# PART 1

# Data

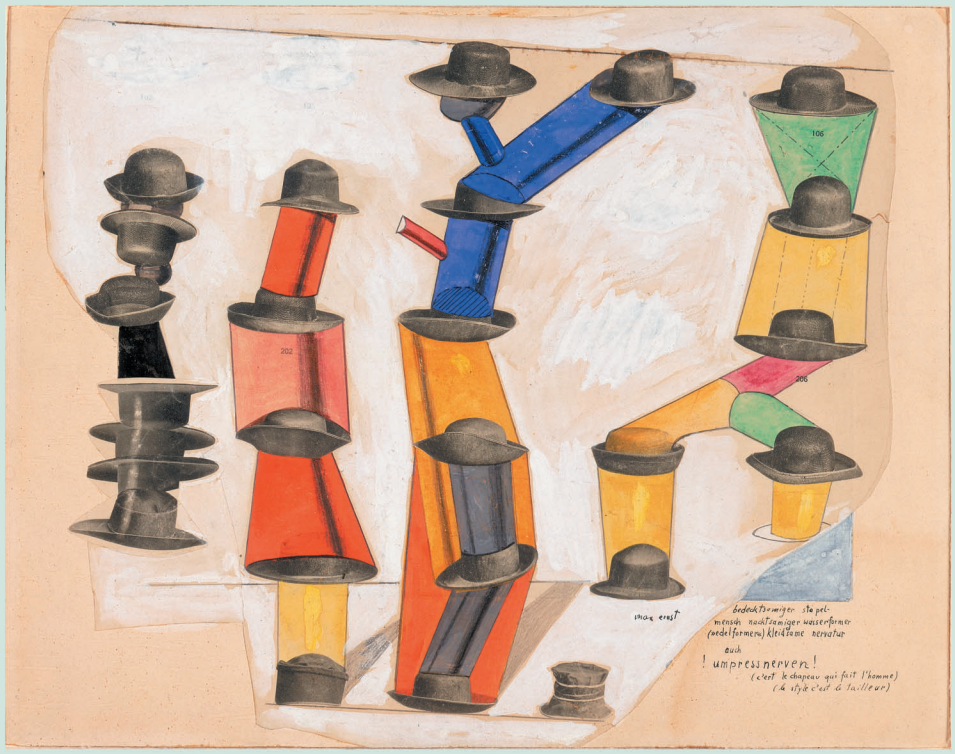
## Prelude: The Case of the Missing Vans

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.

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 vast sums from GM, used most of each loan to pay off the previous loan, 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]

# CHAPTER 1




Max Ernst, *The Hat Makes the Man*, 1920.

# Looking at Data—Distributions

# Introduction

Statistics uses data to gain understanding. Understanding comes from combining knowledge of the background of the data with the ability to use graphs and calculations to see what the data tell us. The numbers in a medical study, for example, mean little without some knowledge of the goals of the study and of what blood pressure, heart rate, and other measurements contribute to those goals. On the other hand, measurements from the study's several hundred subjects are of little value to even the most knowledgeable medical expert until the tools of statistics organize, display, and summarize them. We begin our study of statistics by mastering the art of examining data.

## Variables

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 animals or things.

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. **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 those for whom we actually have data?
2. **Who?** What **individuals** do the data describe? **How many** individuals appear in the data?
3. **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? Weights, for example, might be recorded in pounds, in thousands of pounds, or in kilograms.

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

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 | Job Type | Age | Gender | Race | Salary |
| 2 | Cedillo, Jose | Technical | 27 | Male | White | 52,300 |
| 3 | Chambers, Tonia | Management | 42 | Female | Black | 112,800 |
| 4 | Childers, Amanda | Clerical | 39 | Female | White | 27,500 |
| 5 | Chen, Huabang | Technical | 51 | Male | Asian | 83,600 |
| 6 | | | | | | |

Ready NUM

case

Each row records data on one individual. 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 in years and salary in dollars are quantitative variables.

spreadsheet

Most statistical software uses this format to enter data—each row is an individual, and each column is a variable. This data set appears in a **spreadsheet** program that has rows and columns ready for your use. Spreadsheets are commonly used to enter and transmit data. Most statistical software can read data from the major spreadsheet programs.

## 1.1 Displaying Distributions with Graphs

exploratory data analysis

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 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 more complete description.

## Graphs for categorical variables

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. For example, how well educated are 30-something young adults? Here is the distribution of the highest level of education for people aged 25 to 34 years:[2]

| Education | Count (millions) | Percent |
|---|---|---|
| Less than high school | 4.7 | 12.3 |
| High school graduate | 11.8 | 30.7 |
| Some college | 10.9 | 28.3 |
| Bachelor's degree | 8.5 | 22.1 |
| Advanced degree | 2.5 | 6.6 |

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 five education groups. The heights

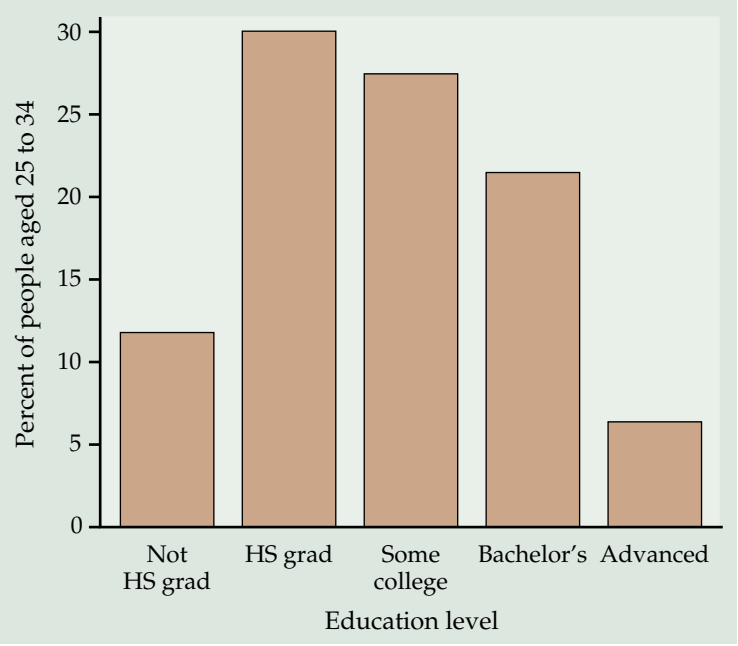



FIGURE 1.1(a) Bar graph of the educational attainment of people aged 25 to 34 years.

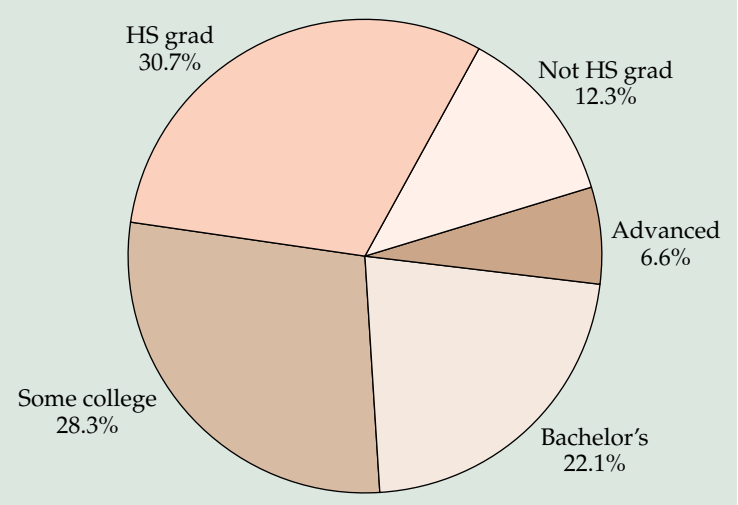



FIGURE 1.1(b) Pie chart of the education data.

pie chart

of the bars show the counts in the five categories. The **pie chart** in Figure 1.1(b) helps us see what part of the whole each group forms. For example, the "HS grad" slice makes up 30.7% of the pie because 30.7% of young adults have only a high school education. Because pie charts lack a scale, we have added the percents to the labels for the wedges. Pie charts require that you 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.

Bar graphs and pie charts help an audience grasp the 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 educational attainment without a graph. We can move on to quantitative variables, where graphs are essential tools.

## Measuring the speed of light

How do we begin to examine the distribution of a single quantitative variable? Let's look at an important scientific study.

EXAMPLE 1.2

Light travels fast, but it is not transmitted instantaneously. Light takes over a second to reach us from the moon and over 10 billion years to reach us from the most distant objects yet observed in the expanding universe. Because radio and radar also travel at the speed of light, an accurate value for that speed is important in communicating with astronauts and orbiting satellites. An accurate value for the speed of light is also important to computer designers because electrical signals travel at light speed. The first reasonably accurate measurements of the speed of light were made over 100 years ago by A. A. Michelson and Simon Newcomb. Table 1.1 contains 66 measurements made by Newcomb between July and September 1882.[3]

TABLE 1.1 Newcomb's measurements of the passage time of light

| | | | | | |
|---|---|---|---|---|---|
| 28 | 22 | 36 | 26 | 28 | 28 |
| 26 | 24 | 32 | 30 | 27 | 24 |
| 33 | 21 | 36 | 32 | 31 | 25 |
| 24 | 25 | 28 | 36 | 27 | 32 |
| 34 | 30 | 25 | 26 | 26 | 25 |
| −44 | 23 | 21 | 30 | 33 | 29 |
| 27 | 29 | 28 | 22 | 26 | 27 |
| 16 | 31 | 29 | 36 | 32 | 28 |
| 40 | 19 | 37 | 23 | 32 | 29 |
| −2 | 24 | 25 | 27 | 24 | 16 |
| 29 | 20 | 28 | 27 | 39 | 23 |

A set of numbers such as those in Table 1.1 is meaningless without some background information. The *individuals* here are Newcomb's 66 repetitions of his experiment. We need to know exactly *what variable* he measured, and in *what units*. Newcomb measured the time in seconds that a light signal took to pass from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back, a total distance of about 7400 meters. Just as you can compute the speed of a car from the time required to drive a mile, Newcomb could compute the speed of light from the passage time. Newcomb's first measurement of the passage time of light was 0.000024828 second, or 24,828 nanoseconds. (There are $10^9$ nanoseconds in a second.) The entries in Table 1.1 record only the deviation from 24,800 nanoseconds. The table entry 28 is short for the original 0.000024828. The entry −44 stands for 24,756 nanoseconds.

## Measurement

instrument

A detailed answer to the question "What variable is being measured?" requires a description of the **instrument** used to make the measurement. Then we must judge whether the variable measured is appropriate for our purpose. This judgment often requires expert knowledge of the particular field of study. For example, Newcomb invented a novel and complicated apparatus to measure the passage time of light. We accept the judgment of physicists that this instrument is appropriate for its intended task and more accurate than earlier instruments.

Newcomb measured a clearly defined and easily understood variable, the time light takes to travel a fixed distance. Questions about measurement are often harder to answer in the social and behavioral sciences than in the physical sciences. We can agree that a tape measure is an appropriate way to measure a person's height. But how shall we measure her intelligence? Questionnaires and interview forms as well as tape measures are measuring instruments. A psychologist wishing to measure "general intelligence" might use the Wechsler Adult Intelligence Scale (WAIS). The WAIS is a standard "IQ

test" that asks subjects to solve a large number of problems. The appropriateness of this instrument as a measure of intelligence is not accepted by all psychologists, many of whom doubt that there is such a thing as "general intelligence." Because questions about measurement usually require knowledge of the particular field of study, we will say little about them.

rate

You should nonetheless always ask yourself if a particular variable really does measure what you want it to. Often, for example, the **rate** at which something occurs is a more meaningful measure than a simple **count** of occurrences.

EXAMPLE 1.3

The government's Fatal Accident Reporting System says that 20,818 passenger car occupants were killed in crashes in 1999, but only 2472 motorcycle riders were killed. Does this mean that motorcycles are safer than cars? Not at all—there are many more cars than motorcycles, and they are driven many more miles.

A better measure of the dangers of driving is a *rate*, the number of deaths divided by the number of miles driven. In 1999, passenger cars drove 1,566,979,000,000 miles in the United States. Because this number is so large, it is usual to give the death rate per 100 million miles driven. The 1999 fatality rate for passenger cars is

$$\frac{\text{traffic deaths}}{\text{hundreds of millions of miles driven}} = \frac{20,818}{15,669.79} = 1.3$$

Motorcycles drove only 10,584,000,000 miles in 1999. There were 23.4 deaths per 100 million miles. Motorcycles are, as we surmise, much more dangerous than cars.[4]

## Variation

Look again at Newcomb's observations in Table 1.1. They are not all the same. Why should this be? After all, each observation records the travel time of light over the same path, measured by a skilled observer using the same apparatus each time. Newcomb knew that careful measurements almost always vary. The environment of every measurement is slightly different. The apparatus changes a bit with the temperature, the density of the atmosphere changes from day to day, and so on. Newcomb did his best to eliminate the sources of variation that he could anticipate, but even the most careful experiments produce variable results. Newcomb took many measurements rather than just one because the average of 66 passage times should be less variable than the result of a single measurement. The average does not depend on the specific conditions at just one time.

Putting ourselves in Newcomb's place, we might rush to compute the average passage time, convert this time to a new and better estimate of the speed of light, and publish the result. The temptation to do a routine calculation and announce an answer arises whenever data have been collected to answer a specific question. Because computers are very good at routine calculations, succumbing to temptation is the easy road. The first step toward statistical sophistication is to resist the temptation to calculate without thinking. Since variation is surely present in any set of data, we must first examine the nature of the variation.

The distribution of a quantitative variable describes the pattern of variation of its values. Just as pie charts and bar graphs display the distribution of categorical variables, distributions of quantitative variables are best displayed graphically.

## Stemplots

A *stemplot* (also called a stem-and-leaf plot) gives a quick picture of the shape of a distribution while including the actual numerical values in the graph. Stemplots work best for small numbers of observations that are all greater than 0.

### 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 1.4

Here are the numbers of home runs that Babe Ruth hit in each of his 15 years with the New York Yankees, 1920 to 1934:

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 54 | 59 | 35 | 41 | 46 | 25 | 47 | 60 | 54 | 46 | 49 | 46 | 41 | 34 | 22 |

To make a stemplot for these data, we use the first digits as stems and the second digits as leaves. Thus 54, for example, appears as a leaf 4 on the stem 5. Figure 1.2 shows the steps in making the plot. We see from the completed stemplot that Ruth hit 54 home runs in two different years.

