The new model will capture the seasonal pattern in the time series.

$$\text{SALES} = \beta_0 + \beta_1 x + \beta_2 X_1 + \cdots + \beta_{12} X_{11} + \epsilon$$

Fitting this multiple regression model to our data, we get the following output:

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.993493346					
5	R Square	0.987029029					
6	Adjusted R Square	0.985639282					
7	Standard Error	964.1133873					
8	Observations	125					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	12	7921942577	660161881.4	710.2221576	1.1832E-99	
13	Residual	112	104105637.8	929514.6235			
14	Total	124	8026048215				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	37472.69024	343.3506365	109.1382577	1.4241E-115	36792.38539	38152.9951
18	x	140.1304508	2.392701887	58.5657794	7.61806E-86	135.3896217	144.87128
19	X1	-24276.1932	421.4213062	-57.60551933	4.57481E-85	-25111.18498	-23441.20142
20	X2	-23748.68729	421.3601691	-56.36196544	4.86201E-84	-24583.55793	-22913.81664
21	X3	-20271.18137	421.3126119	-48.1143474	1.19904E-76	-21105.95779	-19436.40496
22	X4	-20250.49364	421.2786392	-48.06912043	1.32582E-76	-21085.20275	-19415.78454
23	X5	-18517.98773	421.2582543	-43.95875343	1.79155E-72	-19352.65645	-17683.31902
24	X6	-19574.51729	431.403553	-45.37402894	6.21054E-74	-20429.28761	-18719.74698
25	X7	-20323.64775	431.330558	-47.11849733	1.1186E-75	-21178.27343	-19469.02206
26	X8	-19626.8782	431.2708257	-43.1906753	1.15687E-71	-19481.38553	-17772.37087
27	X9	-20878.10865	431.2243614	-48.41588397	6.14847E-77	-21732.52391	-20023.69338
28	X10	-18933.0391	431.1911697	-43.9086893	2.02136E-72	-19787.3886	-18078.6896
29	X11	-13842.16955	431.1712534	-32.10364662	2.57955E-58	-14696.47959	-12987.85951

Figure 13.5 displays the time series plot of Figure 13.1 with the trend-and-season model superimposed. You can see the dramatic improvement over the trend-only model by comparing Figure 13.2 with Figure 13.5. The improved model fit is reflected in the R^2 values for the two models: $R^2 = 0.987$ for the trend-and-season model and $R^2 = 0.429$ for the trend-only model.

APPLY YOUR KNOWLEDGE

13.4 Number of Macs shipped. In Exercise 13.2, you made a time plot of the data in Table 13.2. With only eight quarters, a strong quarterly pattern is hard to detect. Add indicator variables for first, second, and third quarters to the linear trend model fit in Exercise 13.3. Call these indicator variables X_1 , X_2 , and X_3 , respectively. Use statistical software to fit this multiple regression model.

- (a) Write down the estimated trend-and-season model.
- (b) Explain why no indicator variable is needed for fourth quarters.
- (c) What does the ANOVA F – test indicate about this model?

13.5 JCPenney sales. In Exercise 13.1, you made a time plot of the data in Table 13.1. Sales seem to follow a pattern of ups and downs that repeats every four quarters. Add indicator variables for first, second, and third quarters to the linear trend model fit in Exercise 13.2. Call these indicator variables X_1 , X_2 , and X_3 , respectively.

- (a) Write down the estimated trend-and-season model.

- (b) Explain why no indicator variable is needed for fourth quarters.
 (c) Does the intercept still predict sales for a specific quarter? If so, what quarter? Compare the estimated intercept of this model with that of the trend-only model (see Exercise 13.3(b)). Given the pattern of seasonal variation, which appears to be the better estimate?

Using seasonality factors

Using indicator variables to incorporate seasonality into a trend model views the model as a trend component plus a seasonal component.

$$y = \text{TREND} + \text{SEASON}$$

You can see this in Example 13.3 where we *added* the indicator variables to the trend model. A second approach to accounting for seasonal variation is to calculate an “adjustment” factor for each season. The trend is adjusted each particular season by *multiplying* it by the appropriate *seasonality factor*. One seasonality factor is *seasonality factor* calculated for each season observed in the data. Using seasonality factors views the model as a trend component times a seasonal component.

$$y = \text{TREND} \times \text{SEASON}$$

This view of our model suggests that we can identify the seasonal component by calculating

$$\frac{y}{\text{TREND}} = \text{SEASON}$$

Let’s see how this works in practice using the monthly retail sales data of Example 13.1.

EXAMPLE 13.4 Monthly retail sales

We have been treating each month as a season for the monthly retail sales data, so we will need to calculate twelve seasonality factors. In Example 13.2 (page xxx), the trend component was estimated to be

$$\widehat{\text{SALES}} = 18,736 + 145.5x$$

For each of the 125 months in our time series, we calculate the ratio of actual sales divided by predicted sales, $\widehat{\text{SALES}}$. We then average these ratios by month. That is, we compute the average for all the January ratios, then the average for all the February ratios, and so on. These averages become our twelve seasonality factors. The arithmetic needed is straightforward but tedious given that we have 125 months of data. The table below displays the seasonality factors that result.

Month	Seasonality Factor	Month	Seasonality Factor
January	0.784	July	0.931
February	0.805	August	0.994
March	0.931	September	0.913
April	0.935	October	0.985
May	0.995	November	1.166
June	0.958	December	1.662

Our trend-and-season model is then

$$\widehat{\text{SALES}} = (18,736 + 145.5x) \times \text{SF}$$

where SF is the seasonality factor for the appropriate month corresponding to the value of x .

The seasonality factors are a snapshot of the typical ups and downs over the course of a year. The factors are used to adjust the trend model to account for the differences from season to season (month to month in this example). The factors themselves have an interpretation if their average is 1—our factors have an average of 1.005 which is very close to 1.* Think of the factors as percents, then the factors show how each month’s sales compare to average monthly sales for all twelve months. For example, December’s seasonality factor is 1.662, or 166.2%, indicating that December sales are typically 66.2% above the average for all twelve months. January’s seasonality factor is 0.784, or 78.4%, indicating that January sales are typically 21.6% below the annual average. Figure 13.6 plots the seasonality factors in order from January to December. The reference line marked at 1.0 (100%) is a visual aid for interpreting the factors compared to the overall average monthly sales. Notice how the seasonality factors mimic the pattern that is repeated every twelve months in Figure 13.1—dividing the sales data by the estimated trend has isolated the seasonal pattern in the time series.

EXAMPLE 13.5 Next month’s DVD player sales

Case 13.1

Identifying an appropriate model to capture trend and, if present, seasonal variation is an important first step in forecasting future values of a time series. In Case 13.1 (page xxx), we forecasted the number of DVD players that would sell in July 2002 based on data ending in June 2002. Our forecast used a trend-only model. Let’s forecast July 2002 DVD player sales using a trend-and-season model.

The trend-only forecast for DVD player sales in July 2002 ($x = 64$) was calculated as

$$\widehat{\text{UNITS}} = 29,523.65e^{0.069 \times 64} = 2,443,512$$

*If the factors did not have an average close to 1, then we could adjust them so that they did average to 1 by dividing each factor by the total of all twelve factors.

If we calculate the seasonality factors for the DVD player sales data using each month as a season, we get the factors displayed in the following table.

Month	Seasonality Factor	Month	Seasonality Factor
January	0.693	July	0.871
February	0.661	August	0.820
March	0.836	September	1.418
April	0.848	October	1.506
May	0.781	November	1.123
June	0.989	December	1.454

Our July trend-only prediction must be multiplied by the July seasonality factor 0.871 to take seasonal variation into account in our forecast.

$$\widehat{\text{UNITS}} = (29,523.65e^{0.069 \times 64}) \times 0.871 = 2,128,299$$

Taking seasonal variation into account, we forecast July DVD player sales to be 12.9% below what is forecasted by the exponential trend alone. This forecast is less than the trend-only forecast, but it may still be too high given our observation that July sales are always below June sales (except in 1998), and June 2002 sales were only 1,617,098 units. The seasonal pattern is not as strong as the monthly retail sales data of Example 13.1, so we may need more than a trend-and-season model to accurately forecast monthly DVD player sales.

Trend-only and trend-and-season models using indicator variables are both regression models like those used in Chapters 10 and 11. Forecasting with these models uses the same techniques as the regression-based predictions in Chapters 10 and 11. However, the prediction intervals presented in those chapters cannot typically be used with time series data. The regression models used in this section will usually have residuals that indicate our regression assumptions—particularly the assumption of independent deviations ϵ —are not appropriate for our time series data.

APPLY YOUR KNOWLEDGE

13.6 Number of Macs shipped. In Exercise 13.2, you fit a linear trend-only model to the Macs shipped time series. Starting with this trend model, incorporate seasonality factors for each quarter. With only eight quarters of data, these calculations can be done by hand (or computer).

- (a) Calculate the seasonality factor for each quarter.
- (b) Average the four seasonality factors. Is this average close to one? If so, interpret the seasonality factor for first quarters.

(c) Make a scatterplot of seasonality factor versus quarter. Connect the points to see the general pattern of seasonal variation. Also, draw a horizontal line at the average of the four seasonality factors.

13.7 JCPenney sales. In Exercise 13.3, you fit a linear trend-only model to the JCPenney sales data. Starting with this trend model, we want to incorporate seasonality factors to account for the pattern that repeats every four quarters.

- (a) Calculate the seasonality factor for each quarter.
- (b) Average the four seasonality factors. Is this average close to one? If so, interpret the seasonality factor for fourth quarters.
- (c) Make a scatterplot of seasonality factor versus quarter. Connect the points to see the general pattern of seasonal variation. Also, draw a horizontal line at the average of the four seasonality factors.

Seasonally-adjusted data

Many economic time series are seasonally-adjusted to make the overall trend in the numbers more apparent. Government agencies will often release both versions of a time series so be careful to notice whether you are analyzing seasonally-adjusted data or not when using government sources. A seasonally-adjusted time series has had each value divided by the seasonality factor corresponding to the appropriate month.

EXAMPLE 13.6 Seasonally-adjusted monthly retail sales

To calculate seasonally-adjusted monthly retail sales, we simply divide each actual sales value by the appropriate seasonality factor calculated in Example 13.4. Figure 13.7 displays the original time series and the seasonally-adjusted time series. The seasonally-adjusted time series, SALES (SA), does not show the regular ups and downs of the data that is not seasonally-adjusted, SALES (NSA). The seasonally-adjusted time series shows the overall trend and some random variation. Seasonally-adjusted data is easier to interpret because it flattens the regular peaks and valleys that we should expect in our data making it easier to identify unusual peaks and/or valleys in the data.

There are many different approaches to seasonal adjustment. The indicator variable model of Example 13.3 can be used although we won't go through the details. We could also use a trend model that is more complicated than a line. Specialized time series models can also be used to seasonally adjust a time series. Different approaches will lead to different estimates and forecasts; however, the differences will not usually be great.

APPLY YOUR KNOWLEDGE

13.8 Seasonally-adjusted DVD player data.

Case13.1

- (a) Make a scatterplot of the seasonality factors in Example 13.5 with the seasonality factor on the vertical axis and the months (in time order) on the horizontal axis. Connect the points and describe the overall pattern of seasonal variation in the DVD player sales data.
- (b) Calculate the seasonally-adjusted value of the time series for June 2002. Do the calculation by hand.

13.9 Seasonally-adjusted DVD player data.

Case13.1

- (a) Using statistical software, calculate the seasonally-adjusted time series of the DVD player sales data. Make a time plot of the original DVD player time series with the seasonally-adjusted time series superimposed (see Figure 13.7 on page xxx for an example).
- (b) Did seasonally-adjusting the DVD player sales data smooth the time series to the degree that seasonally-adjusting the sales data in Figure 13.7 did? What does this imply about the strength of the seasonal pattern in these two time series?

Looking for autocorrelation

Using a regression model with time series data as we've done in this section will often result in residuals that indicate our regression assumption of independent deviations is not appropriate. The current value of a time series may be heavily influenced by past values of the same series. A simple form of this dependence on past values is known as *autocorrelation*. In Chapter 2, we defined correlation as a measure of the strength of the linear relationship between two variables x and y (page 103). Similarly, autocorrelation measures the strength of the linear relationship between successive values of the same variable.

autocorrelation

AUTOCORRELATION

The correlation between successive values y_{t-1} and y_t of a time series is called *first-order autocorrelation* or, simply, *autocorrelation*.

The residuals from a regression model that uses time as an explanatory variable should be examined for signs of autocorrelation. If successive residuals tend to be both negative or both positive, then the residuals will have positive autocorrelation. If successive residuals tend to be of opposite sign, then the residuals will have negative autocorrelation. In either case, we have evidence that our regression assumption of independent deviations ϵ is inappropriate.

Let's examine the residuals that resulted from fitting an exponential trend to the DVD player sales data from Case 13.1.

EXAMPLE 13.7 DVD player sales

Figure 13.8 plots the residuals from our exponential trend-only model for DVD player sales. Notice the pattern:

- Starting on the left side of the plot, negative residuals tend to be followed by another negative residual.
- In the middle of the plot, positive residuals tend to be followed by another positive residual.
- On the right side of the plot, negative residuals tend to be followed by another negative residual

This pattern indicates positive autocorrelation among the residuals.

An alternative plot for detecting autocorrelation is given in Figure 13.9. Because autocorrelation measures the strength of the linear relationship between successive values in a series, plotting pairs of successive values will show a strong linear pattern when significant autocorrelation exists. Figure 13.9 plots points of the form (e_{t-1}, e_t) where e_t is the residual for time period t . A scatterplot of the “lagged” residuals e_{t-1} against the residuals e_t will show strong positive association for a series that has positive autocorrelation and a strong negative association for a series that has negative autocorrelation. If the points in a *lagged residual plot* (like Figure 13.9) are equally dispersed among the four quadrants of the scatterplot, then we do not have visual evidence of autocorrelation in our residuals.

lagged residual plot

LAGGED RESIDUAL PLOT

A scatterplot of residuals e_t against the same residuals lagged one time period; that is, we plot the points $(e_1, e_2), (e_2, e_3), \dots, (e_{n-1}, e_n)$.

A **lagged residual plot** is used to detect autocorrelation in residuals from a time series. A positive association in the points (e_{t-1}, e_t) indicates positive autocorrelation while a negative association indicates negative autocorrelation.

EXAMPLE 13.8 DVD player sales

The linear pattern with positive slope in Figure 13.9 indicates that our residuals may have positive autocorrelation. The correlation for Figure 13.9 is 0.6165. This is approximately the autocorrelation for the residuals lagged one time period. Statistical software calculates the autocorrelation to be 0.6135 using a modified version of our correlation formula from Chapter 2.

The positive autocorrelation of 0.6135 is another bit of evidence that our residuals are dependent, but is 0.6135 a large enough autocorrelation to give us confidence in our conclusion of dependence? To answer this question conclusively, we would use a statistical test for autocorrelation—the *Durbin-Watson test*. The Durbin-Watson test statistic is provided with the multiple regression output of many statistical software packages. You can consult a multiple regression text for the details of this test.⁵

Durbin-Watson test

Several methods are proposed for handling autocorrelation in a time series. You might add lagged values of your residuals as explanatory variables to a trend-only model. Other approaches suggest clever transformations of the variables in the model to adjust for the autocorrelation. Another approach is to fit a model that uses only past values of the time series itself for forecasting rather than a model that uses other explanatory variables for forecasting the time series variable.

APPLY YOUR KNOWLEDGE

13.10 DVD player sales. In Example 13.7 and 13.8, we found autocorrelation in the residuals from the exponential trend-only model for DVD player sales. The residuals from a trend-and-season model may not exhibit the same pattern of autocorrelation. (a) Calculate the predicted number of units sold using the trend-and-season model based on the seasonality factors displayed in Example 13.5. That is, calculate

$$\widehat{\text{UNITS}} = (29,523.65e^{0.069x}) \times \text{SF}$$

for each month in the time series. Be sure to use the appropriate seasonality factor (SF) for each month.

(b) Calculate the residuals for the model in part (a). That is, calculate

$$e = \text{UNITS} - \widehat{\text{UNITS}}$$

for each month in the time series.

(c) Make a time plot of the residuals. Do we have visual evidence of autocorrelation in this plot. Describe the evidence.

(d) Make a lagged residual plot and calculate the correlation between successive residuals e_{t-1} and e_t . Do we have evidence of autocorrelation? Describe the evidence.

Case 13.1

SECTION 13.1 Summary

Data collected at regular time intervals are called a **time series**. Time series often display a long-run **trend**. Some time series also display a strong, repeating **seasonal pattern**.

Regression methods can be used to model the trend and seasonal variation in a time series. Indicator variables can be used to model the seasons in a time series. Alternatively, **seasonality factors** can be used to adjust trend-only models for seasonal effects.

Seasonally-adjusted data has been divided by seasonality factors to remove the effect of seasonal variation on the time series. Government agencies typically release seasonally-adjusted data for economic time series.

Successive residuals may be linearly related. A **lagged residual plot** is a scatter-plot of residuals plotted against the residuals lagged by one time period. If residuals exhibit **autocorrelation**, then our usual regression assumptions are not appropriate; in particular, the deviations cannot be assumed to be independent. In this case, a model based on past values of the time series should be considered.

SECTION 13.1 Exercises

13.11 JCPenney sales. A linear trend-only model was fit to the JCPenney sales data in Exercise 13.3. Use this model to answer the following:

- On a time plot of the sales data, draw the least-squares line. Comment on any pattern of over-/under-prediction if we were to use this trend-only model for predicting sales.
- Using the equation of least-squares line, forecast sales for the first quarter of 2002 and for the fourth quarter of 2002.
- Which forecast in part (b) do you believe will be more accurate when compared to actual JCPenney sales? Why?

13.12 JCPenney sales. A trend-and-season model with indicator variables was fit to the JCPenney sales data in Exercise 13.5. (a) Using the equation of the trend-and-season model, forecast sales for the first quarter of 2002 and for the fourth quarter of 2002.

- Compare your forecasts to the same forecasts based on the trend-only model of Exercise 13.11.

13.13 JCPenney sales. A trend-and-season model using seasonality factors was calculated in Exercise 13.7. (a) Using the linear trend-only model and the seasonality factors, forecast sales for the first quarter of 2002 and for the fourth quarter of 2002.

- Compare your forecasts to the same forecasts based on the trend-only model of Exercise 13.11.
- Compare your forecasts to the same forecasts based on the trend-and-season model of Exercise 13.12.

13.14 JCPenney sales. The model in part (a) of Exercise 13.5 can be viewed as four models—one for each quarter—by setting the indicator variables to 0 or 1 for each of the four quarters (our seasons for this data set). Do this and record the four models in part (a).

- Write down the four models (one for each quarter) that are derived from the trend-and-season model in Exercise 13.5.
- Each of the models in part (a) is linear. What do you notice about the four

slopes?

- (c) On a time plot of the sales data, sketch each of the four lines. What geometric property do the lines possess?

13.15 Comparing models for JCPenney sales. Compare the trend-only model of Exercise 13.3 with the trend-and-season model of Exercise 13.5.

- (a) Report the value of R^2 for each model and comment on the difference.
- (b) Report the value of s for each model and comment on the difference.
- (c) Make a time plot of the original JCPenney sales figures. On this same plot, overlay both the trend-only predictions and the trend-and-season predictions.
- (d) Taking into account parts (a), (b), and (c), is the trend-and-season model a big improvement over the trend-only model?

13.16 Seasonally-adjusted JCPenney sales. In Exercise 13.7, you calculated seasonality factors for the JCPenney quarterly sales data. Using these factors, complete the following:

- (a) Calculate the seasonally-adjusted value of the time series for the fourth quarter of 2001. Do the calculation by hand.
- (b) Using statistical software, calculate the seasonally-adjusted JCPenney sales time series. Make a time plot of the original sales data with the seasonally-adjusted sales data superimposed (see Figure 13.7 on page xxx for an example).
- (c) Did seasonally-adjusting JCPenney's sales data smooth the time series to the degree that seasonally-adjusting the sales data in Figure 13.7 did? What does this imply about the strength of the seasonal pattern in these two time series?

13.17 Autocorrelation in the JCPenney time series. In Exercise 13.3, a linear trend-only model was fit to the JCPenney sales data. Using the residuals from this model, look for evidence of autocorrelation.

- (a) Make a time plot of the residuals. Describe any pattern you see in this plot.
- (b) Make a lagged residual plot and calculate the correlation between successive residuals e_{t-1} and e_t . Do we have evidence of autocorrelation?

13.18 Autocorrelation in the JCPenney time series. Repeat the previous exercise using the residuals from the trend-and-season model of Exercise 13.5.

13.19 Lagged time series plot. Similar to the lagged residual plot, a lagged time series plot is a scatterplot with y_t on the vertical axis and y_{t-1} on the horizontal axis. Figure 13.10 is a lagged time series plot of the monthly retail sales data introduced in Example 13.1.

- (a) Figure 13.10 consists of three distinct groups of points. One large group containing most of the points and two smaller groups consisting of ten points each. Given what we know about when the ups and downs occur with this time series, we can identify the two smaller groups of points. The group of ten points crossing the lower right corner are the points with January sales as the y -coordinate. The other group

of ten points has which month as the y -coordinate?

(b) The correlation between successive month's sales is 0.4573. If we exclude the two groups of 10 points identified in part (a), this correlation is 0.9206. Does this suggest strong autocorrelation in the time series?

(c) If you looked at the correlation between successive values of the seasonally-adjusted time series, would you expect the correlation to be closer to 0.4573 or 0.9206? Explain your response.

NOTES

1. The home page for the Monthly Retail Trade Survey is found at www.census.gov/mrts/www/mrtshist.html.
2. JCPenney quarterly net sales data were extracted from annual reports found at www.jcpenney.net/company/finance/finarch.htm.
3. Source: *Consumer Electronics Association (CEA)*, www.ce.org.
4. The number of Macintoshes shipped per fiscal quarter can be found in Apple's quarterly earnings press release. Follow the "Financial Releases" link at www.apple.com/find/sitemap.html to find the documents.
5. For example, details of the Durbin-Watson test for autocorrelation can be found in *Applied Linear Statistical Models* (4th edition) by Neter, Kutner, Nachtsheim, and Wasserman (ISBN 0256117365).

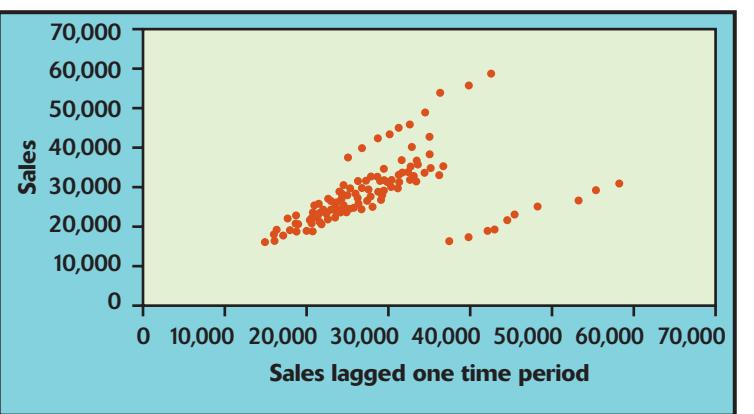


Figure 13.10 Lagged time series plot of U.S. retail sales of general merchandise stores, for Exercise 13.19.

13.2 Time Series Models

The previous section applied regression methods from Chapters 10 and 11 to time series data. A time period variable was used as the explanatory variable, and the time series was the response variable. Models like these may not satisfy our usual regression assumptions—particularly the independence assumption. Specific models have been developed to aid in the analysis of time series data when usual regression methods are not appropriate.