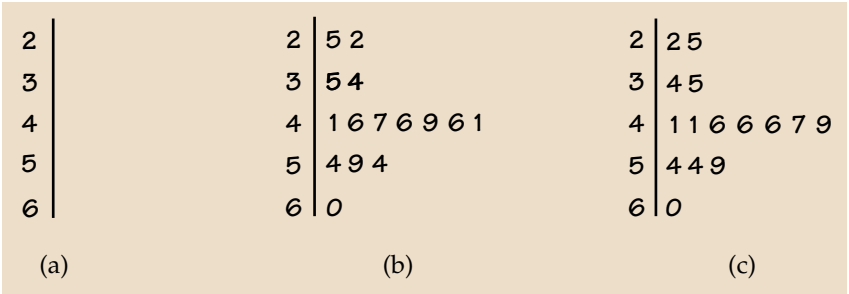



FIGURE 1.2 Making a stemplot of the data in Example 1.4. (a) Write the stems. (b) Go through the data writing each leaf on the proper stem. (c) Arrange the leaves on each stem in order out from the stem.

back-to-back stemplot

When you wish to compare two related distributions, a **back-to-back stemplot** with common stems is useful. The leaves on each side are ordered out from the common stem. The leading contemporary power hitter is Mark McGwire of the St. Louis Cardinals, who retired at the end of the 2001 season.[5] Here is a back-to-back stemplot comparing the home run data for Ruth and McGwire.

| Ruth | | McGwire |
|---|---|---|
| | 0 | 9 9 |
| | 1 | |
| 5 2 | 2 | 2 9 |
| 5 4 | 3 | 2 2 3 9 9 |
| 9 7 6 6 6 1 1 | 4 | 2 9 |
| 9 4 4 | 5 | 2 8 |
| 0 | 6 | 5 |
| | 7 | 0 |

McGwire matches up well with the legendary Ruth. The two low observations are explained by an injury in 1993 and a players' strike in 1994. In fact, McGwire was injured in 2000 and 2001 as well. He hit 32 and 29 home runs in half seasons in those years, respectable full-season totals for most players.

Stemplots do not work well for large data sets, where each stem must hold a large number of leaves. Fortunately, there are several modifications of the basic stemplot that are helpful when plotting the distribution of a moderate number of observations. You can increase the number of stems in a plot by **splitting each stem** into two: one with leaves 0 to 4 and the other with leaves 5 through 9. When the observed values have many digits, it is often best to **round** the numbers to just a few digits before making a stemplot. You must use your judgment in deciding whether to split stems and whether to round, though statistical software will often make these choices for you. Remember that the purpose of a stemplot is to display the shape of a distribution. If a stemplot has fewer than about five stems, you should usually split the stems unless there are few observations. If many stems have no leaves or only one leaf, rounding may help. Here is an example that makes use of both of these modifications.

splitting stems
rounding

EXAMPLE 1.5

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 |

To make a stemplot of this distribution, we first round the purchases to the nearest dollar. We can then use tens of dollars as our stems and dollars as leaves. This gives us the single-digit leaves that a stemplot requires. For example, we round 12.69 to

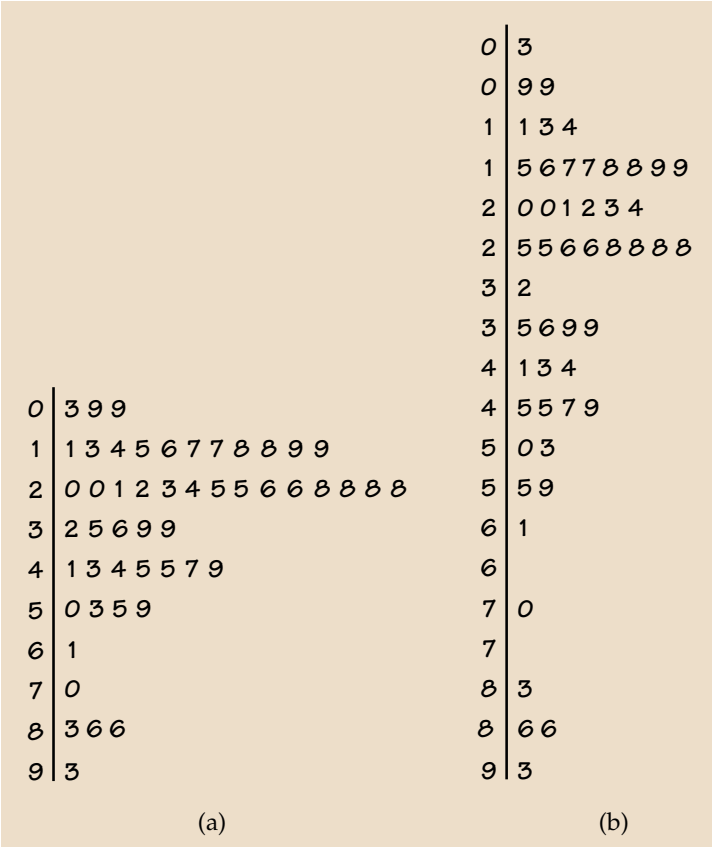



FIGURE 1.3 Stemplots of the amounts spent by 50 supermarket shoppers (in dollars), with and without splitting the stems.

13 and write it as a leaf of 3 on the 1 stem. Figure 1.3(a) displays the stemplot. For comparison, we split each stem to make the stemplot in Figure 1.3(b). Both graphs show the distribution shape, but the split stems provide more detail.

## Examining distributions

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 follows.

### 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 distribution 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 half taking larger values. We can describe the spread of a distribution by giving the range between the *smallest and largest values*. To see the shape of a distribution more clearly, turn the stemplot on its side so that the larger values lie to the right. Some things to look for in describing shape are:

modes
unimodal

- Does the distribution have one or several major peaks, called **modes**? A distribution with one major peak is called **unimodal**.

symmetric
skewed

- Is it approximately symmetric or is it skewed in one direction? A distribution is **symmetric** if the values smaller and larger than its midpoint are mirror images of each other. It is **skewed to the right** if the right tail (larger values) is much longer than the left tail (smaller values).

EXAMPLE 1.6

The distribution of Babe Ruth's home run counts (Figure 1.2) is symmetric and unimodal. Note that when describing data we do not insist on exact symmetry. Our goal is an approximate overall description, not mathematical exactness. The midpoint is 46 home runs (count up to the eighth of the 15 values), and the range is from 22 to 60. There are no outliers. In particular, Ruth's record 60 home runs in one season does not stand outside his overall pattern.

EXAMPLE 1.7

Mark McGwire's 70 home run season is also not an outlier. There are, however, two low outliers in McGwire's career, two seasons in which he hit only 9 home runs. You should search for an explanation for any outlier. Sometimes outliers point to errors made in recording the data. In other cases, the outlying observation may be caused by equipment failure or other unusual circumstances. In McGwire's case, the low outliers are not complete seasons: he played only 27 games in 1993 because of an injury and only 47 games in 1994 due to a players' strike. McGwire's home run counts, like many distributions with only a few observations, have an irregular shape. We might describe the distribution, omitting the outliers, as skewed to the right with midpoint around 40 and range 22 to 70.

EXAMPLE 1.8

The distribution of supermarket spending (Figure 1.3) is skewed to the right. There are many moderate values and a few quite large dollar amounts. The direction of the long tail (to the right) determines the direction of the skewness. Because the largest amounts do not fall outside the overall skewed pattern, we do not call them outliers. The midpoint of the distribution (count up 25 entries in the stemplot) is \$28. The range is \$3 to \$93. The shape revealed by the stemplot will interest the store management. If the big spenders can be studied in more detail, steps to attract them may be profitable. Finally, this distribution is unimodal. Although there are several minor bulges in the stemplot, there is only one major peak, containing the 22 shoppers who spent between \$15 and \$28.

The right-skewed shape in Figure 1.3 is common among distributions of money amounts. There are many moderately priced houses, for example, but the few very expensive mansions give the distribution of house prices a strong right skew.

## Histograms

Stemplots display the actual values of the observations. This feature makes stemplots awkward for large data sets. Moreover, the picture presented by a stemplot divides the observations into groups (stems) determined by the number system rather than by judgment. Histograms do not have these limitations. A **histogram** breaks the range of values of a variable into intervals and displays only the count or percent of the observations that fall into each interval. You can choose any convenient number of intervals, but you should always choose intervals of equal width. Histograms are slower to construct by hand than stemplots and do not display the actual values observed. For these reasons we prefer stemplots for small data sets. The construction of a histogram is best shown by example. Any statistical software package will of course make a histogram for you.

histogram

EXAMPLE 1.9

One of the most striking findings of the 2000 census was the growth of the Hispanic population in the United States. Table 1.2 presents the percent of adults (age 18 and over) in each of the 50 states who identified themselves in the 2000 census as "Spanish/Hispanic/Latino."[6] To make a histogram of this distribution, proceed as follows:

**1.** Divide the range of the data into classes of equal width. The data in Table 1.2 range from 0.6 to 38.7, so we choose as our classes

$$\begin{aligned} 0.0 < \text{percent Hispanic} &\le 5.0 \\ 5.0 < \text{percent Hispanic} &\le 10.0 \\ &\vdots \\ 35.0 < \text{percent Hispanic} &\le 40.0 \end{aligned}$$

TABLE 1.2 Percent of Hispanics in the adult population, by state (2000)

| State | Percent | State | Percent | State | Percent |
|---|---|---|---|---|---|
| Alabama | 1.5 | Louisiana | 2.4 | Ohio | 1.6 |
| Alaska | 3.6 | Maine | 0.6 | Oklahoma | 4.3 |
| Arizona | 21.3 | Maryland | 4.0 | Oregon | 6.5 |
| Arkansas | 2.8 | Massachusetts | 5.6 | Pennsylvania | 2.6 |
| California | 28.1 | Michigan | 2.7 | Rhode Island | 7.0 |
| Colorado | 14.9 | Minnesota | 2.4 | South Carolina | 2.2 |
| Connecticut | 8.0 | Mississippi | 1.3 | South Dakota | 1.2 |
| Delaware | 4.0 | Missouri | 1.8 | Tennessee | 2.0 |
| Florida | 16.1 | Montana | 1.6 | Texas | 28.6 |
| Georgia | 5.0 | Nebraska | 4.5 | Utah | 8.1 |
| Hawaii | 5.7 | Nevada | 16.7 | Vermont | 0.8 |
| Idaho | 6.4 | New Hampshire | 1.4 | Virginia | 4.2 |
| Illinois | 10.7 | New Jersey | 12.3 | Washington | 6.0 |
| Indiana | 3.1 | New Mexico | 38.7 | West Virginia | 0.6 |
| Iowa | 2.3 | New York | 13.8 | Wisconsin | 2.9 |
| Kansas | 5.8 | North Carolina | 4.3 | Wyoming | 5.5 |
| Kentucky | 1.3 | North Dakota | 1.0 | | |

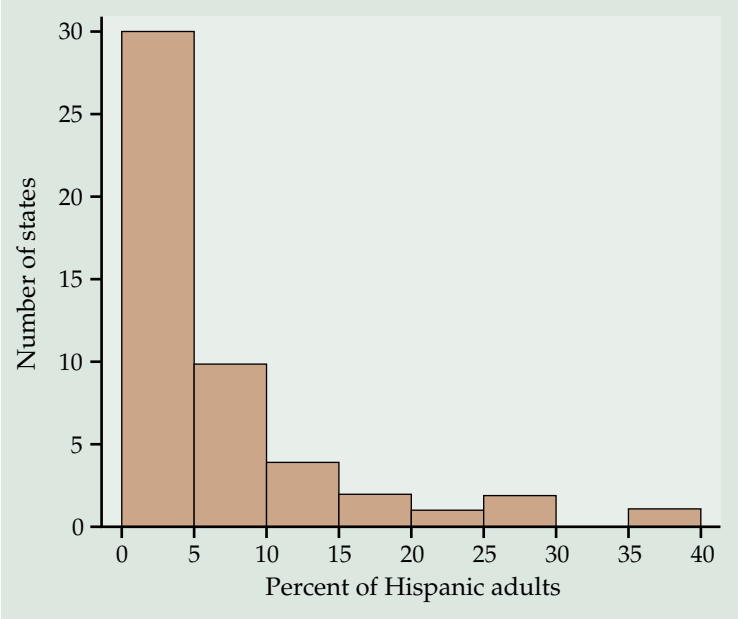



FIGURE 1.4 Histogram of the percent of each state's adult residents who identified themselves as Hispanic in the 2000 census.

Be sure to specify the classes precisely so that each individual falls into exactly one class. A state with 5.0% Hispanic residents would fall into the first class, but 5.1% falls into the second.

frequency

**2.** Count the number of individuals in each class. These counts are called **frequencies**, and a table of frequencies for all classes is a **frequency table**.

| Class | Count | Percent | Class | Count | Percent |
|---|---|---|---|---|---|
| 0.1 to 5.0 | 30 | 60.0 | 20.1 to 25.0 | 1 | 2.0 |
| 5.1 to 10.0 | 10 | 20.0 | 25.1 to 30.0 | 2 | 4.0 |
| 10.1 to 15.0 | 4 | 8.0 | 30.1 to 35.0 | 0 | 0.0 |
| 15.1 to 20.0 | 2 | 4.0 | 35.1 to 40.0 | 1 | 2.0 |

**3.** Draw the histogram. First mark the scale for the variable whose distribution you are displaying on the horizontal axis. That's the percent of adults who are Hispanic. The scale runs from 0 to 40 because that is the span of the classes we chose. 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 1.4 is our histogram.

Large sets of data are often reported in the form of frequency tables when it is not practical to publish the individual observations. In addition to the frequency (count) for each class, the fraction or percent of the observations

relative frequency

that fall in each class can be reported. These fractions are sometimes called **relative frequencies**. Example 1.9 reports both frequencies and relative frequencies. When you report statistical results, however, it is clearer to use the nontechnical terms "number" and "percent." A histogram of relative frequencies looks just like a frequency histogram such as Figure 1.4. Simply relabel the vertical scale to read in percents. Use histograms of relative frequencies for comparing several distributions with different numbers of observations.

Our eyes respond to the *area* of the bars in a histogram. 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 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. Statistical software will choose the classes for you. The software's choice is usually a good one, but you can change it if you want. Our strategy for describing distributions applies to histograms as well as to stemplots.

EXAMPLE 1.10

What does the histogram in Figure 1.4 tell us? **Shape**: The distribution is *skewed to the right* with a *single peak* at the left. Most states have no more than 10% Hispanics, but some states have much higher percents. **Center**: About half the states have fewer than 4% Hispanics among their adult residents and half have more. So the midpoint of the distribution is close to 4%. **Spread**: The spread is from about 0% to 40%, but only one state falls above 30%.

**Outliers**: The one state that stands out is New Mexico, with 38.7% Hispanics. Other states (Arizona, California, Texas) with high percents are part of the long right tail of the distribution, but New Mexico is 10% higher than any other state. Outliers often point to the special nature of some observations. New Mexico is heavily Hispanic by history and location.

Although histograms resemble bar graphs, their details and uses are distinct. A histogram shows the distribution of frequencies or relative frequencies among the values of a single variable. A bar graph compares the size of different items. The horizontal axis of a bar graph need not have any measurement

## Exploring the Web

The U.S. government offers a river of information online. Want to know the percent of each state's residents who are Hispanic? Go to the Census Bureau, `www.census.gov`, and look at the results of the 2000 census. What about fatal traffic accidents? Visit the home of the Fatal Accident Reporting System at the National Highway Traffic Safety Administration, `www.nhtsa.dot.gov`. Doctoral degrees earned by women? The National Center for Education Statistics has the facts at `nces.ed.gov`. Best of all, there is one gateway to the more than 70 federal Web sites: FedStats, at `www.fedstats.gov`.

scale but simply identifies the items being compared. Draw bar graphs with blank space between the bars to separate the items being compared. Draw histograms with no space, to indicate that all values of the variable are covered. Some software not primarily intended for statistics, such as spreadsheet programs, will draw histograms as if they were bar graphs, with space between the bars. You can tell the software to eliminate the space in order to produce a proper histogram.

## Dealing with outliers

What of Simon Newcomb and the speed of light? Scientists know that careful measurements will vary, but they hope to see a symmetric unimodal distribution. If there is no systematic bias in the measurements, the best estimate of the true value of the measured quantity is the center of the distribution. The histogram in Figure 1.5 shows that Newcomb's data do have a symmetric unimodal distribution—but there are two low outliers that stand outside this pattern. What should he do?

The handling of outliers is a matter for judgment. Outliers can point to the unusual nature of some observations, such as Mark McGwire's incomplete seasons in Example 1.7. Newcomb's data, however, are just repeated measurements of the same quantity, so the two outliers ($-44$ and $-2$ in Table 1.1) were disturbing. Outliers in research studies can result from equipment failure or errors in recording the data. If equipment failure or some other abnormal condition caused the outlier, we can simply delete it. We may be able to correct recording errors by looking at the original data record. When no cause is found, the decision is difficult. The outlier may be evidence of an extraordinary occurrence or just a mistake that went unnoticed. Newcomb decided to drop the worse outlier ($-44$) and to keep the other. He based his estimate of

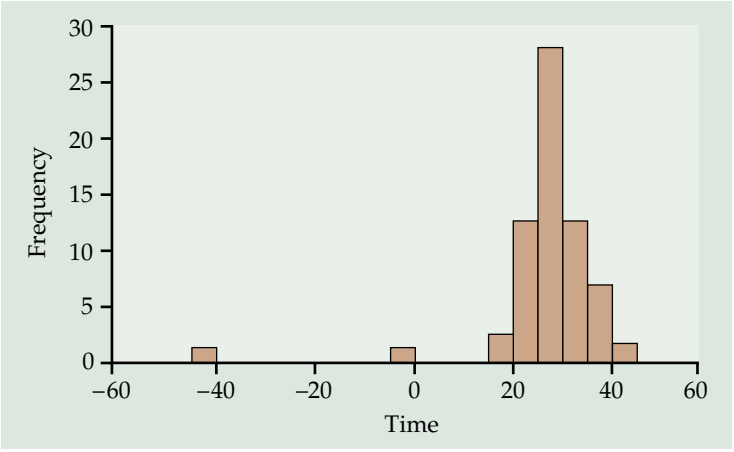



FIGURE 1.5 Histogram of Simon Newcomb's 66 measurements of the passage time of light, from Table 1.1.

the speed of light on the average (the mean, in the language of Section 1.2) of his observations. The mean of all 66 observations is 26.21. The mean of the 65 observations that he kept is 27.29. The strong effect of the single value −44 on the mean is one motivation for discarding it when we are interested in the center of the distribution as a whole.

## Time plots

We can gain more insight into Newcomb's data from a different kind of graph. When data represent similar observations made over time, it is wise to plot them against either time or against the order in which the observations were taken. Figure 1.6 plots Newcomb's passage times for light in the order that they were recorded. There is some suggestion in this plot that the variability (the vertical spread in the plot) is decreasing over time. In particular, both outlying observations were made early on. Perhaps Newcomb became more adept at using his equipment as he gained experience. Learning effects like this are quite common. If we allow Newcomb 20 observations for learning, the mean of the remaining 46 measurements is 28.15. The best modern measurements suggest that the "true value" for the passage time in Newcomb's experiment is 33.02. Allowing for learning does move the average result closer to the true value.

### 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.

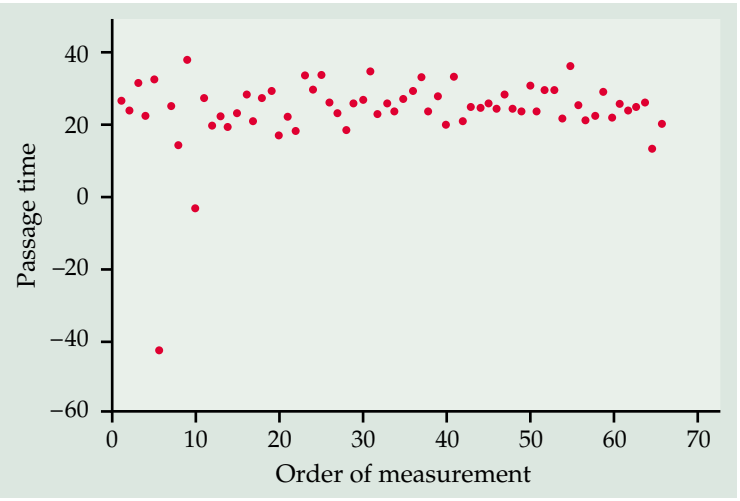



FIGURE 1.6 Plot of Newcomb's measurements against the time order in which they were made.

Whenever data are collected over time, it is a good idea to plot the observations in time order. Summaries of the distribution of a variable that ignore time order, such as stemplots and histograms, can be misleading when there is systematic change over time.

time series

Many interesting data sets are **time series**, measurements of a variable taken at regular intervals over time. Government economic and social data are often published as time series. Some examples are the monthly unemployment rate and the quarterly gross domestic product. Weather records, the demand for electricity, and measurements on the items produced by a manufacturing process are other examples of time series.

Time plots can reveal the main features of a time series. As in our examination of distributions, we 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 1.11

Figure 1.7 is a time plot of the average retail price of gasoline over the period 1988 to April 2001.[7] The plot shows the upward spike in prices due to the 1990 Iraqi invasion of Kuwait, the drop in 1998 when an economic crisis in Asia reduced demand for fuel,

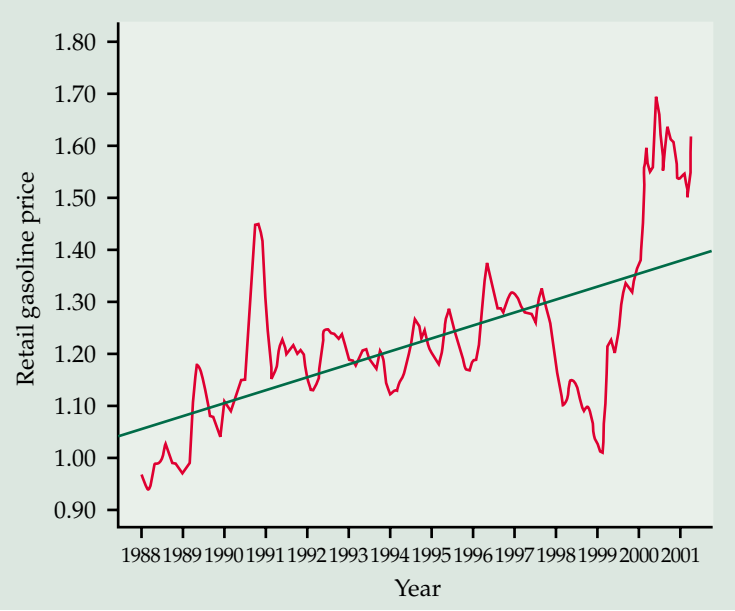



FIGURE 1.7 Time plot of average retail price of gasoline, 1988 to April 2001. The line shows the increasing trend.

and the rapid price increase in 2000 when demand recovered and OPEC reduced production. These deviations are so large that overall patterns are hard to see.

There is nonetheless a clear *trend* of increasing price. Using statistical software, we have drawn a line on the time plot in Figure 1.7 that represents the trend. Superimposed on this trend is *seasonal variation*, a regular rise and fall that recurs each year. Americans drive more in the summer vacation season, so the price of gasoline rises each spring, then drops in the fall as demand goes down.

seasonally adjusted

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.

## Beyond the basics ▸ decomposing time series*

Statistical software can help us examine a time series by "decomposing" the data into systematic patterns such as trends and seasonal variation and the *residuals* that remain after we remove these patterns. Figure 1.8 shows the combined trend and seasonal variation for the time series of gasoline prices

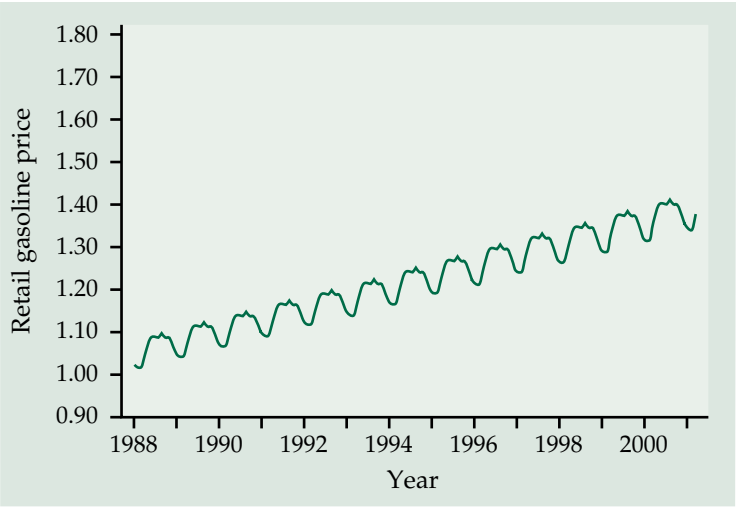



FIGURE 1.8 The estimated trend and seasonal variation in gasoline prices.

*The "Beyond the Basics" sections briefly discuss supplementary topics. Your software may make some of these available to you. For example, the results plotted in Figures 1.8 and 1.9 come from the Minitab statistical software.

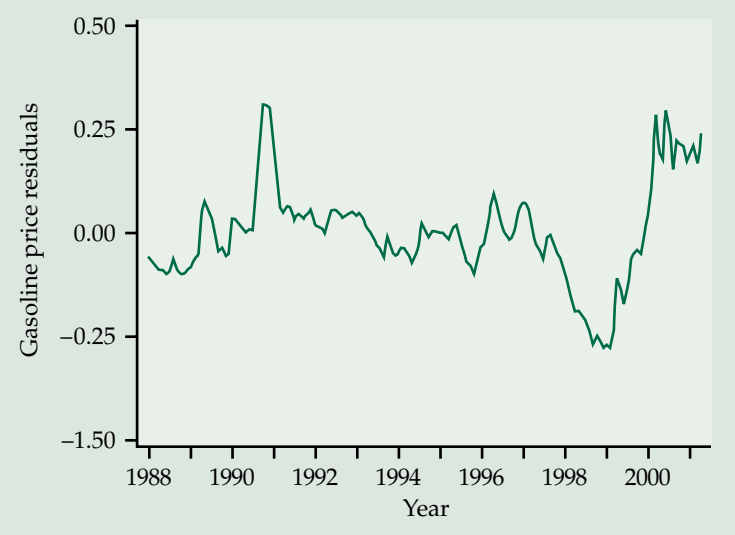



FIGURE 1.9 The erratic fluctuations that remain when both the trend and the seasonal variation are removed from gasoline prices.

in Figure 1.7. The seasonal variation represents an average of the seasonal pattern for all the years in the original data, automatically extracted by software. Figure 1.8 superimposes this average seasonal pattern on the trend. In terms of our strategy for looking at graphs, Figure 1.8 displays the overall pattern of the time plot. This pattern is well hidden by the erratic fluctuations in the data, so decomposing the time series is helpful.

Can we use the overall pattern in Figure 1.8 to predict next year's gasoline prices? Software also subtracts the trend and the seasonal variation from the original data, giving the residuals in Figure 1.9. This is a graph of the deviations of gasoline prices from the overall pattern of Figure 1.8. Figure 1.9 shows a great deal of erratic fluctuation, as well as the major deviations in 1990, 1998, and 2000 that we pointed out in Example 1.11. It is clear that trend and seasonal variation do not allow us to predict gasoline prices at all accurately.

## SUMMARY

A data set contains information on a collection of **individuals**. Individuals may be people, animals, or things. The data for one individual make up a **case**. 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, such as male or female. A quantitative variable has numerical values that measure some characteristic of each individual, such as height in centimeters or annual salary in dollars.

**Exploratory data analysis** uses graphs and numerical summaries to describe the variables in a data set and the relations among them.

A large number of observations on a single variable can be summarized in a table of **frequencies** (counts) or **relative frequencies** (percents or fractions).

**Bar graphs** and **pie charts** display the distributions of categorical variables. These graphs use the frequencies or relative frequencies of the categories.

**Stemplots** and **histograms** display the distributions of quantitative variables. Stemplots separate each observation into a **stem** and a one-digit **leaf**. Histograms plot the frequencies or relative frequencies of classes of values.

When examining a distribution, look for **shape**, **center**, and **spread** and for clear **deviations** from the overall shape.