TIME SERIES MODELS

Time series models use past values of the time series to predict future values of the time series.

In the language of regression, a future time period's value is our “response” while one or more past time periods' values (of the same variable) are the explanatory variable(s). This approach is different from using the values of one variable x to predict the values of a second variable y . A time series model makes forecasts based on past values of the time series itself. For example, we might forecast next month's DVD player sales to be some multiple of this month's DVD player sales. Instead, we might use average DVD player sales for the last three months to predict DVD player sales this month. A wide variety of time series models exist to implement forecasting strategies like these as well as more elaborate forecasting strategies. This section will introduce a few of the most common time series models and illustrate how to use them for forecasting future values of a time series.

Autoregressive models

Can yesterday's stock price help predict today's stock price? Can last quarter's sales be used to predict this quarter's sales? Sometimes the best explanatory variables are simply past values of the response variable.

Autoregressive time series models take advantage of the linear relationship between successive values of a time series to predict future values of the series.

FIRST-ORDER AUTOREGRESSION MODEL

The **first-order autoregression model** specifies a linear relationship between successive values of the time series. The shorthand for this model is AR(1), and the equation is

$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t$$

In Figure 13.9 (page 13-18), we observed a positive, linear relationship between successive residuals from the exponential trend model fit to the DVD player sales data. Autocorrelation in the residuals from a regression model involving time-related variables indicates an AR(1) model might fit the data well.

We have two options for specifying an AR(1) model for the DVD player sales data. We can take the number of units sold per month to be the time series, or since the positive correlation in Figure 13.9 is for the residuals from the exponential trend model, we may want to take the logarithm of the number of units sold per month as the time series. In general, fitting an exponential model $y = ae^{bx}$ is the same as fitting a simple linear regression model with $\log(y)$ as the response variable instead of just y . Taking logarithms of both sides reveals a linear model based on $\log(y)$.

$$\log(y) = \log(ae^{bx}) = \log(a) + \log(e^{bx}) = \log(a) + bx$$

Figure 13.11 plots the number of units sold against the number of units sold lagged one period. Figure 13.12 plots the same points after taking logarithms. Figure 13.12 ($r = 0.95$) shows a stronger linear pattern than

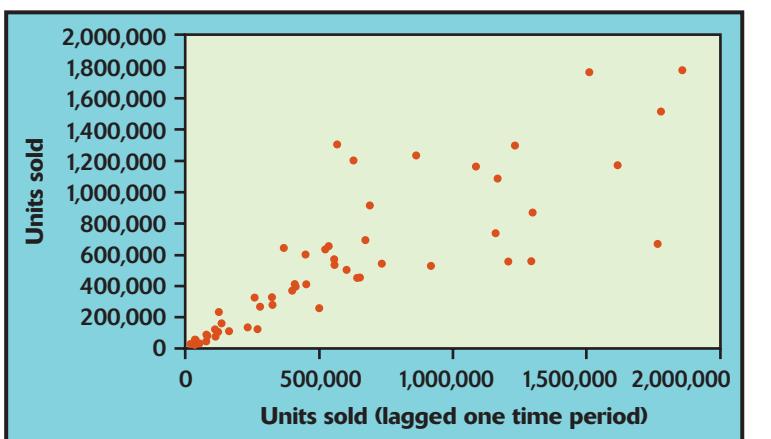


Figure 13.11 Lagged time series plot of the DVD player sales time series ($r = 0.80$).

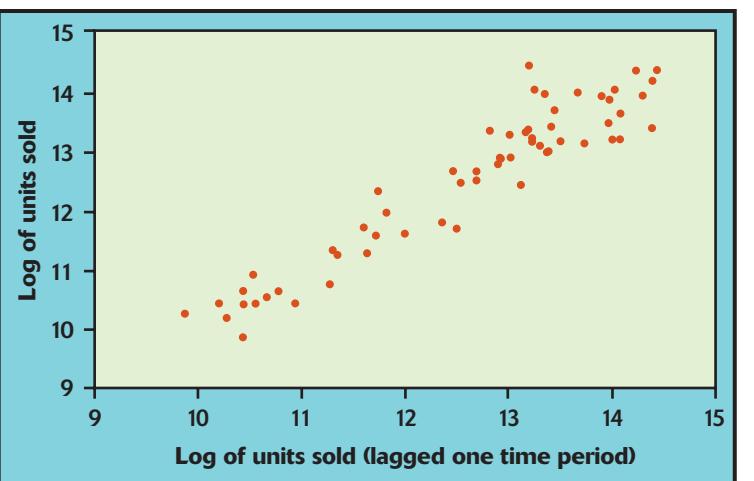


Figure 13.12 Lagged time series plot of the log of the DVD player sales time series ($r = 0.95$).

Figure 13.11 ($r = 0.80$) suggesting an AR(1) model will fit the DVD player sales data better after logarithms have been taken. We will proceed using the transformed time series in an AR(1) model.

EXAMPLE 13.9

DVD player sales

The AR(1) model for the time series $\log(y_t)$ where y_t is the number of DVD players sold in time period t is

$$\log(y_t) = \beta_0 + \beta_1 \log(y_{t-1}) + \epsilon_t$$

Statistical software calculates the fitted model to be

$$\widehat{\log(y_t)} = 0.43717 + 0.96486 \times \log(y_{t-1})$$



maximum likelihood estimation

The model for DVD player sales in Example 13.9 looks like the simple linear regression models found in Chapter 10 except that the explanatory variable is the same as the response variable but lagged one time period. Despite the similarity in the form of the models, the details that made the least-squares line the best line in the setting of Chapter 10 are no longer valid with the autoregression model in Example 13.9. Most statistical software will estimate β_0 and β_1 in an autoregressive model using a method other than least-squares. The most common alternative to least-squares estimation is called **maximum likelihood estimation** or simply MLE. The details of maximum likelihood estimation need not concern us. The general principle of MLE is to use the parameter estimates that are most in agreement with our data. The agreement between parameter estimates and our data is measured in terms of “likelihood.” Maximum likelihood estimates are the answer to the question “For what values of β_0 and β_1 are my data most likely to appear in a random sample?”

For the AR(1) model of Example 13.9, the MLE for β_1 is not too different from the least-squares estimate we would obtain by fitting a simple linear regression model to the time series. The least-squares estimate of β_1 is 0.94963 compared to the MLE of 0.96486. However, the intercept estimates differ more. The least-squares intercept estimate is 0.69 compared to the 0.43717 estimated via maximum likelihood. A difference of this magnitude will affect predictions you make with your model, so be sure to use software that was designed to estimate time series models correctly. Using simple linear regression to fit an AR(1) model is acceptable only if the time series has a mean near zero. The mean of the logarithm of the DVD player sales data is 12.5, so we need software that fits the AR(1) model using maximum likelihood estimation.

EXAMPLE 13.10**Forecasting July DVD player sales**

Our DVD player sales time series ends with June 2002 sales of 1,617,098 units. June 2002 is the 63rd month of our time series, so our notation is $y_{63} = 1,617,098$. The AR(1) model of Example 13.9 relates DVD player sales for July 2002 to DVD player sales in June 2002, so we can forecast July's sales from the reported number of units sold in June.

The model fitted in Example 13.9 is for the time series $\log(y_t)$. This creates one extra step in our task of forecasting July 2002 DVD player sales. Our model will forecast $\log(y_{64})$ and we will need to calculate the forecast of y_{64} as the final step.

First, we use the model to forecast the logarithm of July 2002 DVD player sales.

$$\begin{aligned}\widehat{\log(y_{64})} &= 0.43717 + 0.96486 \times \log(y_{63}) \\ &= 0.43717 + 0.96486 \times \log(1,617,098) \\ &= 0.43717 + 0.96486 \times (14.2961437) \\ &= 14.23095\end{aligned}$$

Second, use your calculator's e^x button (or use software) to calculate the forecast of July 2002 DVD player sales.

$$\widehat{\log(y_{64})} = 14.23095$$

$$e^{\widehat{\log(y_{64})}} = e^{14.23095}$$

$$\widehat{y}_{64} = 1,515,036$$

Our AR(1) model for DVD player sales predicts sales of over 1.5 million units for July 2002.

In Case 13.1 (page 13-7), the exponential trend model predicted sales of 2.4 million units, and in Example 13.5 (page 13-14), the exponential trend-and-season model predicted sales of 2.1 million units. We noted that July sales are always below June sales (except in 1998), but that both these forecasts are greater than the June 2002 sales of 1.6 million. Our AR(1) model's forecast is an improvement in this respect because the forecast of 1.5 million units is below June 2002 sales of 1.6 million units.

Our forecast of July 2002 DVD player sales was based on an *observed* data value—June 2002 sales is a known value in our time series. If we wish to forecast August 2002 DVD player sales, we will have to base our forecast on an *estimated* value because July 2002 sales is not a known value in our time series.

EXAMPLE 13.11**Forecasting August DVD player sales**

Following Example 13.10, we use the model to forecast the logarithm of August 2002 DVD player sales. Because our time series ends with y_{63} , the value of y_{64} is not known. In its place, we will use the value of \hat{y}_{64} calculated in Example 13.10.

$$\begin{aligned}\widehat{\log(y_{65})} &= 0.43717 + 0.96486 \times \log(\hat{y}_{64}) \\ &= 0.43717 + 0.96486 \times \log(1,515,036) \\ &= 0.43717 + 0.96486 \times (14.23095) \\ &= 14.16804\end{aligned}$$

Finally, use your calculator's e^x button or software to calculate the forecast of August 2002 DVD player sales.

$$\begin{aligned}\widehat{\log(y_{65})} &= 14.16804 \\ e^{\widehat{\log(y_{65})}} &= e^{14.16804} \\ \hat{y}_{65} &= 1,422,662\end{aligned}$$

We can continue in this manner to forecast DVD player sales for other future months.



In Chapters 10 and 11, we calculated prediction intervals for the response in our model. Associated with time series forecasts, your software may provide prediction intervals for future time periods. These intervals can be used and interpreted like the prediction intervals we saw in Chapters 10 and 11 although we will not present the details of their calculation.

The widths of time series prediction intervals generally increase as the time period of the prediction moves further beyond the end of the known time series. Predictions far into the future are subject to more uncertainty than predictions close to the end of the time series. Examples 13.10 and 13.11 demonstrate why this is true. Our forecast for July 2002 DVD player sales (y_{64}) depends on our estimated values of β_0 and β_1 and the *known* value of y_{63} . However, our forecast of y_{65} depends on estimated values of β_0 , β_1 , and y_{64} . The additional uncertainty involved in estimating y_{64} will make our prediction interval for y_{65} wider than our prediction interval for y_{64} . Statistical software provides the prediction intervals in the following table. The widths have been calculated—notice how the widths increase as we predict further into the future.

Time Period	Lower 95% PI	Upper 95% PI	Width
July 2002	678,324	3,383,804	2,705,480
August 2002	465,754	4,345,553	3,879,799
September 2002	349,161	5,133,999	4,784,838
October 2002	274,628	5,805,884	5,531,256
November 2002	223,067	6,384,087	6,161,020
December 2002	185,553	6,881,936	6,696,383
January 2003	157,263	7,309,011	7,151,748
February 2003	135,342	7,673,072	7,537,730
March 2003	117,982	7,980,825	7,862,843
April 2003	103,989	8,238,248	8,134,259
May 2003	92,539	8,450,757	8,358,218
June 2003	83,050	8,623,283	8,540,233

Autoregression models are useful when future values of a time series depend linearly on past values of the same time series. With software handling the calculation details, the concepts introduced in Chapters 10 and 11 for regression models apply equally well for autoregression models.