Some distributions have simple shapes, such as **symmetric** and **skewed**. The number of **modes** (major peaks) is another aspect of overall shape. 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 variable vertically. A time plot can reveal **trends** or other changes over time.

## SECTION 1.1 EXERCISES

**1.1** Data from a 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)

**1.2** Here are the first lines of a professor's data set at the end of a statistics course:

| NAME | MAJOR | POINTS | GRADE |
|---|---|---|---|
| ADVANI, SURA | COMM | 397 | B |
| BARTON, DAVID | HIST | 323 | C |
| BROWN, ANNETTE | BIOL | 446 | A |
| CHIU, SUN | PSYC | 405 | B |
| CORTEZ, MARIA | PSYC | 461 | A |

What are the individuals and the variables in these data? Which variables are categorical and which are quantitative?

**1.3** Here is a small part of a data set that describes mutual funds available to the public:

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

What individuals does this data set describe? 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?

**1.4** You want to measure the amount of "leisure time" that college students enjoy. Write a brief discussion of two issues:

**(a)** How will you define "leisure time"?

**(b)** Once you have defined leisure time, how will you measure Sally's leisure time this week?

**1.5** You want to measure the "physical fitness" of college students. Describe several variables you might use to measure fitness. What instrument or instruments does each measurement require?

**1.6** Popular magazines rank colleges and universities on their "academic quality" in serving undergraduate students. Describe five variables that you would like to see measured for each college if you were choosing where to study. Give reasons for each of your choices.

**1.7** Congress wants the medical establishment to show that progress is being made in fighting cancer. Some variables that might be used are:[8]

**(a)** Total deaths from cancer. These have risen over time, from 331,000 in 1970 to 505,000 in 1980 to 552,000 in 2000.

**(b)** The percent of all Americans who die from cancer. The percent of deaths due to cancer has also risen steadily, from 17.2% in 1970 to 20.9% in 1980 to 23.0% in 2000.

**(c)** The percent of cancer patients who survive for five years from the time the disease is discovered. These rates are rising slowly. For whites, the five-year survival rate was 50.9% in the 1974 to 1979 period and 61.5% from 1989 to 1996.

Discuss the usefulness of each of these variables as a measure of the effectiveness of cancer treatment. In particular, explain why both (a) and (b) could increase even if treatment is getting more effective, and why (c) could increase even if treatment is getting less effective.

**1.8** Is driving becoming more dangerous? Traffic deaths declined for years, bottoming out at 39,250 killed in 1992, then began to increase again. In 1999, 41,611 people died in traffic accidents. But more people drove more miles in 1999 than in 1992. In fact, the government says that motor vehicles traveled 2247 billion miles in 1992 and 2691 billion miles in 1999. Reports on transportation use deaths per 100 million miles as a measure of risk. Compare the rates for 1992 and 1999. What do you conclude?

**1.9** The National Highway Traffic Safety Administration says that an average of 11 children die each year in school bus accidents, and an average of 600 school-age children die each year in auto accidents during school hours. These numbers suggest that riding the bus is safer than driving to school with a parent. The *counts* aren't fully convincing, however. What *rates* would you like to know to compare the safety of bus and private auto?

**1.10** 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. Describe an appropriate variable to measure the reliability of a make of dishwasher, and compute the values of this variable for Brand A and for Brand B.

**1.11** In 1997, there were 12,298,000 undergraduate students in U.S. colleges. According to the U.S. Department of Education, there were 127,000 American Indian or Alaskan Native students, 737,000 Asian or Pacific Islander, 1,380,000 non-Hispanic black, 1,108,000 Hispanic, and 8,682,000 non-Hispanic white students. In addition, 265,000 foreign undergraduates were enrolled in U.S. colleges.[9]

roundoff error

**(a)** Each number, including the total, is rounded to the nearest thousand. Separate rounding may cause **roundoff errors**, so that the sum of the counts does not equal the total given. Are roundoff errors present in these data?

**(b)** Present the data in a graph.

**1.12** The number of deaths among persons aged 15 to 24 years in the United States in 1999 due to the eight leading causes of death for this age group were: accidents, 13,602; homicide, 4989; suicide, 3885; cancer, 1724; heart disease, 1048; congenital defects, 430; respiratory disease, 208; AIDS, 197.[10]

**(a)** Make a bar graph to display these data.

**(b)** What additional information do you need to make a pie chart?

**1.13** According to the 2000 census, there are 105.5 million households in the United States. A household consists of people living together in the same residence, regardless of their relationship to each other. Of these, 71.8 million were "family households" in which at least one other person was related to the householder by blood, marriage, or adoption. The family households include 54.5 million headed by a married couple and 17.3 million other families (for example, a single parent with children). The other 33.7 million households are "nonfamily households." Of these, 27.2 million contain a person living alone, 5.5 million are unmarried couples living together, and 1 million consist of other unrelated people living together.[11] Be creative: make a bar graph that displays these facts, including the distinction between family and nonfamily households.

**1.14** Here are the percents of doctoral degrees in each of several subjects that were earned by women in 1997–98: psychology, 67.5%; education, 63.2%; life sciences, 42.5%; business, 31.4%; physical sciences, 25.2%; engineering, 12.2%.[12]

**(a)** Explain clearly why we cannot use a pie chart to display these data, even if we knew the percent female for every academic subject.

**(b)** Make a bar graph of the data. (Comparisons are easier if you order the bars by height, which is the order in which we give the percents. A bar graph ordered from tallest to shortest bar is sometimes called a **Pareto chart**, after the Italian economist who recommended this procedure.)

Pareto chart

**1.15** Here is a stemplot of the percents of residents aged 25 to 34 in each of the 50 states. The stems are whole percents and the leaves are tenths of a percent.

```
10 | 9
11 | 0
12 | 1 3 4 4 6 7 7 8 8 9
13 | 0 0 1 2 4 5 5 5 6 6 7 8 9 9 9 9
14 | 1 1 2 2 2 3 4 4 4 4 5 7 8 9
15 | 2 4 4 7 8 9 9 9
```

**(a)** Montana and Wyoming have the smallest percents of young adults, perhaps because they lack job opportunities. What are the percents for these two states?

**(b)** Ignoring Montana and Wyoming, describe the shape, center, and spread of this distribution.

**1.16** Make another stemplot of the percent of residents aged 25 to 34 in each of the 50 states by splitting the stems in the plot from the previous exercise. Which plot do you prefer? Why?

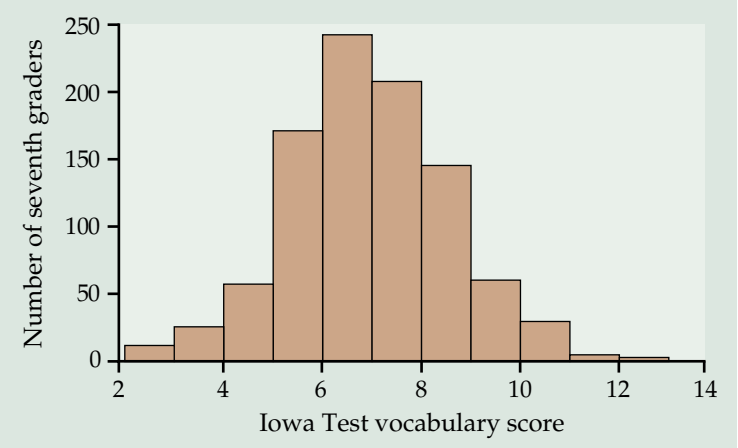



FIGURE 1.10 Histogram of the Iowa Test of Basic Skills vocabulary scores of seventh-grade students in Gary, Indiana, for Exercise 1.17.

**1.17** Figure 1.10 displays the scores of all 947 seventh-grade students in the public schools of Gary, Indiana, on the vocabulary part of the Iowa Test of Basic Skills.[13] Give a brief description of the overall pattern (shape, center, spread) of this distribution.

**1.18** Figure 1.11 is a histogram of the lengths of words used in Shakespeare's plays. Because there are so many words in the plays, we use a relative

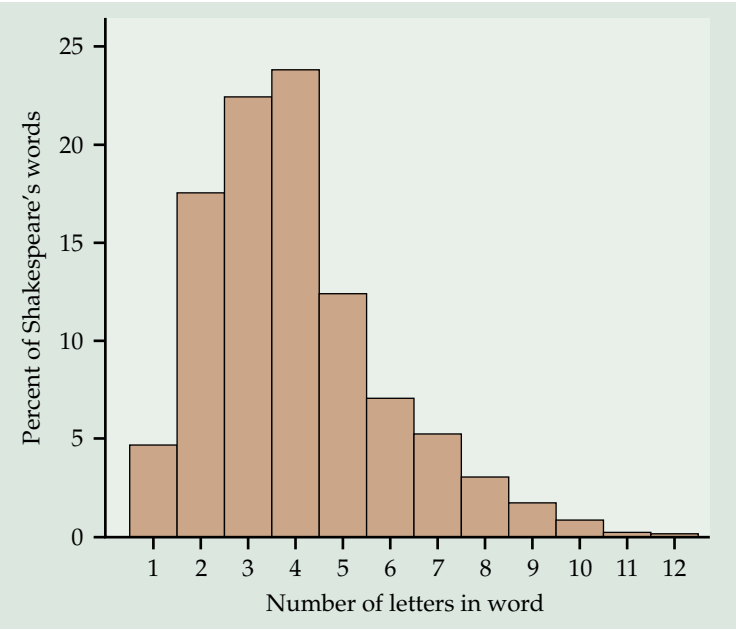



FIGURE 1.11 Histogram of lengths of words used in Shakespeare's plays, for Exercise 1.18.

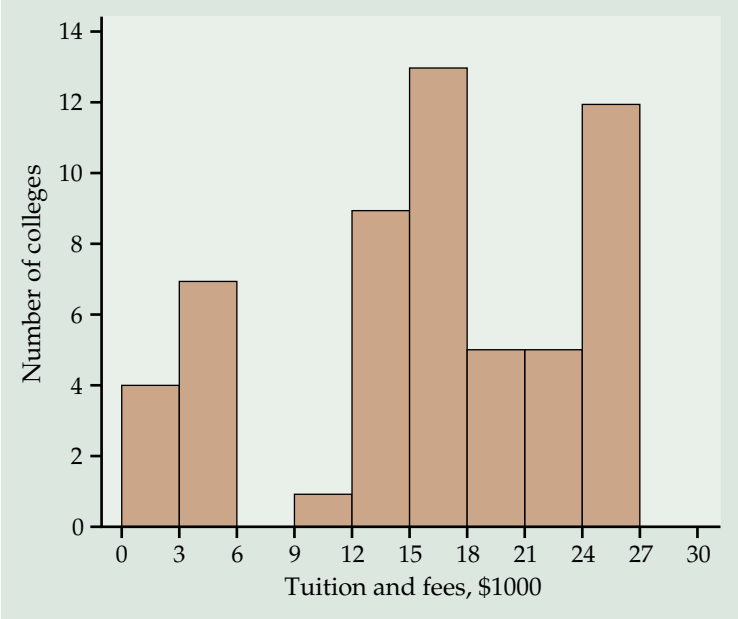



FIGURE 1.12 Histogram of the tuition and fees charged by all four-year colleges in Massachusetts, for Exercise 1.19.

frequency histogram. What is the overall shape of this distribution? What does this shape say about word lengths in Shakespeare? Do you expect other authors to have word length distributions of the same general shape? Why?

**1.19** Jeanna plans to attend college in her home state of Massachusetts. In the College Board's *Annual Survey of Colleges*, she finds data on college tuition and fees for the 2000–2001 academic year. Figure 1.12 displays the costs for all 56 four-year colleges in Massachusetts (omitting art schools and other special colleges). For state schools, we used the in-state tuition. What is the most important aspect of the overall pattern of this distribution? Why do you think this pattern appears?

**1.20** Outliers are sometimes the most interesting feature of a distribution. Figure 1.13 displays the distribution of batting averages for all 167 American League baseball players who batted at least 200 times in the 1980 season. The outlier is the .390 batting average of George Brett, the highest batting average in the major leagues since Ted Williams hit .406 in 1941. (See Exercise 1.87 on page 86 for a comparison of Brett and Williams.) Is the overall shape (ignoring the outlier) roughly symmetric or clearly skewed? What is the approximate midpoint of American League batting averages? What is the range if we ignore the outlier?

**1.21** How much oil the wells in a given field will ultimately produce is key information in deciding whether to drill more wells. Here are the estimated

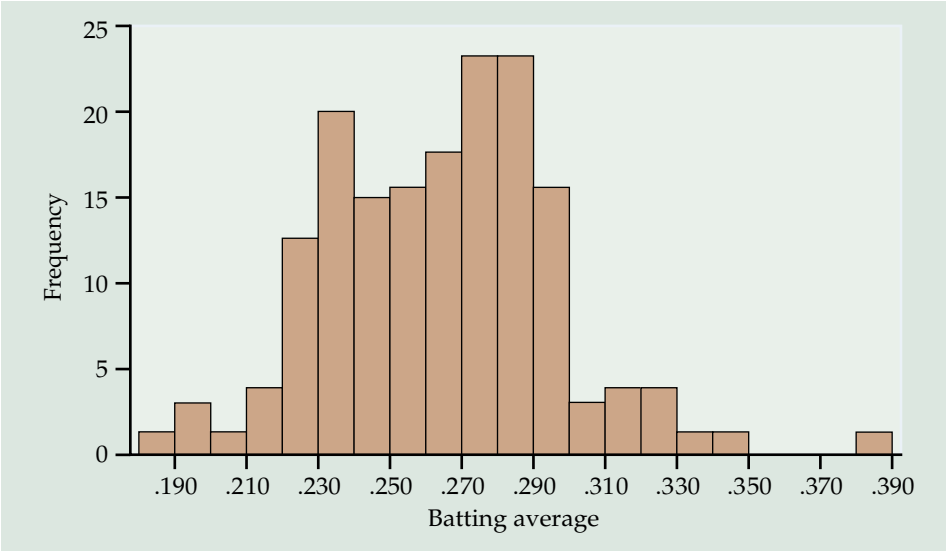



FIGURE 1.13 The distribution of batting averages of American League players in 1980, for Exercise 1.20.

total amounts of oil recovered from 64 wells in the Devonian Richmond Dolomite area of the Michigan basin, in thousands of barrels:[14]

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 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 |

Graph the distribution and describe its main features.