APPLY YOUR KNOWLEDGE

- 13.20 Existing home sales. The National Association of Realtors tracks monthly sales of existing homes in the United States. The file *ex13_020.dat* on the CD that accompanies this book has the existing home sales time series beginning in January 1968 and ending with July 2001. Use statistical software to analyze this time series.
- (a) Make a time plot of the existing home sales time series.
 - (b) Describe any important features of the time series. Is there a strong, clear trend? If so, describe it. What about seasonal variation?
 - (c) Make a lagged time series plot (see Exercise 13.19 for an example) of the existing home sales time series. Does this plot suggest that an AR(1) model is appropriate for this time series? Why or why not?
- 13.21 A model for existing home sales. For this exercise, let y_t denote the number of existing homes sold (in thousands of units) during time period t . Use statistical software and the *ex13_020.dat* data file.
- (a) Fit a simple linear regression model using y_t as the response variable and y_{t-1} as the explanatory variable. Record the estimated regression equation. Use this model to forecast August 2001 existing home sales.
 - (b) Fit an AR(1) model to the existing home sales time series. Record the estimated autoregression equation. Use this model to forecast August 2001 existing home sales.
 - (c) Compare the two estimated equations as well as the August 2001 forecasts from parts (a) and (b). State which of the two estimated models is preferred and briefly explain why.

Moving average models

The autoregression model AR(1) looks back *one* time period and uses that value of the time series in the forecast for the current time period. This

works well when consecutive values of the time series are linearly related, but not all time series fit this pattern. With some time series, our forecasts can be improved by using the *average of many* past time periods. *Moving average* models use the average of the last several values of the time series to forecast the next value.

MOVING AVERAGE FORECAST MODEL

The **moving average forecast model** uses the average of the last k values of the time series as the forecast for time period t . The equation is

$$\hat{y}_t = \frac{y_{t-1} + y_{t-2} + \cdots + y_{t-k}}{k}$$

The number of preceding values included in the moving average is called the **span** of the moving average.

Some care should be taken in choosing the span k for a moving average forecast model. As a general rule, larger spans “smooth” the time series more than smaller spans by averaging many ups and downs in each calculation. Smaller spans tend to follow the ups and downs of the time series. With seasonal data, the length of the season is often used for the value of k .

EXAMPLE 13.12

Moving averages for JCPenney sales

Exercise 13.1 (page 13-5) asked you to plot and comment on quarterly JCPenney sales data. Figure 13.13 displays the JCPenney sales time series with the moving average forecast model based on a span of $k = 4$ overlaid. Note that the seasonal pattern in the time series is not present in the moving averages. Our moving averages are a smoothed version of our original time series. The moving averages follow the general movements of the time series but not every up and down.

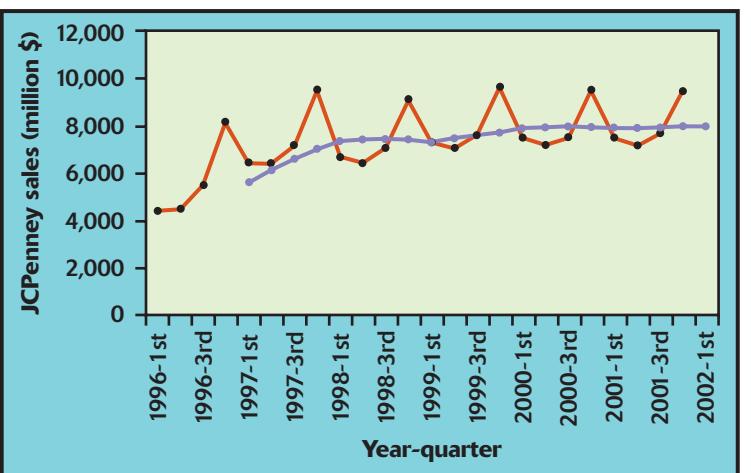
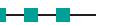


Figure 13.13 Moving averages based on a span of $k = 4$ (purple) overlaying the JCPenney quarterly sales data (red), for Example 13.12.

The JCPenney sales time series begins with the first quarter of 1996 and ends with the fourth quarter of 2001 for a total of 24 quarters. Figure 13.13 includes the forecasted value for the first quarter of 2002. The forecast is calculated as

$$\hat{y}_{25} = \frac{y_{24} + y_{23} + y_{22} + y_{21}}{4} = \frac{32,004}{4} = 8001$$



When a strong seasonal pattern is present, moving average models with the span set equal to the length of the season are similar to the trend-only models presented in Section 13.1. The moving averages will follow the long-run trend of the time series, but the ups and downs of the seasons are ignored. We would not expect the moving average model of Example 13.12 to forecast fourth-quarter JCPenney sales well.

EXAMPLE 13.13
Chicago Cubs attendance per game

Major League Baseball's Chicago Cubs have been playing their home games at Wrigley Field since 1916. The file *eg13_013.dat* on the CD that accompanies this book has the "attendance per home game" time series beginning in 1916 and ending in 2001.

Figure 13.14 displays the time series (red), moving averages based on a span of 5 years (purple), and moving averages based on a span of 20 years (black). The moving averages based on a span of 20 years highlight the general upward trend in attendance per home game at Wrigley Field without following the ups and downs in the data. The 5-year moving averages are not as smooth as the 20-year moving averages. The 5-year moving averages follow the larger ups and downs while smoothing the smaller changes in the time series.

The attendance time series has 86 observations ending with attendance per home game in 2001. Our interest is in forecasting attendance per home game for

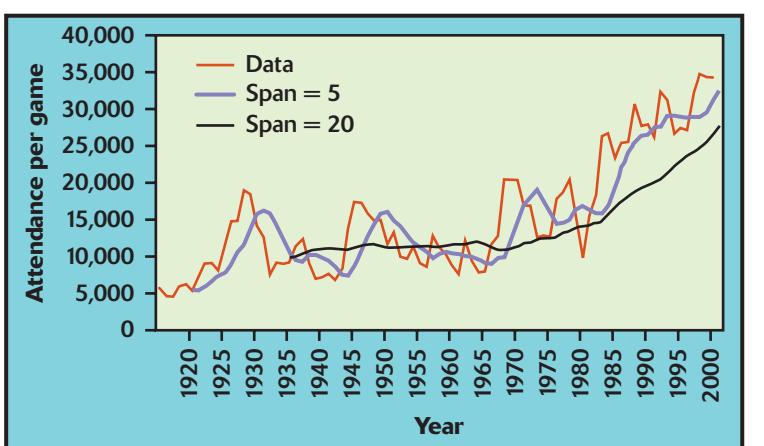


Figure 13.14 Chicago Cubs annual attendance per game time series (red) with 5-year (purple) and 20-year (black) moving averages overlaid, for Example 13.13.

2002. We will do this using both the 5-year moving averages and the 20-year moving averages.

$$\hat{y}_{87} = \frac{y_{86} + y_{85} + \cdots + y_{82}}{5} = \frac{162,522}{5} = 32,504$$

$$\hat{y}_{87} = \frac{y_{86} + y_{85} + \cdots + y_{67}}{20} = \frac{553,681}{20} = 27,684$$

The forecast based on the 20-year moving average is less than the forecast based on the 5-year moving average because the 20-year average includes the lower attendance figures of the 1980s. We might guess from looking at Figure 13.14 that the 5-year moving average forecast will be more accurate than the 20-year moving average forecast. In this case, actual 2002 attendance figures for Wrigley Field are now available. Total attendance was 2,693,096 for 81 home games, so the average attendance per home game was $2,693,096/81$ or 33,248. The 5-year moving average forecast was more accurate.



Once a new value becomes available (like 2002 attendance figures in Example 13.13), we update our time series to include the new observation. We can then calculate a forecast for the next time period (2003 attendance per home game in Example 13.13).

APPLY YOUR KNOWLEDGE

13.22 Winter wheat prices. The United States Department of Agriculture (USDA) tracks prices received by Montana farmers for winter wheat crops. The prices are tracked monthly in dollars per bushel. The file *ex13_022.dat* on the CD that accompanies this book has the wheat prices time series beginning in July 1929 and ending with October 2002 (880 months). Use statistical software to analyze this time series.

- (a) Make a time plot of the wheat prices time series.
- (b) Describe any important features of the time series. Be sure to comment on trend, seasonal patterns, and significant shifts in the series.
- (c) Calculate 12-month moving averages and plot them on your time series plot from part (a).
- (d) Calculate 120-month moving averages and plot them together with the time series from part (a) and the 12-month moving averages from part (c).
- (e) Compare the 12-month and 120-month moving averages. Which features of the wheat prices time series does each capture? Which features does each smooth?

13.23 Forecasting winter wheat prices.

- (a) Using the winter wheat prices from the previous exercise, calculate and compare the 12-month and 120-month moving average forecasts for November 2002.
- (b) Record the actual winter wheat price received by Montana farmers for November 2002 from the Web page www.nass.usda.gov/mt/economic/prices/wwheatpr.htm. Which moving average forecast model provided the most accurate forecast for November 2002?

Exponential smoothing models

Moving average forecast models appeal to our intuition. Using the average of several of the most recent data values to forecast the next value of the time series is easy to understand conceptually. However, two criticisms can be made against moving average models. First, our forecast for the next time period ignores all but the last k observations in our data set. If you have 100 observations and use a span of $k = 5$, your forecast will not use 95% of your data! Second, the data values used in our forecast are all weighted equally. In many settings, the current value of a time series depends more on the most recent value and less on past values. We may improve our forecasts if we give the most recent values greater “weight” in our forecast calculation. *Exponential smoothing* models address both these criticisms.

*exponential
smoothing*

There are several variations on the basic exponential smoothing model. We will look at the details of the *simple exponential smoothing model*, which we will refer to as, simply, the *exponential smoothing model*. More complex variations exist to handle time series with specific features, but the details of these models are beyond the scope of this chapter. We will only mention the scenarios for which these more complex models are appropriate.

EXPONENTIAL SMOOTHING MODEL

The **exponential smoothing model** uses a weighted average of the observed value y_{t-1} and the forecasted value \hat{y}_{t-1} as the forecast for time period t . The forecasting equation is

$$\hat{y}_t = w y_{t-1} + (1 - w) \hat{y}_{t-1}$$

The weight w is called the **smoothing constant** for the exponential smoothing model. The smoothing constant is a value between 0 and 1. Choosing w close to 1 puts more weight on the most recent value.

Choosing the smoothing constant w in the exponential smoothing model is similar to choosing the span k in the moving average model—both relate directly to the smoothness of the model. Smaller values of w correspond to greater smoothing of the ups and downs in the time series. Larger values of w put most of the weight on the most recent observed value, so the forecasts tend to follow the ups and downs of the series more closely.

EXAMPLE 13.14

Philip Morris returns

Table 1.10 (page 49) gives the monthly returns on Philip Morris stock for the period from June 1990 to July 2001 (134 months). Figure 13.15 displays the returns (red), the exponential smoothing model with $w = 0.5$ (purple), and the exponential smoothing model with $w = 0.1$ (black).

Using a smoothing constant of 0.5 puts more weight on the most recent values of the time series than does using a smoothing constant of 0.1. As a result, the

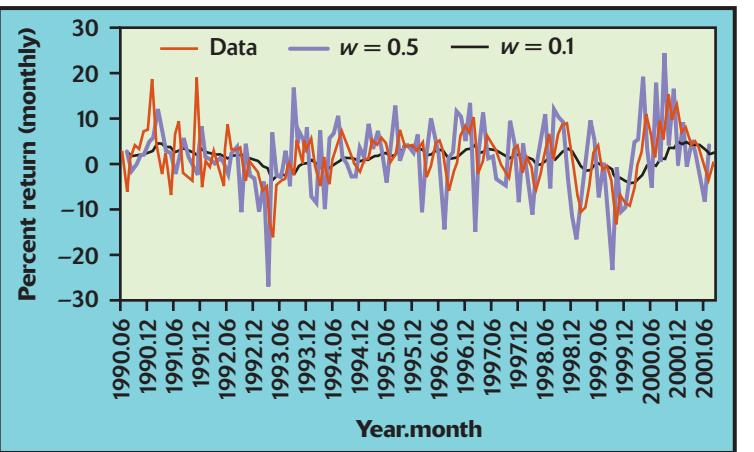


Figure 13.15 Philip Morris monthly stock returns (red) with two overlays, exponential smoothing model with $w = 0.5$ (purple) and exponential smoothing model with $w = 0.1$ (black), for Example 13.14.

purple curve tends to follow the ups and downs of the time series more closely than the smoother black curve. With a smoothing constant close to 0, the model will follow only the major changes in the time series.