**1.22** 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.

**1.23** The Survey of Study Habits and Attitudes (SSHA) is a psychological test that evaluates college students' motivation, study habits, and attitudes toward school. A selective private college gives the SSHA to a sample of 18 of its incoming first-year college women. Their scores are

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 154 | 109 | 137 | 115 | 152 | 140 | 154 | 178 | 101 |
| 103 | 126 | 126 | 137 | 165 | 165 | 129 | 200 | 148 |

The college also administers the test to a sample of 20 first-year college men. Their scores are

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 108 | 140 | 114 | 91 | 180 | 115 | 126 | 92 | 169 | 146 |
| 109 | 132 | 75 | 88 | 113 | 151 | 70 | 115 | 187 | 104 |

**(a)** Make a back-to-back stemplot of the men's and women's scores. The overall shapes of the distributions are indistinct, as often happens when only a few observations are available. Are there any outliers?

**(b)** Compare the midpoints and the ranges of the two distributions. What is the most noticeable contrast between the female and male scores?

**1.24** There is some evidence that increasing the amount of calcium in the diet can lower blood pressure. In a medical experiment one group of men was given a daily calcium supplement, while a control group received a placebo (a dummy pill). The seated systolic blood pressure of all the men was measured before the treatments began and again after 12 weeks. The blood pressure distributions in the two groups should have been similar at the beginning of the experiment. Here are the initial blood pressure readings for the two groups:

Calcium group

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 107 | 110 | 123 | 129 | 112 | 111 | 107 | 112 | 136 | 102 |

Placebo group

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 123 | 109 | 112 | 102 | 98 | 114 | 119 | 112 | 110 | 117 | 130 |

Make a back-to-back stemplot of these data. Does your plot show any major differences in the two groups before the treatments began? In particular, are the centers of the two blood pressure distributions close together?

**1.25** Do women study more than men? We asked the students in a large first-year college class how many minutes they studied on a typical week night. Here are the responses of random samples of 30 women and 30 men from the class:

| Women | | | | | Men | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 180 | 120 | 180 | 360 | 240 | 90 | 120 | 30 | 90 | 200 |
| 120 | 180 | 120 | 240 | 170 | 90 | 45 | 30 | 120 | 75 |
| 150 | 120 | 180 | 180 | 150 | 150 | 120 | 60 | 240 | 300 |
| 200 | 150 | 180 | 150 | 180 | 240 | 60 | 120 | 60 | 30 |
| 120 | 60 | 120 | 180 | 180 | 30 | 230 | 120 | 95 | 150 |
| 90 | 240 | 180 | 115 | 120 | 0 | 200 | 120 | 120 | 180 |

**(a)** Examine the data. Why are you not surprised that most responses are multiples of 10 minutes? We eliminated one student who claimed to study 30,000 minutes per night. Are there any other responses you consider suspicious?

**(b)** Make a back-to-back stemplot of these data. Report the approximate midpoints of both groups. Does it appear that women study more than men (or at least claim that they do)?

**1.26** The Degree of Reading Power (DRP) test is often used to measure the reading ability of children. Here are the DRP scores of 44 third-grade students, measured during research on ways to improve reading performance:[15]

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 40 | 26 | 39 | 14 | 42 | 18 | 25 | 43 | 46 | 27 | 19 |
| 47 | 19 | 26 | 35 | 34 | 15 | 44 | 40 | 38 | 31 | 46 |
| 52 | 25 | 35 | 35 | 33 | 29 | 34 | 41 | 49 | 28 | 52 |
| 47 | 35 | 48 | 22 | 33 | 41 | 51 | 27 | 14 | 54 | 45 |

Make a stemplot of these data. Then make a histogram. Which display do you prefer, and why? Describe the main features of the distribution.

**1.27** In 1798 the English scientist Henry Cavendish measured the density of the earth by careful work with a torsion balance. The variable recorded was the density of the earth as a multiple of the density of water. Here are Cavendish's 29 measurements:[16]

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5.50 | 5.61 | 4.88 | 5.07 | 5.26 | 5.55 | 5.36 | 5.29 | 5.58 | 5.65 |
| 5.57 | 5.53 | 5.62 | 5.29 | 5.44 | 5.34 | 5.79 | 5.10 | 5.27 | 5.39 |
| 5.42 | 5.47 | 5.63 | 5.34 | 5.46 | 5.30 | 5.75 | 5.68 | 5.85 | |

Present these measurements graphically by either a stemplot or a histogram and explain the reason for your choice. Then briefly discuss the main features of the distribution. In particular, what is your estimate of the density of the earth based on these measurements?

**1.28** Table 1.3 gives the number of medical doctors per 100,000 people in each state.[17]

**(a)** 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?

**(b)** Make a graph to display the distribution of doctors per 100,000 people. Write a brief description of the distribution. Are there any outliers? If so, can you explain them?

TABLE 1.3 Medical doctors per 100,000 population, by state (1998)

| State | Doctors | State | Doctors | State | Doctors |
|---|---|---|---|---|---|
| Alabama | 198 | Louisiana | 246 | Ohio | 235 |
| Alaska | 167 | Maine | 223 | Oklahoma | 169 |
| Arizona | 202 | Maryland | 374 | Oregon | 225 |
| Arkansas | 190 | Massachusetts | 412 | Pennsylvania | 291 |
| California | 247 | Michigan | 224 | Rhode Island | 338 |
| Colorado | 238 | Minnesota | 249 | South Carolina | 207 |
| Connecticut | 354 | Mississippi | 163 | South Dakota | 184 |
| Delaware | 234 | Missouri | 230 | Tennessee | 246 |
| Florida | 238 | Montana | 190 | Texas | 203 |
| Georgia | 211 | Nebraska | 218 | Utah | 200 |
| Hawaii | 265 | Nevada | 173 | Vermont | 305 |
| Idaho | 154 | New Hampshire | 237 | Virginia | 241 |
| Illinois | 260 | New Jersey | 295 | Washington | 235 |
| Indiana | 195 | New Mexico | 212 | West Virginia | 215 |
| Iowa | 173 | New York | 387 | Wisconsin | 227 |
| Kansas | 203 | North Carolina | 232 | Wyoming | 171 |
| Kentucky | 209 | North Dakota | 222 | D.C. | 737 |

TABLE 1.4 Monthly percent returns on Philip Morris stock (July 1990 to May 1997)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| −5.7 | 1.2 | 4.1 | 3.2 | 7.3 | 7.5 | 18.6 | 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 | −2.9 | −2.3 | 3.5 | −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 | 1.3 | | | | | | | |

**1.29** Table 1.4 gives the monthly percent returns on Philip Morris stock for the period from July 1990 to May 1997. (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 is one clear outlier. What is the value of this observation? (It 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 outlier.

TABLE 1.5 Survival times (days) of guinea pigs in a medical experiment

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 43 | 45 | 53 | 56 | 56 | 57 | 58 | 66 | 67 | 73 |
| 74 | 79 | 80 | 80 | 81 | 81 | 81 | 82 | 83 | 83 |
| 84 | 88 | 89 | 91 | 91 | 92 | 92 | 97 | 99 | 99 |
| 100 | 100 | 101 | 102 | 102 | 102 | 103 | 104 | 107 | 108 |
| 109 | 113 | 114 | 118 | 121 | 123 | 126 | 128 | 137 | 138 |
| 139 | 144 | 145 | 147 | 156 | 162 | 174 | 178 | 179 | 184 |
| 191 | 198 | 211 | 214 | 243 | 249 | 329 | 380 | 403 | 511 |
| 522 | 598 | | | | | | | | |

**(c)** The data appear in time order reading from left to right across each row in turn, beginning with the −5.7% return in July 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?

**1.30** Table 1.5 gives the survival times in days of 72 guinea pigs after they were injected with tubercle bacilli in a medical experiment.[18] Make a suitable graph and describe the shape, center, and spread of the distribution of survival times. Are there any outliers?

**1.31** Table 1.6 presents data on 78 seventh-grade students in a rural midwestern school.[19] The researcher was interested in the relationship between the students' "self-concept" and their academic performance. The data we give here include each student's grade point average (GPA), score on a standard IQ test, and gender, taken from school records. Gender is coded as F for female and M for male. The students are identified only by an observation number (OBS). The missing OBS numbers show that some students dropped out of the study. The final variable is each student's score on the Piers-Harris Children's Self-Concept Scale, a psychological test administered by the researcher.

**(a)** How many variables does this data set contain? Which are categorical variables and which are quantitative variables?

**(b)** Make a stemplot of the distribution of GPA, after rounding to the nearest tenth of a point.

**(c)** Describe the shape, center, and spread of the GPA distribution. Identify any suspected outliers from the overall pattern.

**(d)** Make a back-to-back stemplot of the rounded GPAs for female and male students. Write a brief comparison of the two distributions.

**1.32** Make a graph of the distribution of IQ scores for the seventh-grade students in Table 1.6. Describe the shape, center, and spread of the distribution, as well as any outliers. IQ scores are usually said to be centered at 100. Is the midpoint for these students close to 100, clearly above, or clearly below?

TABLE 1.6 Educational data for 78 seventh-grade students

| OBS | GPA | IQ | Gender | Self-concept | OBS | GPA | IQ | Gender | Self-concept |
|---|---|---|---|---|---|---|---|---|---|
| 001 | 7.940 | 111 | M | 67 | 043 | 10.760 | 123 | M | 64 |
| 002 | 8.292 | 107 | M | 43 | 044 | 9.763 | 124 | M | 58 |
| 003 | 4.643 | 100 | M | 52 | 045 | 9.410 | 126 | M | 70 |
| 004 | 7.470 | 107 | M | 66 | 046 | 9.167 | 116 | M | 72 |
| 005 | 8.882 | 114 | F | 58 | 047 | 9.348 | 127 | M | 70 |
| 006 | 7.585 | 115 | M | 51 | 048 | 8.167 | 119 | M | 47 |
| 007 | 7.650 | 111 | M | 71 | 050 | 3.647 | 97 | M | 52 |
| 008 | 2.412 | 97 | M | 51 | 051 | 3.408 | 86 | F | 46 |
| 009 | 6.000 | 100 | F | 49 | 052 | 3.936 | 102 | M | 66 |
| 010 | 8.833 | 112 | M | 51 | 053 | 7.167 | 110 | M | 67 |
| 011 | 7.470 | 104 | F | 35 | 054 | 7.647 | 120 | M | 63 |
| 012 | 5.528 | 89 | F | 54 | 055 | 0.530 | 103 | M | 53 |
| 013 | 7.167 | 104 | M | 54 | 056 | 6.173 | 115 | M | 67 |
| 014 | 7.571 | 102 | F | 64 | 057 | 7.295 | 93 | M | 61 |
| 015 | 4.700 | 91 | F | 56 | 058 | 7.295 | 72 | F | 54 |
| 016 | 8.167 | 114 | F | 69 | 059 | 8.938 | 111 | F | 60 |
| 017 | 7.822 | 114 | F | 55 | 060 | 7.882 | 103 | F | 60 |
| 018 | 7.598 | 103 | F | 65 | 061 | 8.353 | 123 | M | 63 |
| 019 | 4.000 | 106 | M | 40 | 062 | 5.062 | 79 | M | 30 |
| 020 | 6.231 | 105 | F | 66 | 063 | 8.175 | 119 | M | 54 |
| 021 | 7.643 | 113 | M | 55 | 064 | 8.235 | 110 | M | 66 |
| 022 | 1.760 | 109 | M | 20 | 065 | 7.588 | 110 | M | 44 |
| 024 | 6.419 | 108 | F | 56 | 068 | 7.647 | 107 | M | 49 |
| 026 | 9.648 | 113 | M | 68 | 069 | 5.237 | 74 | F | 44 |
| 027 | 10.700 | 130 | F | 69 | 071 | 7.825 | 105 | M | 67 |
| 028 | 10.580 | 128 | M | 70 | 072 | 7.333 | 112 | F | 64 |
| 029 | 9.429 | 128 | M | 80 | 074 | 9.167 | 105 | M | 73 |
| 030 | 8.000 | 118 | M | 53 | 076 | 7.996 | 110 | M | 59 |
| 031 | 9.585 | 113 | M | 65 | 077 | 8.714 | 107 | F | 37 |
| 032 | 9.571 | 120 | F | 67 | 078 | 7.833 | 103 | F | 63 |
| 033 | 8.998 | 132 | F | 62 | 079 | 4.885 | 77 | M | 36 |
| 034 | 8.333 | 111 | F | 39 | 080 | 7.998 | 98 | F | 64 |
| 035 | 8.175 | 124 | M | 71 | 083 | 3.820 | 90 | M | 42 |
| 036 | 8.000 | 127 | M | 59 | 084 | 5.936 | 96 | F | 28 |
| 037 | 9.333 | 128 | F | 60 | 085 | 9.000 | 112 | F | 60 |
| 038 | 9.500 | 136 | M | 64 | 086 | 9.500 | 112 | F | 70 |
| 039 | 9.167 | 106 | M | 71 | 087 | 6.057 | 114 | M | 51 |
| 040 | 10.140 | 118 | F | 72 | 088 | 6.057 | 93 | F | 21 |
| 041 | 9.999 | 119 | F | 54 | 089 | 6.938 | 106 | M | 56 |

**1.33** Based on a suitable graph, briefly describe the distribution of self-concept scores for the students in Table 1.6. Be sure to identify any suspected outliers.

**1.34** The distribution of the ages of a nation's population has a strong influence on economic and social conditions. The following table shows the age distribution of U.S. residents in 1950 and 2050, in millions of people. The 1950 data come from that year's census, while the 2050 data are projections made by the Census Bureau.

| Age group | 1950 | 2050 |
|---|---|---|
| Under 10 years | 29.3 | 53.3 |
| 10 to 19 years | 21.8 | 53.2 |
| 20 to 29 years | 24.0 | 51.2 |
| 30 to 39 years | 22.8 | 50.5 |
| 40 to 49 years | 19.3 | 47.5 |
| 50 to 59 years | 15.5 | 44.8 |
| 60 to 69 years | 11.0 | 40.7 |
| 70 to 79 years | 5.5 | 30.9 |
| 80 to 89 years | 1.6 | 21.7 |
| 90 to 99 years | 0.1 | 8.8 |
| 100 to 109 years | — | 1.1 |
| Total | 151.1 | 403.7 |

**(a)** Because the total population in 2050 is much larger than the 1950 population, comparing percents (relative frequencies) 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 2050.