A little algebra is needed to see that exponential smoothing models address the criticisms of moving average models. We will start with the forecasting equation for the exponential smoothing model and imagine forecasting the value of the time series for the time period $n + 1$ where n is the number of observed values in the time series. A specific example would be forecasting the August 2001 Philip Morris return y_{135} using the observed returns for the preceding $n = 134$ months.

$$\begin{aligned}
 \hat{y}_{n+1} &= wy_n + (1 - w)\hat{y}_n \\
 &= wy_n + (1 - w)[wy_{n-1} + (1 - w)\hat{y}_{n-1}] \\
 &= wy_n + (1 - w)wy_{n-1} + (1 - w)^2\hat{y}_{n-1} \\
 &= wy_n + (1 - w)wy_{n-1} + (1 - w)^2[wy_{n-2} + (1 - w)\hat{y}_{n-2}] \\
 &= wy_n + (1 - w)wy_{n-1} + (1 - w)^2wy_{n-2} + (1 - w)^3\hat{y}_{n-2} \\
 &\vdots \\
 &= wy_n + (1 - w)wy_{n-1} + (1 - w)^2wy_{n-2} + \cdots + (1 - w)^{n-2}wy_2 \\
 &\quad + (1 - w)^{n-1}y_1
 \end{aligned}$$

Careful substitution and multiplication reveal a version of the forecast equation showing exactly how our forecast depends on the values of the time series. First, notice that the calculation of the forecast for y_{n+1} uses *all* available values of the time series y_1, y_2, \dots, y_n , not just the most recent k values as a moving average model would. Second, the values are not equally weighted. Using a value of w close to 1 puts greater weight on the most

recent observation. In this version of the forecast equation, we also see how the model gets its name. The coefficients in the forecast model decrease exponentially in value as you read the equation from left to right (with the exception of the last coefficient, $(1 - w)^{n-1}$). Exercise 13.25 has you explore what happens to the coefficients as you change the value of the smoothing constant w .

While the second version of our forecasting equation reveals some important properties of the exponential smoothing model, it is easier to use the first version of the equation for calculating forecasts.

EXAMPLE 13.15**Forecasting Philip Morris returns**

Consider forecasting the Philip Morris August 2001 return using an exponential smoothing model with $w = 0.5$.

$$\hat{y}_{135} = 0.5 \times y_{134} + (1 - 0.5) \times \hat{y}_{134}$$

We need the forecasted value \hat{y}_{134} to finish our calculation. However, to calculate \hat{y}_{134} we will need the forecasted value of \hat{y}_{133} ! In fact, this pattern continues, and we need to calculate all past forecasts before we can calculate \hat{y}_{135} . We will calculate the first few forecasts here and leave the remaining calculations for software. The forecast for the first time period \hat{y}_1 is always taken to be the actual value y_1 of the time series in the first time period.

$$\begin{aligned}\hat{y}_2 &= 0.5 \times y_1 + (1 - 0.5) \times \hat{y}_1 \\ &= (0.5)(3) + (0.5)(3) \\ &= 3\end{aligned}$$

$$\begin{aligned}\hat{y}_3 &= 0.5 \times y_2 + (1 - 0.5) \times \hat{y}_2 \\ &= (0.5)(-5.7) + (0.5)(3) \\ &= -1.35\end{aligned}$$

$$\begin{aligned}\hat{y}_4 &= 0.5 \times y_3 + (1 - 0.5) \times \hat{y}_3 \\ &= (0.5)(1.2) + (0.5)(-1.35) \\ &= -0.075\end{aligned}$$

Software continues our calculations to arrive at a forecast for y_{134} of -3.812 . We use this value to complete our forecast calculation for y_{135} .

$$\begin{aligned}\hat{y}_{135} &= 0.5 \times y_{134} + (1 - 0.5) \times \hat{y}_{134} \\ &= (0.5)(4.2) + (0.5)(-3.812) \\ &= 0.194\end{aligned}$$

Our model forecasts a 0.194% return for Philip Morris stock in August 2001.



With the forecasted value for August 2001 from Example 13.15, calculating the forecast for September 2001 requires only one calculation.

$$\hat{y}_{136} = (0.5)(y_{135}) + (0.5)(0.194)$$

Once we observe the actual Philip Morris return for August 2001 y_{135} , we can enter that value into the forecast equation above. Updating forecasts from exponential smoothing models requires only that we keep track of last period's forecast and last period's observed value. In contrast, moving average models require that we keep track of the last k observed values of the time series. For this reason, exponential smoothing models are often preferred over moving average models especially if data storage is an issue.

The exponential smoothing model is best suited for forecasting time series with no strong trend or seasonal variation. Variations on the exponential smoothing model have been developed to handle time series with a trend (double exponential smoothing and Holt's exponential smoothing), with seasonality (seasonal exponential smoothing), and with both trend and seasonality (Winters' exponential smoothing). Your software may offer one or more of these smoothing models.

APPLY YOUR KNOWLEDGE

13.24 Philip Morris returns.

- Using the Philip Morris returns in Table 1.10 (page 49), calculate the first 4 forecasts $\hat{y}_1, \dots, \hat{y}_4$ using an exponential smoothing model with $w = 0.1$. Do not use statistical software for these calculations.
- Now, use software to fit an exponential smoothing model with $w = 0.1$. Use the forecasts provided by the software to verify your hand calculations in part (a). Are your forecasts the same as those provided by your software?
- Provide a forecast for the August 2001 Philip Morris return based on your exponential smoothing model with $w = 0.1$ and compare this to the forecast calculated in Example 13.15.
- Write down the forecast equation for the September 2001 Philip Morris return based on the exponential smoothing model with $w = 0.1$.

13.25 It's exponential. Exponential smoothing models are so named because the coefficients

$$w, (1-w)w, (1-w)^2w, \dots, (1-w)^{n-2}w$$

decrease in value exponentially. For this exercise, take $n = 11$. Use software to do the calculations.

- Calculate the coefficients for a smoothing constant of $w = 0.1$.
- Calculate the coefficients for a smoothing constant of $w = 0.5$.
- Calculate the coefficients for a smoothing constant of $w = 0.9$.
- Plot each set of coefficients from parts (a), (b), and (c). The coefficient values should be measured on the vertical axis while the horizontal axis can simply be numbered 1, 2, ..., 9, 10 for the 10 coefficients from each part. Be sure to use a different plotting symbol and/or color to distinguish the three sets of coefficients and connect the points for each set. Also, label the plot so that it is clear which curve corresponds to each value of w used.
- Describe each curve in part (d). Which curve puts more weight on the most recent value of the time series when you are calculating a forecast?

- (f) The coefficient of y_1 in the exponential smoothing model is $(1 - w)^{n-1}$. Calculate the coefficient of y_1 for each of the values of w in parts (a), (b), and (c). How do these values compare to the first 10 coefficients you calculated for each value of w ? Which value of w puts the greatest weight on y_1 when you are calculating a forecast?

BEYOND THE BASICS: SPLINE FITS

smoothing spline

Modern computing capabilities have made possible many statistical tools that would otherwise be impossible or impractical. One such tool is the general purpose **smoothing spline**. A spline curve can be fit to any (x, y) data set. For a time series, we take x to be the time variable and y to be the time series values. A spline curve is a single curve consisting of a large number of polynomial curves pieced together in an optimal manner. The more polynomials used, the more flexible and less smooth the spline curve is.

The smoothness of a spline fit is determined by the choice of a positive-valued smoothing constant similar to the choice of w in the exponential smoothing model. The spline smoothing constant is often denoted by the lowercase Greek letter lambda (λ). Choosing a value of λ close to zero results in a very flexible and not very smooth spline curve. As greater values of λ are used, the spline curve becomes less flexible and more smooth.

Figure 13.16 displays three spline curves. The red spline curve offers the least amount of smoothing for the data ($\lambda = 0.000001$). With λ nearly 0, the spline curve passes through every point. The purple spline curve is based on $\lambda = 10,000$. This spline curve follows the ups and downs of the time series without trying to pass through each point. The black spline curve offers the

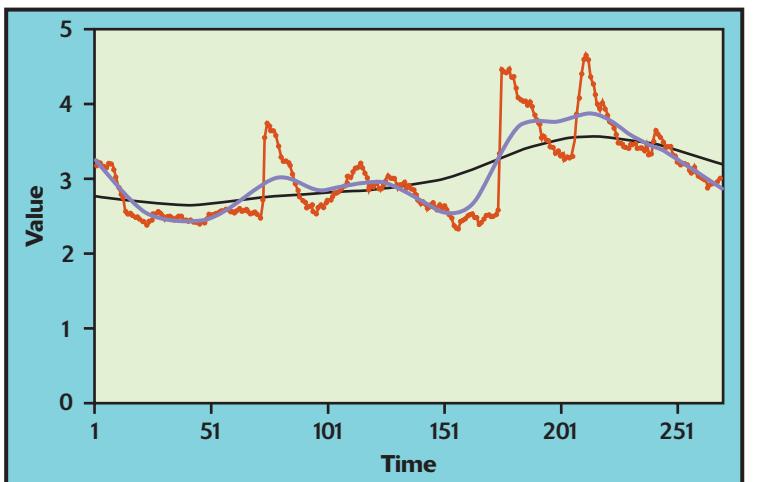


Figure 13.16 Spline curve models for three values of the smoothing constant λ : $\lambda = 0.000001$ (red), $\lambda = 10,000$ (purple), and $\lambda = 1,000,000$ (black).

highest degree of smoothing with $\lambda = 1,000,000$. The black spline curve reveals the overall pattern in the time series without following the smaller ups and downs.

Spline fits can help you quickly identify overall trends and seasonal variation as well as less regular patterns in your time series data. They are a powerful, modern exploratory data analysis tool.

SECTION 13.2 SUMMARY

- Regression models relating a time series to a time variable are not always appropriate. Specifically, the residuals from such a model may exhibit high correlation. **Time series models** use past values of the time series to forecast future values of the time series.
- The **first-order autoregression model**, AR(1), is appropriate when successive values of a time series are linearly related. The parameters of the model are often estimated using **maximum likelihood estimation** rather than least-squares estimation.
- **Moving average forecast models** use the average of the last k observed values to forecast next period's value. k is called the **span** of the moving average. Larger values of k result in a smoother model.
- The forecast equation for the **exponential smoothing model** is a weighted average of last period's observed value and last period's forecasted value. The degree of smoothing is determined by the choice of a smoothing constant w between 0 and 1. Values of w close to 0 result in a smoother model.

SECTION 13.2 EXERCISES

13.26 A closer look at oranges. Example 1.7 (page 20) looked at the trend and seasonal variation in the average monthly price of oranges. Figure 1.7 (page 21) is a time series plot of the data. The data is found in the file *fg01_007.dat* on the CD that accompanies this book.

- Make a lagged time series plot (see Exercise 13.19 for an example).
- Does the plot in part (a) suggest that an AR(1) model might be appropriate for the orange prices time series?
- Calculate the correlation between successive values of orange prices. Does the value of this correlation support your conclusion in part (b)? Why or why not?

13.27 Least-squares or maximum likelihood? Continue analyzing the orange prices data from the previous exercise. Let y_t denote the average price of oranges in time period t .

- Fit a simple linear regression model using y_t as the response variable and y_{t-1} as the explanatory variable. Record the estimated regression equation, the value of R^2 , and the regression standard error s .
- Fit an AR(1) model to the orange prices time series. Record the estimated autoregression equation. Your software should also report a model R^2

and a model standard deviation (or model standard error)—record these values also.