**(b)** Make a relative frequency histogram 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.

**(c)** Make a relative frequency histogram of the projected age distribution for the year 2050. 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 century between 1950 and 2050?

**1.35** **(Optional)** Sometimes you want to make a histogram from data that are already grouped into classes of unequal width. A report on the recent graduates of a large state university includes the following relative frequency table of the first-year salaries of last year's graduates. Salaries are in $1000 units, and it is understood that each class includes its left endpoint but not its right endpoint—for example, a salary of exactly $20,000 belongs in the second class.

| Salary | 15–20 | 20–25 | 25–30 | 30–35 | 35–40 | 40–50 | 50–60 | 60–80 |
|---|---|---|---|---|---|---|---|---|
| Percent | 7 | 14 | 29 | 23 | 13 | 9 | 4 | 1 |

The last three classes are wider than the others. An accurate histogram must take this into account. If the base of each bar in the histogram covers a class and the height is the percent of graduates with salaries in that class, the areas of the three rightmost bars will overstate the percent who have salaries in these classes. To make a correct histogram, the area of each bar must be proportional to the percent in that class. Most classes are $5000 wide. A class *twice* as wide ($10,000) should have a bar *half* as tall as the percent in that class. This keeps the area proportional to the percent. How should you treat the height of the bar for a class $20,000 wide? Make a correct histogram with the heights of the bars for the last three classes adjusted so that the areas of the bars reflect the percent in each class.

**1.36** "Major hurricanes account for just over 20% of the tropical storms and hurricanes that strike the United States but cause more than 80% of the damage." So say investigators who have shown that major hurricanes (with sustained wind speeds at least 50 meters per second) are tied to ocean temperature and other variables. These variables change slowly, so the high level of hurricane activity that began in 1995 "is likely to persist for an additional 10 to 40 years." This is bad news for people with beach houses on the Atlantic coast. Table 1.7 gives the counts of major hurricanes for each year between 1944 and 2000.[20]

**(a)** What is the average number of major hurricanes per year during the period 1944 to 2000?

TABLE 1.7 Major hurricanes (sustained winds in excess of 50 meters per second), 1944 to 2000

| Year | Count | Year | Count | Year | Count | Year | Count | Year | Count |
|---|---|---|---|---|---|---|---|---|---|
| 1944 | 3 | 1956 | 2 | 1968 | 0 | 1980 | 2 | 1992 | 1 |
| 1945 | 2 | 1957 | 2 | 1969 | 3 | 1981 | 3 | 1993 | 1 |
| 1946 | 1 | 1958 | 4 | 1970 | 2 | 1982 | 1 | 1994 | 0 |
| 1947 | 2 | 1959 | 2 | 1971 | 1 | 1983 | 1 | 1995 | 5 |
| 1948 | 4 | 1960 | 2 | 1972 | 0 | 1984 | 1 | 1996 | 6 |
| 1949 | 3 | 1961 | 6 | 1973 | 1 | 1985 | 3 | 1997 | 1 |
| 1950 | 7 | 1962 | 0 | 1974 | 2 | 1986 | 0 | 1998 | 3 |
| 1951 | 2 | 1963 | 2 | 1975 | 3 | 1987 | 1 | 1999 | 5 |
| 1952 | 3 | 1964 | 5 | 1976 | 2 | 1988 | 3 | 2000 | 3 |
| 1953 | 3 | 1965 | 1 | 1977 | 1 | 1989 | 2 | | |
| 1954 | 2 | 1966 | 3 | 1978 | 2 | 1990 | 1 | | |
| 1955 | 5 | 1967 | 1 | 1979 | 2 | 1991 | 2 | | |

**(b)** Make a time plot of the count of major hurricanes. Draw a line across your plot at the average number of hurricanes per year. This helps divide the plot into three periods. Describe the pattern you see.

**1.37** 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:[21]

| 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 | | |

**(a)** Make a time plot of the returns paid by Treasury bills in these years.

cycles

**(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?

**1.38** Time series data often display the effects of changes in policy. Here are data on motor vehicle deaths in the United States. As in Example 1.3, we look at the death rate per 100 million miles driven.

| Year | Rate | Year | Rate | Year | Rate | Year | Rate |
|---|---|---|---|---|---|---|---|
| 1960 | 5.1 | 1970 | 4.7 | 1980 | 3.3 | 1990 | 2.1 |
| 1962 | 5.1 | 1972 | 4.3 | 1982 | 2.8 | 1992 | 1.7 |
| 1964 | 5.4 | 1974 | 3.5 | 1984 | 2.6 | 1994 | 1.7 |
| 1966 | 5.5 | 1976 | 3.2 | 1986 | 2.5 | 1996 | 1.7 |
| 1968 | 5.2 | 1978 | 3.3 | 1988 | 2.3 | 1998 | 1.6 |

**(a)** Make a time plot of these death rates. During these years, safety requirements for motor vehicles became stricter and interstate highways replaced older roads. How does the pattern of your plot reflect these changes?

**(b)** In 1974 the national speed limit was lowered to 55 miles per hour in an attempt to conserve gasoline after the 1973 Arab-Israeli War. In the

mid-1980s most states raised speed limits on interstate highways to 65 miles per hour. Some said that the lower speed limit saved lives. Is the effect of lower speed limits between 1974 and the mid-1980s visible in your plot?

**(c)** Does it make sense to make a histogram of these 20 death rates? Explain your answer.

**1.39** 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 more time plots of the data in Exercise 1.38, one that makes motor vehicle death rates appear to decrease very rapidly and one that shows only a slow decrease. The moral of this exercise is: pay close attention to the scales when you look at a time plot.

**1.40** The following table gives the times (in minutes, rounded to the nearest minute) for the winning man in the Boston Marathon in the years 1959 to 2001:

| Year | Time | Year | Time | Year | Time | Year | Time |
|---|---|---|---|---|---|---|---|
| 1959 | 143 | 1970 | 131 | 1981 | 129 | 1992 | 128 |
| 1960 | 141 | 1971 | 139 | 1982 | 129 | 1993 | 130 |
| 1961 | 144 | 1972 | 136 | 1983 | 129 | 1994 | 127 |
| 1962 | 144 | 1973 | 136 | 1984 | 131 | 1995 | 129 |
| 1963 | 139 | 1974 | 134 | 1985 | 134 | 1996 | 129 |
| 1964 | 140 | 1975 | 130 | 1986 | 128 | 1997 | 131 |
| 1965 | 137 | 1976 | 140 | 1987 | 132 | 1998 | 128 |
| 1966 | 137 | 1977 | 135 | 1988 | 129 | 1999 | 130 |
| 1967 | 136 | 1978 | 130 | 1989 | 129 | 2000 | 130 |
| 1968 | 142 | 1979 | 129 | 1990 | 128 | 2001 | 130 |
| 1969 | 134 | 1980 | 132 | 1991 | 131 | | |

Women were allowed to enter the race in 1972. Here are the times for the winning woman from 1972 to 2001:

| Year | Time | Year | Time | Year | Time | Year | Time |
|---|---|---|---|---|---|---|---|
| 1972 | 190 | 1980 | 154 | 1988 | 145 | 1996 | 147 |
| 1973 | 186 | 1981 | 147 | 1989 | 144 | 1997 | 146 |
| 1974 | 167 | 1982 | 150 | 1990 | 145 | 1998 | 143 |
| 1975 | 162 | 1983 | 143 | 1991 | 144 | 1999 | 143 |
| 1976 | 167 | 1984 | 149 | 1992 | 144 | 2000 | 146 |
| 1977 | 168 | 1985 | 154 | 1993 | 145 | 2001 | 144 |
| 1978 | 165 | 1986 | 145 | 1994 | 142 | | |
| 1979 | 155 | 1987 | 146 | 1995 | 145 | | |

Make time plots of the men's and women's winning times on the same graph for easy comparison. Give a brief description of the pattern of Boston Marathon winning times over these years. Have times stopped improving in recent years? If so, when did improvement end for men and for women?

## 1.2 Describing Distributions with Numbers

Interested in an exotic car? Worried that it may use too much gas? The Environmental Protection Agency lists most such vehicles in its "minicompact" or "two-seater" categories. Table 1.8 gives the city and highway gas mileage for all cars in these groups.[22] (The mileages are for the basic engine and transmission combination for each car.) We want to compare minicompacts with two-seaters and city mileage with highway mileage. We can begin with graphs, but numerical summaries make the comparisons more specific.

A brief description of a distribution should include its *shape* and numbers describing its *center* and *spread*. We describe the shape of a distribution based on inspection of a histogram or a stemplot. Now we will learn specific ways to use numbers to measure the center and spread of a distribution. We can calculate these numerical measures for any quantitative variable. But to interpret measures of center and spread, and to choose among the several measures

TABLE 1.8 Fuel economy (miles per gallon) for model year 2001 cars

| Minicompact cars | | | Two-seater cars | | |
|---|---|---|---|---|---|
| Model | City | Highway | Model | City | Highway |
| Audi TT Coupe | 22 | 31 | Acura NSX | 17 | 24 |
| BMW 325CI Convertible | 19 | 27 | Audi TT Roadster | 22 | 30 |
| BMW 330CI Convertible | 20 | 28 | BMW Z3 Coupe | 21 | 28 |
| BMW M3 Convertible | 16 | 23 | BMW Z3 Roadster | 20 | 27 |
| Jaguar XK8 Convertible | 17 | 24 | BMW Z8 | 13 | 21 |
| Jaguar XKR Convertible | 16 | 22 | Chevrolet Corvette | 18 | 26 |
| Mercedes-Benz CLK320 | 20 | 28 | Dodge Viper | 11 | 21 |
| Mercedes-Benz CLK430 | 18 | 24 | Ferrari Modena | 11 | 16 |
| Mitsubishi Eclipse | 22 | 30 | Ferrari Maranello | 8 | 13 |
| Porsche 911 Carrera | 17 | 25 | Honda Insight | 61 | 68 |
| Porsche 911 Turbo | 15 | 22 | Honda S2000 | 20 | 26 |
| | | | Lamborghini Diablo | 10 | 13 |
| | | | Mazda Miata | 22 | 28 |
| | | | Mercedes-Benz SL500 | 16 | 23 |
| | | | Mercedes-Benz SL600 | 13 | 19 |
| | | | Mercedes-Benz SLK320 | 21 | 27 |
| | | | Plymouth Prowler | 17 | 23 |
| | | | Porsche Boxster | 19 | 27 |
| | | | Toyota MR2 | 25 | 30 |

we will learn, you must think about the shape of the distribution and the meaning of the data. The numbers, like graphs, are aids to understanding, not "the answer" in themselves.

## Measuring center: the mean

Numerical description of a distribution begins with a measure of its center or average. The two common measures of center are the *mean* and the *median*. The mean is the "average value" and the median is the "middle value." These are two different ideas for "center," and the two measures behave differently. We need precise recipes for the mean and the median.

### The Mean $\bar{x}$

To find the **mean** $\bar{x}$ of a set of observations, add their values and divide by the number of observations. If the $n$ observations are $x_1, x_2, \ldots, x_n$, 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 bar over the $x$ indicates the mean of all the $x$-values. Pronounce the mean $\bar{x}$ as "x-bar." This notation is so common that writers who are discussing data use $\bar{x}$, $\bar{y}$, etc. without additional explanation. 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.

EXAMPLE 1.12

The mean highway mileage for the 19 two-seaters in Table 1.8 is

$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{24 + 30 + 28 + \cdots + 30}{19} \\ &= \frac{490}{19} = 25.8 \text{ miles per gallon}\end{aligned}$$

In practice, you can key the data into your calculator and hit the $\bar{x}$ key.

The data for Example 1.12 contain an outlier: the Honda Insight is a hybrid gas-electric car that doesn't belong in the same category as the 18 gasoline-powered two-seater cars. If we exclude the Insight, the mean highway mileage drops to 23.4 MPG. The single outlier adds 2.4 MPG to the mean highway mileage. This illustrates an important weakness of the mean as a

resistant measure

measure of center: it is sensitive to the influence of a few 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 influence of extreme observations, we say that it is not a **resistant measure** of center. A measure that is resistant does more than limit the influence of outliers. Its value does not respond strongly to changes in a few observations, no matter how large those changes may be. The mean fails this requirement because we can make the mean as large as we wish by making a large enough increase in just one observation.

## Measuring center: the median

We used the midpoint of a distribution as an informal measure of center in the previous section. 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 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 computer software or a graphing calculator to automate finding the median.

EXAMPLE 1.13

To find the median highway mileage for 2001 model two-seater cars, arrange the data in increasing order:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 13 | 13 | 16 | 19 | 21 | 21 | 23 | 23 | 24 | **26** |
| 26 | 27 | 27 | 27 | 28 | 28 | 30 | 30 | 68 | |

The median is the bold 26, the 10th observation in the ordered list. You can find the median by eye—there are 9 observations to the left and 9 to the right. Or you can use the recipe $(n + 1)/2 = 20/2 = 10$ to locate the median in the list.

What happens if we drop the Honda Insight? The remaining 18 cars have highway mileages

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 13 | 13 | 16 | 19 | 21 | 21 | 23 | 23 | **24** | **26** | 26 | 27 | 27 |
| 27 | 28 | 28 | 30 | 30 | | | | | | | | |

Because the number of observations $n = 18$ is even, there is no center observation. There is a center pair—the bold 24 and 26 have 8 observations to their left and 8 to their right. The median $M$ is the mean of the center pair:

$$M = \frac{24 + 26}{2} = \frac{50}{2} = 25$$

The recipe $(n + 1)/2 = 19/2 = 9.5$ for the position of the median in the list says that the median is at location "nine and one-half," that is, halfway between the 9th and 10th observations.

You see that the median is more resistant than the mean. Removing the Honda Insight reduces the median highway mileage by just 1 MPG. And if we mistakenly enter the Insight's mileage as 680 rather than 68, the median remains 25. The very high value is simply one observation to the right of center. The *Mean and Median* applet available on the Web site `www.whfreeman.com/ips` is an excellent way to compare the resistance of $M$ and $\bar{x}$. See Exercises 1.51 and 1.52 for use of this applet.


## Mean versus median

The median and mean are the most common measures of the center of a distribution. 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 ("middle value") rather than the mean ("arithmetic average value"). 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.