- (c) Compare the two estimated slopes and intercepts.
- (d) Compare the model R^2 -values and the model standard errors s . Do these values give any clear indication as to which fitting method (least-squares or maximum likelihood) is preferable?

13.28 Moving averages for orange prices.

- (a) Calculate and plot (on a single time series plot) moving averages for spans of $k = 5, 10$, and 20 .
- (b) Comment on the smoothness of each moving average model in part (a). Which model would be best for forecasting monthly ups and downs in orange prices?
- (c) Calculate and compare forecasts for January 2001 orange prices for each of the models in part (a). Which model provided the most accurate forecast? (The actual value of the orange prices time series for January 2001 is 224.2.)

13.29 Exponential smoothing for orange prices.

- (a) Calculate and plot (on a single time series plot) exponential smoothing models using smoothing constants of $\omega = 0.1, 0.5$, and 0.9 .
- (b) Comment on the smoothness of each exponential smoothing model in part (a). Which model would be best for forecasting monthly ups and downs in orange prices?
- (c) Calculate and compare forecasts for January 2001 orange prices for each of the models in part (a). Which model provided the most accurate forecast? (The actual value of the orange prices time series for January 2001 is 224.2.)
- (d) Update your data by appending the January 2001 observed value of 224.2. Now forecast the February 2001 orange price with each of the models from part (a). Which model provided the most accurate forecast? (The actual value of the orange prices time series for February 2001 is 229.6.)

**random walk
models**

13.30 A special AR(1) model. Random walk models for various financial time series are often mentioned in business literature. A simple random walk model specifies that one-period *differences* in the time series can be modeled as a constant term plus a random-error term. The equation for this random walk model is

$$y_t - y_{t-1} = \beta_0 + \epsilon$$

If we rewrite this equation solving for y_t , we get

$$y_t = \beta_0 + y_{t-1} + \epsilon$$

which is our AR(1) model with $\beta_1 = 1$. If we fit an AR(1) model and find that the β_1 estimate is close to 1, then a simple random walk model for the time series is another modeling option.

The CD that accompanies this book contains a data file named *ex13_030.dat*. Daily “USD to Euro” exchange rates beginning July 24, 2001 and ending July 23, 2002 are contained in the data file. The first exchange rate is 1.1509. This means that on July 24, 2001 a single U.S. dollar (USD) was worth 1.1509 euro (EUR).

- (a) Fit an AR(1) model to the exchange rate time series. Record the estimated forecast equation, the R^2 -value, and the model standard error estimate. What evidence do you have that a simple random walk model for these exchange rates is appropriate?
 - (b) Use the AR(1) forecast equation to predict the exchange rate for July 24, 2002.
 - (c) For a simple random walk model, the estimate of β_0 is simply the average of all one period differences $y_t - y_{t-1}$ (call this average \bar{y}_{diff}) and the forecast equation is $\hat{y}_t = y_{t-1} + \bar{y}_{\text{diff}}$. Calculate the one-period differences and their average and use these to provide a “random walk forecast” for the exchange rate on July 24, 2002. Compare the random walk forecast to the AR(1) forecast.
 - (d) You can get the actual exchange rate for July 24, 2002 at the Web site www.oanda.com/convert/fxhistory. Compare both forecasts to the actual exchange rate. Which model was more accurate?
- 13.31 **Moving averages for exchange rates.** Use statistical software to fit moving average models to the exchange rate data in the file *ex13_030.dat* on the CD that accompanies this book.
- (a) What would the moving average forecast equation be if we used a span of $k = 1$? (Your response to this part is not specific to the exchange rate data.)
 - (b) Choose a value for k that is larger than 1 and less than 6 and calculate the corresponding moving averages. Be sure to report what value of k you chose.
 - (c) Choose a value of k that is larger than 35 and less than 365 and calculate the corresponding moving averages. Be sure to report what value of k you chose.
 - (d) Plot both sets of moving averages on a time series plot of the exchange rates. Be sure to label the plot clearly. Comment on the smoothness of both sets of moving averages.
- 13.32 **Exchange rate moving averages continued.**
- (a) Using a span of $k = 1$, forecast the USD to Euro exchange rate on July 24, 2002.
 - (b) Use your model from part (b) of the previous exercise to forecast the USD to euro exchange rate on July 24, 2002.
 - (c) Use your model from part (c) of the previous exercise to forecast the USD to euro exchange rate on July 24, 2002.
 - (d) What would our moving average forecast equation be if we used a span of $k = 365$? What would this model forecast for July 24, 2002?
 - (e) You can get the actual exchange rate for July 24, 2002 at the Web site www.oanda.com/convert/fxhistory. Compare these four forecasts to the actual exchange rate. Which model was closest to the actual exchange rate? Which model performed the worst on this forecast?
- 13.33 **Exponential smoothing of exchange rates.** Use statistical software to fit exponential smoothing models to the exchange rate data in the file *ex13_030.dat* on the CD that accompanies this book.
- (a) What would the exponential smoothing forecast equation be if we used a smoothing constant of $w = 1$? What about $w = 0$? (Your response to this part is not specific to the exchange rate data.)

- (b) Choose a value for w such that $0 < w < 0.3$ and fit the corresponding exponential smoothing model. Be sure to report what value of w you chose.
- (c) Choose a value of w such that $0.7 < w < 1$ and fit the corresponding exponential smoothing model. Be sure to report what value of w you chose.
- (d) Plot both exponential smoothing models on a time series plot of the exchange rates. Be sure to label the plot clearly. Comment on the smoothness of each model.
- (e) Using the models from parts (b) and (c), forecast the July 24, 2002 exchange rate and compare both these forecasts to the actual exchange rate on that day. (You can get the actual exchange rate for July 24, 2002 at the Web site www.oanda.com/convert/fxhistory.)

STATISTICS IN SUMMARY

Standard regression methods like those described in Chapters 10 and 11 can be used with time series data with time as the explanatory variable and the time series as the response variable. These methods can help model both trend and seasonality in a time series although the models will generally not satisfy the regression assumptions needed for inference. Special time series models relate the current period's value to past values of the time series. These models are often referred to as "smoothing models." We rely on statistical software for fitting time series models and forecasting future values using these models. Here are the specific skills you should develop from studying this chapter.

A. TRENDS

1. Identify a long-run trend in a time series plot.
2. Use software to fit a regression model to the long-run trend.
3. Use a regression model to forecast future values of a time series based on a trend-only model.

B. SEASONS

1. Identify seasonal variation in a time series plot.
2. Model seasonal variation using indicator variables in a regression model.
3. Model seasonal variation using seasonality factors.
4. Forecast future values of a time series taking into account trend and seasonality.
5. Recognize seasonally adjusted data in a time series plot.

C. TIME SERIES MODELS

1. Identify autocorrelation in a lagged residual plot or a lagged time series plot.
2. Use software to calculate the autocorrelation of a time series.

TABLE A Table entry for z is under the standard normal curve to the left of z .

TABLE A	Standard normal probabilities									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

TABLE A Table entry for z is under the standard normal curve to the left of z .

TABLE B Random digits

Line								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335
106	68417	35013	15529	72765	85089	57067	50211	47487
107	82739	57890	20807	47511	81676	55300	94383	14893
108	60940	72024	17868	24943	61790	90656	87964	18883
109	36009	19365	15412	39638	85453	46816	83485	41979
110	38448	48789	18338	24697	39364	42006	76688	08708
111	81486	69487	60513	09297	00412	71238	27649	39950
112	59636	88804	04634	71197	19352	73089	84898	45785
113	62568	70206	40325	03699	71080	22553	11486	11776
114	45149	32992	75730	66280	03819	56202	02938	70915
115	61041	77684	94322	24709	73698	14526	31893	32592
116	14459	26056	31424	80371	65103	62253	50490	61181
117	38167	98532	62183	70632	23417	26185	41448	75532
118	73190	32533	04470	29669	84407	90785	65956	86382
119	95857	07118	87664	92099	58806	66979	98624	84826
120	35476	55972	39421	65850	04266	35435	43742	11937
121	71487	09984	29077	14863	61683	47052	62224	51025
122	13873	81598	95052	90908	73592	75186	87136	95761
123	54580	81507	27102	56027	55892	33063	41842	81868
124	71035	09001	43367	49497	72719	96758	27611	91596
125	96746	12149	37823	71868	18442	35119	62103	39244
126	96927	19931	36089	74192	77567	88741	48409	41903
127	43909	99477	25330	64359	40085	16925	85117	36071
128	15689	14227	06565	14374	13352	49367	81982	87209
129	36759	58984	68288	22913	18638	54303	00795	08727
130	69051	64817	87174	09517	84534	06489	87201	97245
131	05007	16632	81194	14873	04197	85576	45195	96565
132	68732	55259	84292	08796	43165	93739	31685	97150
133	45740	41807	65561	33302	07051	93623	18132	09547
134	27816	78416	18329	21337	35213	37741	04312	68508
135	66925	55658	39100	78458	11206	19876	87151	31260
136	08421	44753	77377	28744	75592	08563	79140	92454
137	53645	66812	61421	47836	12609	15373	98481	14592
138	66831	68908	40772	21558	47781	33586	79177	06928
139	55588	99404	70708	41098	43563	56934	48394	51719
140	12975	13258	13048	45144	72321	81940	00360	02428
141	96767	35964	23822	96012	94591	65194	50842	53372
142	72829	50232	97892	63408	77919	44575	24870	04178
143	88565	42628	17797	49376	61762	16953	88604	12724
144	62964	88145	83083	69453	46109	59505	69680	00900
145	19687	12633	57857	95806	09931	02150	43163	58636
146	37609	59057	66967	83401	60705	02384	90597	93600
147	54973	86278	88737	74351	47500	84552	19909	67181
148	00694	05977	19664	65441	20903	62371	22725	53340
149	71546	05233	53946	68743	72460	27601	45403	88692
150	07511	88915	41267	16853	84569	79367	32337	03316

TABLE B Random digits (*continued*)

Line								
151	03802	29341	29264	80198	12371	13121	54969	43912
152	77320	35030	77519	41109	98296	18984	60869	12349
153	07886	56866	39648	69290	03600	05376	58958	22720
154	87065	74133	21117	70595	22791	67306	28420	52067
155	42090	09628	54035	93879	98441	04606	27381	82637
156	55494	67690	88131	81800	11188	28552	25752	21953
157	16698	30406	96587	65985	07165	50148	16201	86792
158	16297	07626	68683	45335	34377	72941	41764	77038
159	22897	17467	17638	70043	36243	13008	83993	22869
160	98163	45944	34210	64158	76971	27689	82926	75957
161	43400	25831	06283	22138	16043	15706	73345	26238
162	97341	46254	88153	62336	21112	35574	99271	45297
163	64578	67197	28310	90341	37531	63890	52630	76315
164	11022	79124	49525	63078	17229	32165	01343	21394
165	81232	43939	23840	05995	84589	06788	76358	26622
166	36843	84798	51167	44728	20554	55538	27647	32708
167	84329	80081	69516	78934	14293	92478	16479	26974
168	27788	85789	41592	74472	96773	27090	24954	41474
169	99224	00850	43737	75202	44753	63236	14260	73686
170	38075	73239	52555	46342	13365	02182	30443	53229
171	87368	49451	55771	48343	51236	18522	73670	23212
172	40512	00681	44282	47178	08139	78693	34715	75606
173	81636	57578	54286	27216	58758	80358	84115	84568
174	26411	94292	06340	97762	37033	85968	94165	46514
175	80011	09937	57195	33906	94831	10056	42211	65491
176	92813	87503	63494	71379	76550	45984	05481	50830
177	70348	72871	63419	57363	29685	43090	18763	31714
178	24005	52114	26224	39078	80798	15220	43186	00976
179	85063	55810	10470	08029	30025	29734	61181	72090
180	11532	73186	92541	06915	72954	10167	12142	26492
181	59618	03914	05208	84088	20426	39004	84582	87317
182	92965	50837	39921	84661	82514	81899	24565	60874
183	85116	27684	14597	85747	01596	25889	41998	15635
184	15106	10411	90221	49377	44369	28185	80959	76355
185	03638	31589	07871	25792	85823	55400	56026	12193
186	97971	48932	45792	63993	95635	28753	46069	84635
187	49345	18305	76213	82390	77412	97401	50650	71755
188	87370	88099	89695	87633	76987	85503	26257	51736
189	88296	95670	74932	65317	93848	43988	47597	83044
190	79485	92200	99401	54473	34336	82786	05457	60343
191	40830	24979	23333	37619	56227	95941	59494	86539
192	32006	76302	81221	00693	95197	75044	46596	11628
193	37569	85187	44692	50706	53161	69027	88389	60313
194	56680	79003	23361	67094	15019	63261	24543	52884
195	05172	08100	22316	54495	60005	29532	18433	18057
196	74782	27005	03894	98038	20627	40307	47317	92759
197	85288	93264	61409	03404	09649	55937	60843	66167
198	68309	12060	14762	58002	03716	81968	57934	32624
199	26461	88346	52430	60906	74216	96263	69296	90107
200	42672	67680	42376	95023	82744	03971	96560	55148