We now have two general strategies for dealing with outliers and other irregularities in data. The first strategy asks us to detect outliers and investigate their causes. We can then correct outliers if they are wrongly recorded, delete them for good reason, or otherwise give them individual attention. The second strategy is to use resistant methods, so that outliers have little influence over our conclusions. In general, we prefer to examine data carefully and consider any irregularity in the light of the individual situation. This will sometimes lead to a decision to employ resistant methods.

EXAMPLE 1.14

Newcomb's data are supposed to be repeated measurements of the same quantity, but they contain unexplained outliers. Even though we can't explain the outlying values, we should not give them much influence over our estimate of the velocity of light. We can delete the outliers from a computation of the mean, or we can report the median. If the two outliers are omitted, the mean of the remaining 64 measurements is $\bar{x} = 27.75$. The median of 66 observations has position $(66 + 1)/2 = 33.5$ in the ordered list of passage times from Table 1.1. Its value is $M = 27$. Either of these values is preferable to the overall mean ($\bar{x} = 26.21$) as a statement of Newcomb's value for the travel time of light.

## Measuring spread: the quartiles

A measure of center alone can be misleading. Two nations with the same median family income are very different if one has extremes of wealth and poverty and the other has little variation among families. 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.**

percentile

We can describe the spread or variability of a distribution by 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. (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 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.

EXAMPLE 1.15

The highway mileages of the 18 gasoline-powered two-seater cars, arranged in increasing order, are

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 13 | 13 | 16 | 19 | 21 | 21 | 23 | 23 | 24 | \| | 26 | 26 | 27 | 27 |
| 27 | 28 | 28 | 30 | 30 | | | | | | | | | |

The median is midway between the center pair of observations. We have marked its position in the list by |. The first quartile is the median of the 9 observations to the left of the position of the median. It is the 5th of these, $Q_1 = 21$. Similarly, the third quartile is the median of the 9 observations to the right of the |. Check that $Q_3 = 27$.

When there is an odd number of observations, the median is the unique center observation, and the rule for finding the quartiles excludes this center value. The highway mileages of the 11 minicompact cars in Table 1.8 are (in order)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 22 | 22 | 23 | 24 | 24 | **25** | 27 | 28 | 28 | 30 | 31 |

The median is the bold 25. The first quartile is the median of the 5 observations falling to the left of this point in the list, $Q_1 = 23$. Similarly, $Q_3 = 28$.

We find other percentiles more informally. For example, we take the 90th percentile of the 18 two-seater mileages to be the 16th in the ordered list, because $0.90 \times 18 = 16.2$, which we round to 16. The 90th percentile is therefore 28 MPG.

EXAMPLE 1.16

Statistical software often provides several numerical measures in response to a single command. Figure 1.14 displays such output from the Minitab and SPSS software for the shopping data from Example 1.5. Both give us the number of observations, the mean and median, and the quartiles, as well as other measures. Minitab has been set to round its output to two decimal places. You can check that the values for the quartiles differ slightly from those given by our rule. For example, the first quartile is \$19.27 by hand and \$19.06 by software. Software often uses more elaborate rules that aren't suited for hand calculation. Moreover, not all software uses the same rule. For example, the Microsoft Excel spreadsheet gives the first quartile as 19.3275. These differences are too small to affect conclusions based on the data. Just use the values that your software gives you.

## The five-number summary and boxplots

In Section 1.1, we used the smallest and largest observations to indicate the spread of a distribution. These single 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.

SPSS

| | | |
|---|---|---|
| **N** | **Valid** | 50 |
| | **Missing** | 0 |
| Mean | | 34.7022 |
| Median | | 27.8550 |
| Std. Deviation | | 21.6974 |
| Range | | 90.23 |
| Minimum | | 3.11 |
| Maximum | | 93.34 |
| Percentiles | 25 | 19.0600 |
| | 50 | 27.8550 |
| | 75 | 45.7225 |

Minitab

Descriptive Statistics

| Variable | N | Mean | Median | TrMean | StDev | SEMean |
|---|---|---|---|---|---|---|
| spending | 50 | 34.70 | 27.85 | 32.92 | 21.70 | 3.07 |

| Variable | Min | Max | Q1 | Q3 |
|---|---|---|---|---|
| spending | 3.11 | 93.34 | 19.06 | 45.72 |

FIGURE 1.14 Numerical descriptions of the unrounded shopping data from the SPSS and Minitab software.

### The Five-Number Summary

The **five-number summary** of a set of observations 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

$$\text{Minimum} \quad Q_1 \quad M \quad Q_3 \quad \text{Maximum}$$

These five numbers offer a reasonably complete description of center and spread. The five-number summaries for highway gas mileages are

| | | | | |
|---|---|---|---|---|
| 13 | 21 | 25 | 27 | 30 |

for two-seaters and

| | | | | |
|---|---|---|---|---|
| 22 | 23 | 25 | 28 | 31 |

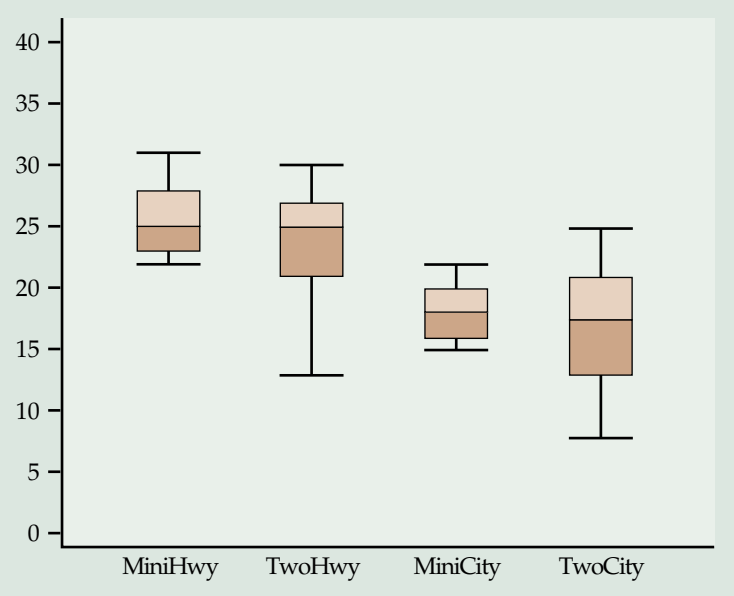



FIGURE 1.15 Boxplots of the highway and city gas mileages for cars classified as two-seaters and as minicompacts by the Environmental Protection Agency.

for minicompacts. The median describes the center of the distribution; the quartiles show the spread of the center half of the data; the minimum and maximum show the full spread of the data. The five-number summary leads to another visual representation of a distribution, the *boxplot*. Figure 1.15 shows boxplots for both city and highway gas mileages for our two groups of cars.

### Boxplot

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

- A central box spans the quartiles $Q_1$ and $Q_3$.
- A line in the box marks the median $M$.
- Lines extend from the box out to the smallest and largest observations.

Because boxplots show less detail than histograms or stemplots, they are best used for side-by-side comparison of more than one distribution, as in Figure 1.15. 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 see at once that city mileages are lower than highway mileages. The median gas mileages for the two groups of cars are very close together, but the two-seaters are more variable. In particular, the two-seater group contains some cars with quite low gas mileages.

## The 1.5 × *IQR* criterion for suspected outliers

Look again at the distribution of the percent of each state's adult population that is Hispanic, given in Table 1.2 (page 14) and displayed in the histogram in Figure 1.4. You can check that the five-number summary for this distribution is

| | | | | |
|---|---|---|---|---|
| 0.6 | 2.0 | 4.1 | 7.0 | 38.7 |

The largest observation (New Mexico, 38.7% Hispanic) is an outlier. How shall we describe the spread of this distribution? The range of all the observations depends entirely on the smallest and largest observations and does not describe the spread of the majority of the data. The distance between the quartiles, the range of the center half of the data, is a more resistant measure of spread. This distance is called the *interquartile range*.

### The Interquartile Range *IQR*

The **interquartile range *IQR*** is the distance between the first and third quartiles,

$$IQR = Q_3 - Q_1$$

For our data on Hispanics in the states, $IQR = 7.0 - 2.0 = 5.0$. The quartiles and the *IQR* are not affected by changes in either tail of the distribution. They are therefore resistant, because changes in a few data points have no further effect once these points move outside the quartiles. You should be aware that no single numerical measure of spread, such as *IQR*, is very useful for describing skewed distributions. The two sides of a skewed distribution have different spreads. We can often detect skewness from the five-number summary by comparing how far the first quartile and the minimum are from the median (left tail) with how far the third quartile and the maximum (right tail) are from the median. The interquartile range is mainly used as the basis of a rule of thumb for identifying suspected outliers.

### The 1.5 × *IQR* Criterion for Outliers

Call an observation a suspected outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile.

EXAMPLE 1.17

For the Hispanics data,

$$1.5 \times IQR = 1.5 \times 5 = 7.5$$

Any values below $2.0 - 7.5 = -5.5$ or above $7.0 + 7.5 = 14.5$ are flagged as possible outliers. There are no low outliers, but 7 states are flagged as possible high outliers.

This distribution is strongly skewed and has a quite compact middle half (small *IQR*). The histogram in Figure 1.4 (page 15) suggests that only New Mexico is truly an outlier in the sense of deviating from the overall pattern of the distribution. The other 6 states are just part of the long right tail. You see that the $1.5 \times IQR$ rule does not remove the need to look at the distribution and use judgment. It is useful mainly when large volumes of data must be scanned automatically.

modified boxplot

We can modify boxplots to plot suspected outliers individually. In a **modified boxplot**, the lines extend out from the central box only to the smallest and largest observations that are not suspected outliers. Observations more than $1.5 \times IQR$ outside the box are plotted as individual points. Figure 1.16 is a modified boxplot of the data from Table 1.2.

The modified boxplot in Figure 1.16, and especially the histogram in Figure 1.4, tell us much more about the distribution than the five-number summary or other numerical measures. The routine methods of statistics compute numerical measures and draw conclusions based on their values. These methods are very useful, and we will study them carefully in later chapters. But they cannot be applied blindly, by feeding data to a computer program, because **statistical measures and methods based on them are**

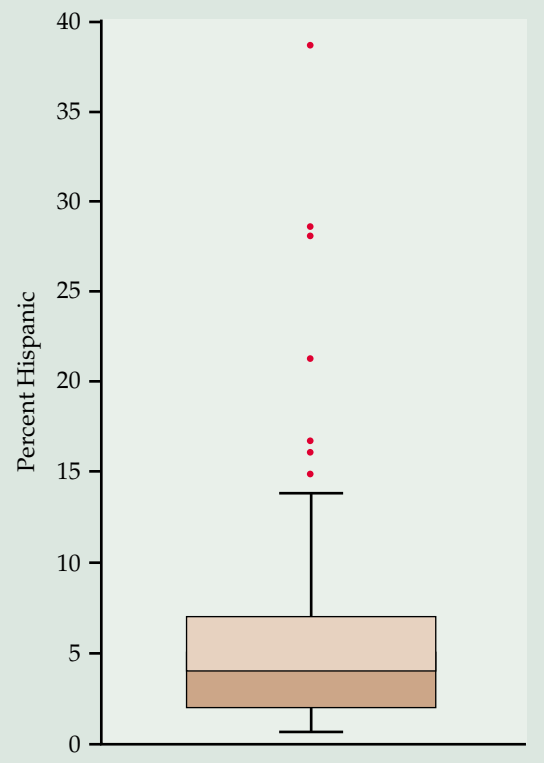



FIGURE 1.16 Modified boxplot of the percent of adults in each state who identified themselves as Hispanic in the 2000 census.

**generally meaningful only for distributions of sufficiently regular shape**. This principle will become clearer as we progress, but it is good to be aware at the beginning that quickly resorting to fancy calculations is the mark of a statistical amateur. Look, think, and choose your calculations selectively.

## 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 $x_1, x_2, \ldots, x_n$ is

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

or, in more compact notation,

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

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

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

The idea behind the variance and the standard deviation as measures of spread is as follows: The deviations $x_i - \bar{x}$ display the spread of the values $x_i$ about their mean $\bar{x}$. Some of these deviations will be positive and some negative because some of the observations fall on each side of the mean. In fact, *the sum of the deviations of the observations from their mean will always be zero*. Squaring the deviations makes them all positive, so that observations far from the mean in either direction have large positive squared deviations. The variance is the average squared deviation. Therefore, $s^2$ and $s$ will be large if the observations are widely spread about their mean, and small if the observations are all close to the mean.

EXAMPLE 1.18

A person's metabolic rate is the rate at which the body consumes energy. Metabolic rate is important in studies of weight gain, dieting, and exercise. Here are the metabolic rates of 7 men who took part in a study of dieting. (The units are calories per 24 hours. These are the same calories used to describe the energy content of foods.)

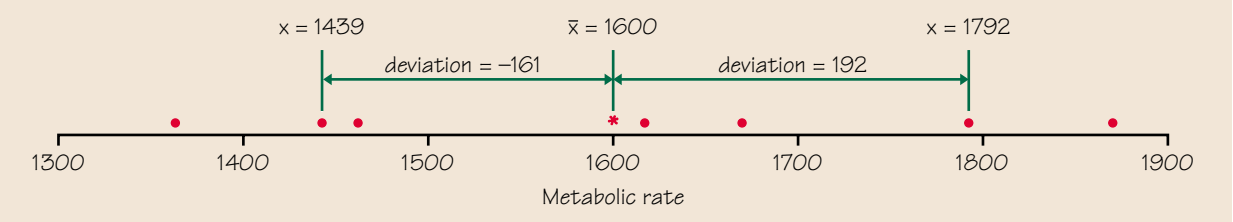



FIGURE 1.17 Metabolic rates for seven men, with the mean (∗) and the deviations of two observations from the mean.

| | | | | | | |
|---|---|---|---|---|---|---|
| 1792 | 1666 | 1362 | 1614 | 1460 | 1867 | 1439 |

Enter these data into your calculator or software and verify that

$$\bar{x} = 1600 \text{ calories} \quad s = 189.24 \text{ calories}$$

Figure 1.17 plots these data as dots on the calorie scale, with their mean marked by an asterisk (∗). The arrows mark two of the deviations from the mean. If you were calculating $s$ by hand, you would find the first deviation as

$$x_1 - \bar{x} = 1792 - 1600 = 192$$

Exercise 1.62 asks you to calculate the seven deviations, square them, and find $s^2$ and $s$ directly from the deviations. Working one or two short examples by hand helps you understand how the standard deviation is obtained. In practice you will always use either software or a calculator that will find $s$ from keyed-in data.