TABLE C Binomial probabilities

TABLE C**Binomial probabilities (continued)**

		Entry is $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$									
		p									
n	k	.10	.15	.20	.25	.30	.35	.40	.45	.50	
2	0	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500	
	1	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000	
	2	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500	
3	0	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250	
	1	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750	
	2	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750	
	3	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250	
4	0	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625	
	1	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500	
	2	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750	
	3	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500	
	4	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625	
5	0	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0313	
	1	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1563	
	2	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125	
	3	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125	
	4	.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562	
	5	.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312		
6	0	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156	
	1	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938	
	2	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344	
	3	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125	
	4	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344	
	5	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0937	
	6		.0001	.0002	.0007	.0018	.0041	.0083	.0156		
7	0	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078	
	1	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547	
	2	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641	
	3	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734	
	4	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734	
	5	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641	
	6		.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547	
	7			.0001	.0002	.0006	.0016	.0037	.0078		
8	0	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039	
	1	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0313	
	2	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094	
	3	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188	
	4	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734	
	5	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188	
	6		.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094	
	7			.0001	.0004	.0012	.0033	.0079	.0164	.0312	
	8				.0001	.0002	.0007	.0017	.0039		

TABLE C Binomial probabilities (*continued*)

TABLE C Binomial probabilities (*continued*)

TABLE C Binomial probabilities (*continued*)

TABLE D Table entry for p and C is the critical value t^* with probability p lying to its right and probability c lying between $-t^*$ and t^*

TABLE D *t* distribution critical values

TABLE E Table entry for p is the critical value F^* with probability p lying to its right

		Degrees of freedom in the numerator									
		p	1	2	3	4	5	6	7	8	9
Degrees of freedom in the denominator	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5	
	.001	405284	500000	540379	562500	576405	585937	592873	598144	602284	
	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39	
Degrees of freedom in the denominator	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86	
	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	
Degrees of freedom in the denominator	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	
	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	
	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	
Degrees of freedom in the denominator	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	

TABLE E Table entry for p is the critical value F^* with probability p lying to its right

TABLE E <i>F</i> critical values (continued)										
Degrees of freedom in the numerator										
10	12	15	20	25	30	40	50	60	120	1000
60.19	60.71	61.22	61.74	62.05	62.26	62.53	62.69	62.79	63.06	63.30
241.88	243.91	245.95	248.01	249.26	250.10	251.14	251.77	252.20	253.25	254.19
968.63	976.71	984.87	993.10	998.08	1001.4	1005.6	1008.1	1009.8	1014.0	1017.7
6055.8	6106.3	6157.3	6208.7	6239.8	6260.6	6286.8	6302.5	6313.0	6339.4	6362.7
605621	610668	615764	620908	624017	626099	628712	630285	631337	633972	636301
9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.47	9.48	9.49
19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.48	19.49	19.49
39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.48	39.49	39.50
99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.48	99.49	99.50
999.40	999.42	999.43	999.45	999.46	999.47	999.47	999.48	999.48	999.49	999.50
5.23	5.22	5.20	5.18	5.17	5.17	5.16	5.15	5.15	5.14	5.13
8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.58	8.57	8.55	8.53
14.42	14.34	14.25	14.17	14.12	14.08	14.04	14.01	13.99	13.95	13.91
27.23	27.05	26.87	26.69	26.58	26.50	26.41	26.35	26.32	26.22	26.14
129.25	128.32	127.37	126.42	125.84	125.45	124.96	124.66	124.47	123.97	123.53
3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.80	3.79	3.78	3.76
5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.70	5.69	5.66	5.63
8.84	8.75	8.66	8.56	8.50	8.46	8.41	8.38	8.36	8.31	8.26
14.55	14.37	14.20	14.02	13.91	13.84	13.75	13.69	13.65	13.56	13.47
48.05	47.41	46.76	46.10	45.70	45.43	45.09	44.88	44.75	44.40	44.09
3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.12	3.11
4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.44	4.43	4.40	4.37
6.62	6.52	6.43	6.33	6.27	6.23	6.18	6.14	6.12	6.07	6.02
10.05	9.89	9.72	9.55	9.45	9.38	9.29	9.24	9.20	9.11	9.03
26.92	26.42	25.91	25.39	25.08	24.87	24.60	24.44	24.33	24.06	23.82
2.94	2.90	2.87	2.84	2.81	2.80	2.78	2.77	2.76	2.74	2.72
4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.75	3.74	3.70	3.67
5.46	5.37	5.27	5.17	5.11	5.07	5.01	4.98	4.96	4.90	4.86
7.87	7.72	7.56	7.40	7.30	7.23	7.14	7.09	7.06	6.97	6.89
18.41	17.99	17.56	17.12	16.85	16.67	16.44	16.31	16.21	15.98	15.77
2.70	2.67	2.63	2.59	2.57	2.56	2.54	2.52	2.51	2.49	2.47
3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.32	3.30	3.27	3.23
4.76	4.67	4.57	4.47	4.40	4.36	4.31	4.28	4.25	4.20	4.15
6.62	6.47	6.31	6.16	6.06	5.99	5.91	5.86	5.82	5.74	5.66
14.08	13.71	13.32	12.93	12.69	12.53	12.33	12.20	12.12	11.91	11.72

TABLE E Table entry for p is the critical value F^* with probability p lying to its right

		Degrees of freedom in the numerator									
		<i>p</i>	1	2	3	4	5	6	7	8	9
Degrees of freedom in the denominator	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	
	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	
Degrees of freedom in the denominator	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	
	.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	
	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	
	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	
	.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	
	.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	
	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	
Degrees of freedom in the denominator	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	
	.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	
	.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	
	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	
	.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	
	.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	
	.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	
	.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	
Degrees of freedom in the denominator	.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	
	.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	
	.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	
	.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	
	.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	
	.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	
	.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	
	.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	
	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	
Degrees of freedom in the denominator	.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	
	.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	
	.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	
	.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	
	.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	
	.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	
	.010	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	
	.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	

TABLE E Table entry for p is the critical value F^* with probability p lying to its right

TABLE E <i>F</i> critical values (continued)											
Degrees of freedom in the numerator											
10	12	15	20	25	30	40	50	60	120	1000	
2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.35	2.34	2.32	2.30	
3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.02	3.01	2.97	2.93	
4.30	4.20	4.10	4.00	3.94	3.89	3.84	3.81	3.78	3.73	3.68	
5.81	5.67	5.52	5.36	5.26	5.20	5.12	5.07	5.03	4.95	4.87	
11.54	11.19	10.84	10.48	10.26	10.11	9.92	9.80	9.73	9.53	9.36	
2.42	2.38	2.34	2.30	2.27	2.25	2.23	2.22	2.21	2.18	2.16	
3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.80	2.79	2.75	2.71	
3.96	3.87	3.77	3.67	3.60	3.56	3.51	3.47	3.45	3.39	3.34	
5.26	5.11	4.96	4.81	4.71	4.65	4.57	4.52	4.48	4.40	4.32	
9.89	9.57	9.24	8.90	8.69	8.55	8.37	8.26	8.19	8.00	7.84	
2.32	2.28	2.24	2.20	2.17	2.16	2.13	2.12	2.11	2.08	2.06	
2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.64	2.62	2.58	2.54	
3.72	3.62	3.52	3.42	3.35	3.31	3.26	3.22	3.20	3.14	3.09	
4.85	4.71	4.56	4.41	4.31	4.25	4.17	4.12	4.08	4.00	3.92	
8.75	8.45	8.13	7.80	7.60	7.47	7.30	7.19	7.12	6.94	6.78	
2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.04	2.03	2.00	1.98	
2.85	2.79	2.72	2.65	2.60	2.57	2.53	2.51	2.49	2.45	2.41	
3.53	3.43	3.33	3.23	3.16	3.12	3.06	3.03	3.00	2.94	2.89	
4.54	4.40	4.25	4.10	4.01	3.94	3.86	3.81	3.78	3.69	3.61	
7.92	7.63	7.32	7.01	6.81	6.68	6.52	6.42	6.35	6.18	6.02	
2.19	2.15	2.10	2.06	2.03	2.01	1.99	1.97	1.96	1.93	1.91	
2.75	2.69	2.62	2.54	2.50	2.47	2.43	2.40	2.38	2.34	2.30	
3.37	3.28	3.18	3.07	3.01	2.96	2.91	2.87	2.85	2.79	2.73	
4.30	4.16	4.01	3.86	3.76	3.70	3.62	3.57	3.54	3.45	3.37	
7.29	7.00	6.71	6.40	6.22	6.09	5.93	5.83	5.76	5.59	5.44	
2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.92	1.90	1.88	1.85	
2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.31	2.30	2.25	2.21	
3.25	3.15	3.05	2.95	2.88	2.84	2.78	2.74	2.72	2.66	2.60	
4.10	3.96	3.82	3.66	3.57	3.51	3.43	3.38	3.34	3.25	3.18	
6.80	6.52	6.23	5.93	5.75	5.63	5.47	5.37	5.30	5.14	4.99	
2.10	2.05	2.01	1.96	1.93	1.91	1.89	1.87	1.86	1.83	1.80	
2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.24	2.22	2.18	2.14	
3.15	3.05	2.95	2.84	2.78	2.73	2.67	2.64	2.61	2.55	2.50	
3.94	3.80	3.66	3.51	3.41	3.35	3.27	3.22	3.18	3.09	3.02	
6.40	6.13	5.85	5.56	5.38	5.25	5.10	5.00	4.94	4.77	4.62	
2.06	2.02	1.97	1.92	1.89	1.87	1.85	1.83	1.82	1.79	1.76	
2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.18	2.16	2.11	2.07	
3.06	2.96	2.86	2.76	2.69	2.64	2.59	2.55	2.52	2.46	2.40	
3.80	3.67	3.52	3.37	3.28	3.21	3.13	3.08	3.05	2.96	2.88	
6.08	5.81	5.54	5.25	5.07	4.95	4.80	4.70	4.64	4.47	4.33	
2.03	1.99	1.94	1.89	1.86	1.84	1.81	1.79	1.78	1.75	1.72	
2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.12	2.11	2.06	2.02	
2.99	2.89	2.79	2.68	2.61	2.57	2.51	2.47	2.45	2.38	2.32	
3.69	3.55	3.41	3.26	3.16	3.10	3.02	2.97	2.93	2.84	2.76	
5.81	5.55	5.27	4.99	4.82	4.70	4.54	4.45	4.39	4.23	4.08	
2.00	1.96	1.91	1.86	1.83	1.81	1.78	1.76	1.75	1.72	1.69	
2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.08	2.06	2.01	1.97	
2.92	2.82	2.72	2.62	2.55	2.50	2.44	2.41	2.38	2.32	2.26	
3.59	3.46	3.31	3.16	3.07	3.00	2.92	2.87	2.83	2.75	2.66	
5.58	5.32	5.05	4.78	4.60	4.48	4.33	4.24	4.18	4.02	3.87	

TABLE E Table entry for p is the critical value F^* with probability p lying to its right

TABLE E		<i>F</i> critical values (continued)								
Degrees of freedom in the denominator	<i>p</i>	Degrees of freedom in the numerator								
		1	2	3	4	5	6	7	8	9
18	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56
	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
19	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39
	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
20	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24
	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
21	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11
	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
22	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99
	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89
23	.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80
	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
24	.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71
	.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
25	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
	.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64
	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
27	.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
	.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57

TABLE E Table entry for p is the critical value F^* with probability p lying to its right

TABLE E <i>F</i> critical values (continued)										
Degrees of freedom in the numerator										
10	12	15	20	25	30	40	50	60	120	1000
1.98	1.93	1.89	1.84	1.80	1.78	1.75	1.74	1.72	1.69	1.66
2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.04	2.02	1.97	1.92
2.87	2.77	2.67	2.56	2.49	2.44	2.38	2.35	2.32	2.26	2.20
3.51	3.37	3.23	3.08	2.98	2.92	2.84	2.78	2.75	2.66	2.58
5.39	5.13	4.87	4.59	4.42	4.30	4.15	4.06	4.00	3.84	3.69
1.96	1.91	1.86	1.81	1.78	1.76	1.73	1.71	1.70	1.67	1.64
2.38	2.31	2.23	2.16	2.11	2.07	2.03	2.00	1.98	1.93	1.88
2.82	2.72	2.62	2.51	2.44	2.39	2.33	2.30	2.27	2.20	2.14
3.43	3.30	3.15	3.00	2.91	2.84	2.76	2.71	2.67	2.58	2.50
5.22	4.97	4.70	4.43	4.26	4.14	3.99	3.90	3.84	3.68	3.53
1.94	1.89	1.84	1.79	1.76	1.74	1.71	1.69	1.68	1.64	1.61
2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.97	1.95	1.90	1.85
2.77	2.68	2.57	2.46	2.40	2.35	2.29	2.25	2.22	2.16	2.09
3.37	3.23	3.09	2.94	2.84	2.78	2.69	2.64	2.61	2.52	2.43
5.08	4.82	4.56	4.29	4.12	4.00	3.86	3.77	3.70	3.54	3.40
1.92	1.87	1.83	1.78	1.74	1.72	1.69	1.67	1.66	1.62	1.59
2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.94	1.92	1.87	1.82
2.73	2.64	2.53	2.42	2.36	2.31	2.25	2.21	2.18	2.11	2.05
3.31	3.17	3.03	2.88	2.79	2.72	2.64	2.58	2.55	2.46	2.37
4.95	4.70	4.44	4.17	4.00	3.88	3.74	3.64	3.58	3.42	3.28
1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.65	1.64	1.60	1.57
2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.91	1.89	1.84	1.79
2.70	2.60	2.50	2.39	2.32	2.27	2.21	2.17	2.14	2.08	2.01
3.26	3.12	2.98	2.83	2.73	2.67	2.58	2.53	2.50	2.40	2.32
4.83	4.58	4.33	4.06	3.89	3.78	3.63	3.54	3.48	3.32	3.17
1.89	1.84	1.80	1.74	1.71	1.69	1.66	1.64	1.62	1.59	1.55
2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.88	1.86	1.81	1.76
2.67	2.57	2.47	2.36	2.29	2.24	2.18	2.14	2.11	2.04	1.98
3.21	3.07	2.93	2.78	2.69	2.62	2.54	2.48	2.45	2.35	2.27
4.73	4.48	4.23	3.96	3.79	3.68	3.53	3.44	3.38	3.22	3.08
1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.62	1.61	1.57	1.54
2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.86	1.84	1.79	1.74
2.64	2.54	2.44	2.33	2.26	2.21	2.15	2.11	2.08	2.01	1.94
3.17	3.03	2.89	2.74	2.64	2.58	2.49	2.44	2.40	2.31	2.22
4.64	4.39	4.14	3.87	3.71	3.59	3.45	3.36	3.29	3.14	2.99
1.87	1.82	1.77	1.72	1.68	1.66	1.63	1.61	1.59	1.56	1.52
2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.84	1.82	1.77	1.72
2.61	2.51	2.41	2.30	2.23	2.18	2.12	2.08	2.05	1.98	1.91
3.13	2.99	2.85	2.70	2.60	2.54	2.45	2.40	2.36	2.27	2.18
4.56	4.31	4.06	3.79	3.63	3.52	3.37	3.28	3.22	3.06	2.91
1.86	1.81	1.76	1.71	1.67	1.65	1.61	1.59	1.58	1.54	1.51
2.22	2.15	2.07	1.99	1.94	1.90	1.85	1.82	1.80	1.75	1.70
2.59	2.49	2.39	2.28	2.21	2.16	2.09	2.05	2.03	1.95	1.89
3.09	2.96	2.81	2.66	2.57	2.50	2.42	2.36	2.33	2.23	2.14
4.48	4.24	3.99	3.72	3.56	3.44	3.30	3.21	3.15	2.99	2.84
1.85	1.80	1.75	1.70	1.66	1.64	1.60	1.58	1.57	1.53	1.50
2.20	2.13	2.06	1.97	1.92	1.88	1.84	1.81	1.79	1.73	1.68
2.57	2.47	2.36	2.25	2.18	2.13	2.07	2.03	2.00	1.93	1.86
3.06	2.93	2.78	2.63	2.54	2.47	2.38	2.33	2.29	2.20	2.11
4.41	4.17	3.92	3.66	3.49	3.38	3.23	3.14	3.08	2.92	2.78

TABLE E Table entry for p is the critical value F^* with probability p lying to its right

		Degrees of freedom in the numerator									
		<i>p</i>	1	2	3	4	5	6	7	8	9
Degrees of freedom in the denominator	28	.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
		.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
		.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
		.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
		.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50
	29	.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
		.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
		.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
		.010	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
		.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45
Degrees of freedom in the denominator	30	.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
		.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
		.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
		.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
		.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39
	40	.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
		.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
		.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45
		.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
		.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02
Degrees of freedom in the denominator	50	.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76
		.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
		.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38
		.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
		.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82
	60	.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
		.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
		.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
		.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
		.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69
Degrees of freedom in the denominator	100	.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69
		.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
		.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24
		.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
		.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44
	200	.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66
		.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
		.025	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18
		.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50
		.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26
Degrees of freedom in the denominator	1000	.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64
		.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89
		.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13
		.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43
		.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13

TABLE E Table entry for p is the critical value F^* with probability p lying to its right

TABLE E <i>F</i> critical values (continued)										
Degrees of freedom in the numerator										
10	12	15	20	25	30	40	50	60	120	1000
1.84	1.79	1.74	1.69	1.65	1.63	1.59	1.57	1.56	1.52	1.48
2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.79	1.77	1.71	1.66
2.55	2.45	2.34	2.23	2.16	2.11	2.05	2.01	1.98	1.91	1.84
3.03	2.90	2.75	2.60	2.51	2.44	2.35	2.30	2.26	2.17	2.08
4.35	4.11	3.86	3.60	3.43	3.32	3.18	3.09	3.02	2.86	2.72
1.83	1.78	1.73	1.68	1.64	1.62	1.58	1.56	1.55	1.51	1.47
2.18	2.10	2.03	1.94	1.89	1.85	1.81	1.77	1.75	1.70	1.65
2.53	2.43	2.32	2.21	2.14	2.09	2.03	1.99	1.96	1.89	1.82
3.00	2.87	2.73	2.57	2.48	2.41	2.33	2.27	2.23	2.14	2.05
4.29	4.05	3.80	3.54	3.38	3.27	3.12	3.03	2.97	2.81	2.66
1.82	1.77	1.72	1.67	1.63	1.61	1.57	1.55	1.54	1.50	1.46
2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.76	1.74	1.68	1.63
2.51	2.41	2.31	2.20	2.12	2.07	2.01	1.97	1.94	1.87	1.80
2.98	2.84	2.70	2.55	2.45	2.39	2.30	2.25	2.21	2.11	2.02
4.24	4.00	3.75	3.49	3.33	3.22	3.07	2.98	2.92	2.76	2.61
1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.48	1.47	1.42	1.38
2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.66	1.64	1.58	1.52
2.39	2.29	2.18	2.07	1.99	1.94	1.88	1.83	1.80	1.72	1.65
2.80	2.66	2.52	2.37	2.27	2.20	2.11	2.06	2.02	1.92	1.82
3.87	3.64	3.40	3.14	2.98	2.87	2.73	2.64	2.57	2.41	2.25
1.73	1.68	1.63	1.57	1.53	1.50	1.46	1.44	1.42	1.38	1.33
2.03	1.95	1.87	1.78	1.73	1.69	1.63	1.60	1.58	1.51	1.45
2.32	2.22	2.11	1.99	1.92	1.87	1.80	1.75	1.72	1.64	1.56
2.70	2.56	2.42	2.27	2.17	2.10	2.01	1.95	1.91	1.80	1.70
3.67	3.44	3.20	2.95	2.79	2.68	2.53	2.44	2.38	2.21	2.05
1.71	1.66	1.60	1.54	1.50	1.48	1.44	1.41	1.40	1.35	1.30
1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.56	1.53	1.47	1.40
2.27	2.17	2.06	1.94	1.87	1.82	1.74	1.70	1.67	1.58	1.49
2.63	2.50	2.35	2.20	2.10	2.03	1.94	1.88	1.84	1.73	1.62
3.54	3.32	3.08	2.83	2.67	2.55	2.41	2.32	2.25	2.08	1.92
1.66	1.61	1.56	1.49	1.45	1.42	1.38	1.35	1.34	1.28	1.22
1.93	1.85	1.77	1.68	1.62	1.57	1.52	1.48	1.45	1.38	1.30
2.18	2.08	1.97	1.85	1.77	1.71	1.64	1.59	1.56	1.46	1.36
2.50	2.37	2.22	2.07	1.97	1.89	1.80	1.74	1.69	1.57	1.45
3.30	3.07	2.84	2.59	2.43	2.32	2.17	2.08	2.01	1.83	1.64
1.63	1.58	1.52	1.46	1.41	1.38	1.34	1.31	1.29	1.23	1.16
1.88	1.80	1.72	1.62	1.56	1.52	1.46	1.41	1.39	1.30	1.21
2.11	2.01	1.90	1.78	1.70	1.64	1.56	1.51	1.47	1.37	1.25
2.41	2.27	2.13	1.97	1.87	1.79	1.69	1.63	1.58	1.45	1.30
3.12	2.90	2.67	2.42	2.26	2.15	2.00	1.90	1.83	1.64	1.43
1.61	1.55	1.49	1.43	1.38	1.35	1.30	1.27	1.25	1.18	1.08
1.84	1.76	1.68	1.58	1.52	1.47	1.41	1.36	1.33	1.24	1.11
2.06	1.96	1.85	1.72	1.64	1.58	1.50	1.45	1.41	1.29	1.13
2.34	2.20	2.06	1.90	1.79	1.72	1.61	1.54	1.50	1.35	1.16
2.99	2.77	2.54	2.30	2.14	2.02	1.87	1.77	1.69	1.49	1.22

TABLE F Table entry for p is the critical value (χ^2)^{*} with probability p lying to its right

		Tail probability p											
		.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
df													
1		1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2		2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3		4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4		5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5		6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6		7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7		9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8		10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9		11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10		12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11		13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12		14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13		15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14		17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15		18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16		19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17		20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18		21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19		22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20		23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21		24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22		26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23		27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24		28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25		29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26		30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27		31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28		32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29		33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30		34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40		45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50		56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60		66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80		88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
100		109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2