The idea of the variance is straightforward: it is the average of the squares of the deviations of the observations from their mean. The details we have just presented, however, raise some questions.

*Why do we square the deviations?* Why not just average the distances of the observations from their mean? There are two reasons, neither of them obvious. First, the sum of the squared deviations of any set of observations from their mean is the smallest that the sum of squared deviations from any number can possibly be. This is not true of the unsquared distances. So squared deviations point to the mean as center in a way that distances do not. Second, the standard deviation turns out to be the natural measure of spread for a particularly important class of symmetric unimodal distributions, the *normal distributions*. We will meet the normal distributions in the next section. We commented earlier that the usefulness of many statistical procedures is tied to distributions of particular shapes. This is distinctly true of the standard deviation.

*Why do we emphasize the standard deviation rather than the variance?* One reason is that $s$, not $s^2$, is the natural measure of spread for normal distributions. There is also a more general reason to prefer $s$ to $s^2$. Because the variance involves squaring the deviations, it does not have the same unit of measurement as the original observations. The variance of the metabolic rates, for example, is measured in squared calories. Taking the square root remedies this. The standard deviation $s$ measures spread about the mean in the original scale.

*Why do we average by dividing by $n - 1$ rather than $n$ in calculating the variance?* Because the sum of the deviations is always zero, the last deviation can be found once we know the other $n - 1$. So we are not averaging $n$ unrelated numbers. 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 $n - 1$.

degrees of freedom

## Properties of the standard deviation

Here are the basic properties of the standard deviation $s$ as a measure of spread.

### Properties 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.
- $s$, like the mean $\bar{x}$, is not resistant. A few outliers can make $s$ very large.

The use of squared deviations renders $s$ even more sensitive than $\bar{x}$ to a few extreme observations. For example, dropping the Honda Insight from our list of two-seater cars reduces the mean highway mileage from 25.8 MPG to 23.4 MPG. It cuts the standard deviation more than half, from 11.4 MPG with the Insight to 5.3 MPG without it. Distributions with outliers and strongly skewed distributions have large standard deviations. The number $s$ does not give much helpful information about such distributions.

## 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 strong outliers. Use $\bar{x}$ and $s$ only for reasonably symmetric distributions that are free of outliers.

EXAMPLE 1.19

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 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.

Are $\bar{x}$ and $s$ good summaries for distributions of investment returns? Figure 1.18 displays stemplots of the annual returns for both investments. (Because stock returns are so much more spread out, 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 modes or gaps, for example. **Always plot your data**.

## Changing the unit of measurement

The same variable can be recorded in different units of measurement. Americans commonly record distances in miles and temperatures in degrees Fahrenheit, while the rest of the world measures distances in kilometers and temperatures in degrees Celsius. Fortunately, it is easy to convert numerical descriptions of a distribution from one unit of measurement to another. This is true because a change in the measurement unit is a *linear transformation* of the measurements.

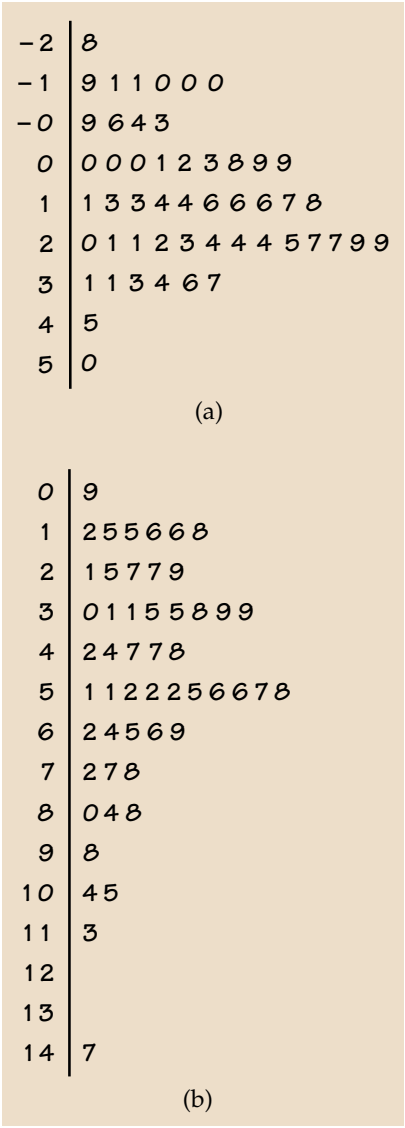



FIGURE 1.18 Stemplots of annual returns for stocks and Treasury bills, 1950 to 2000. (a) Stock returns, in whole percents. (b) Treasury bill returns, in percents and tenths of a percent.

## Linear Transformations

A **linear transformation** changes the original variable $x$ into the new variable $x_{\text{new}}$ given by an equation of the form

$$x_{\text{new}} = a + bx$$

Adding the constant $a$ shifts all values of $x$ upward or downward by the same amount. In particular, such a shift changes the origin (zero point) of the variable. Multiplying by the positive constant $b$ changes the size of the unit of measurement.

EXAMPLE 1.20

**(a)** If a distance $x$ is measured in kilometers, the same distance in miles is

$$x_{\text{new}} = 0.62x$$

For example, a 10-kilometer race covers 6.2 miles. This transformation changes the units without changing the origin—a distance of 0 kilometers is the same as a distance of 0 miles.

**(b)** A temperature $x$ measured in degrees Fahrenheit must be reexpressed in degrees Celsius to be easily understood by the rest of the world. The transformation is

$$x_{\text{new}} = \frac{5}{9}(x - 32) = -\frac{160}{9} + \frac{5}{9}x$$

Thus, the high of 95° F on a hot American summer day translates into 35° C. In this case

$$a = -\frac{160}{9} \quad \text{and} \quad b = \frac{5}{9}$$

This linear transformation changes both the unit size and the origin of the measurements. The origin in the Celsius scale (0° C, the temperature at which water freezes) is 32° in the Fahrenheit scale.

*Linear transformations do not change the shape of a distribution*. If measurements on a variable $x$ have a right-skewed distribution, any new variable $x_{\text{new}}$ obtained by a linear transformation $x_{\text{new}} = a + bx$ (for $b > 0$) will also have a right-skewed distribution. If the distribution of $x$ is symmetric and unimodal, the distribution of $x_{\text{new}}$ remains symmetric and unimodal.

Although a linear transformation preserves the basic shape of a distribution, the center and spread will change. Because linear changes of measurement scale are common, we must be aware of their effect on numerical descriptive measures of center and spread. Fortunately, the changes follow a simple pattern.

EXAMPLE 1.21

Mary and John both measure the weights of the same five newly hatched pythons. Mary measures in ounces and John uses grams. There are 28.35 grams in an ounce, so each of John's measurements is 28.35 times as large as Mary's measurement of the same python. Here are their results:

| Python | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Mary | 1.13 oz | 1.02 oz | 1.23 oz | 1.06 oz | 1.16 oz |
| John | 32 g | 29 g | 35 g | 30 g | 33 g |

These two sets of numbers measure the same five weights in different units. Their means represent the same average weight in different units, so John's mean (in grams) must be 28.35 times as large as Mary's mean (in ounces). The spread of John's numbers is likewise 28.35 times the spread of Mary's, so John's standard deviation (again in grams) is 28.35 times Mary's standard deviation (in ounces). You can verify these facts (up to roundoff error) with your calculator.

Here is a summary of such common sense relations in more formal language.

### Effect of a Linear Transformation

To see the effect of a linear transformation on measures of center and spread, apply these rules:

- Multiplying each observation by a positive number $b$ multiplies both measures of center (mean and median) and measures of spread (interquartile range and standard deviation) by $b$.
- Adding the same number $a$ (either positive or negative) to each observation adds $a$ to measures of center and to quartiles and other percentiles but does not change measures of spread.

The measures of spread $IQR$ and $s$ do not change when we add the same number $a$ to all of the observations because adding a constant changes the location of the distribution but leaves the spread unaltered. You can find the effect of a linear transformation $x_{\text{new}} = a + bx$ by combining these rules. For example, if $x$ has mean $\overline{x}$, the transformed variable $x_{\text{new}}$ has mean $a + b\overline{x}$.

## SUMMARY

A numerical summary of a distribution should report its **center** and its **spread** or **variability**.

The **mean** $\overline{x}$ 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 their midpoint.

When you use the median to describe 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 **interquartile range** is the difference between the quartiles. It is the spread of the center half of the data. The **1.5 × *IQR* criterion** flags observations more than $1.5 \times IQR$ beyond the quartiles as possible outliers.

The **five-number summary** consisting of the median, the quartiles, and the smallest and largest individual observations 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. In a **modified boxplot**, points identified by the $1.5 \times IQR$ criterion are plotted individually.

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.

**Linear transformations** have the form $x_{\text{new}} = a + bx$. A linear transformation changes the origin if $a \neq 0$ and changes the size of the unit of measurement if $b > 0$. Linear transformations do not change the overall shape of a distribution. A linear transformation multiplies a measure of spread by $b$ and changes a percentile or measure of center $m$ into $a + bm$.

Numerical measures of particular aspects of a distribution, such as center and spread, do not report the entire shape of most distributions. In some cases, particularly distributions with multiple peaks and gaps, these measures may not be very informative.

## SECTION 1.2 EXERCISES

**1.41** Figure 1.12 (page 27) is a histogram of the tuition and fees charged by all 56 four-year colleges in the state of Massachusetts. Here are those charges (in dollars), arranged in increasing order:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2,508 | 2,830 | 2,883 | 2,914 | 3,018 | 3,044 | 3,357 | 4,129 | 4,222 | 4,255 |
| 5,212 | 11,894 | 12,520 | 13,170 | 13,600 | 13,660 | 13,800 | 14,062 | 14,645 | 14,720 |
| 14,800 | 15,500 | 15,504 | 15,618 | 15,830 | 16,412 | 16,450 | 16,775 | 17,270 | 17,300 |
| 17,320 | 17,410 | 17,500 | 17,820 | 18,320 | 18,910 | 19,590 | 20,014 | 20,890 | 23,262 |
| 23,485 | 23,520 | 23,584 | 23,815 | 24,154 | 24,450 | 24,790 | 24,850 | 25,044 | 25,128 |
| 25,360 | 25,714 | 26,050 | 26,080 | 26,125 | 26,241 | | | | |

Find the five-number summary and make a boxplot. What distinctive feature of the histogram do these summaries miss? Remember that numerical summaries are not a substitute for looking at the data.

*The INDIVIDUALS data set, described in the Data Appendix, contains the income and level of education of 55,899 people between the ages of 25 and 65. The boxplots in Figure 1.19 compare the distributions of income for people with six levels of education. This figure is a variation on the boxplot idea: because large data sets often contain very extreme observations, the lines extend from the central box only to the 5th and 95th percentiles. Exercises 1.42 to 1.44 concern these data.*

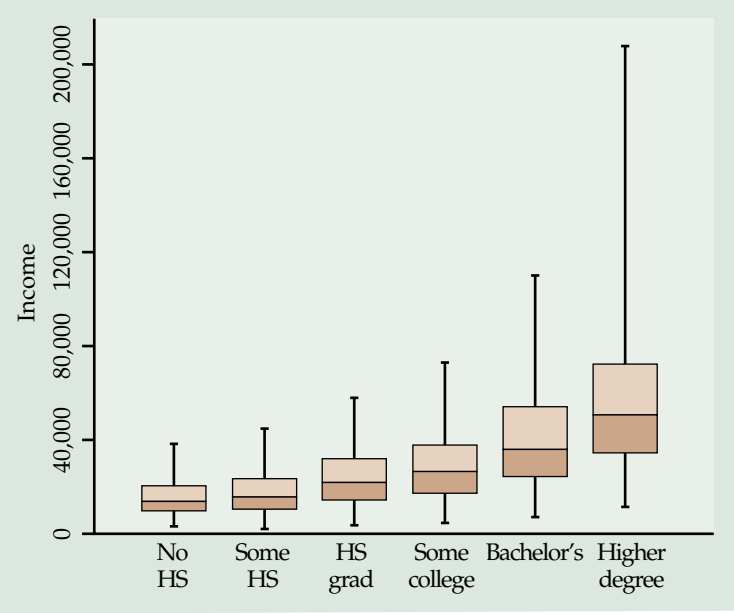



FIGURE 1.19 Boxplots comparing the distributions of income for people aged 25 to 65 years with six different levels of education. The lines extend from the quartiles to the 5th and 95th percentiles.

**1.42** The data include 10,991 people whose highest level of education is a bachelor's degree.

**(a)** What is the position of the median in the ordered list of incomes (1 to 10,991)? From the plot, about what is the median income of people with a bachelor's degree?

**(b)** What is the position of the first and third quartiles in the ordered list of incomes for these people? About what are the numerical values of $Q_1$ and $Q_3$?

**1.43** About what are the positions of the 5th and 95th percentiles in the ordered list of incomes of people with a bachelor's degree? Incomes outside this range do not appear in the boxplot. About what are the numerical values of the 5th and 95th percentiles of income? (For comparison, the largest income among these 10,991 people was $419,304. That one person made this much tells us less about the group than does the 95th percentile.)

**1.44** Write a brief description of how the distribution of income changes with the highest level of education reached. Be sure to discuss center, spread, and skewness. Give some specifics read from the graph to back up your statements.

**1.45** 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:[23]

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 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 |

**(a)** 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?

**(b)** Which observations are suspected outliers by the $1.5 \times IQR$ criterion? Which observations would you call outliers based on the stemplot?

**1.46** Stemplots help you find the five-number summary because they arrange the observations in increasing order. Exercise 1.15 (page 25) includes a stemplot of the percent of residents aged 25 to 34 in each of the 50 states.

**(a)** Find the five-number summary of this distribution.

**(b)** Does the $1.5 \times IQR$ criterion flag Montana and Wyoming as suspected outliers?

**(c)** How much does the median change if you omit Montana and Wyoming?

**1.47** Look at the histogram of lengths of words in Shakespeare's plays, Figure 1.11 (page 26). The heights of the bars tell us what percent of words have each length. What is the median length of words used by Shakespeare? Similarly, what are the quartiles? Give the five-number summary for Shakespeare's word lengths.

**1.48** Exercise 1.21 (page 27) gives data on the total oil recovered from 64 wells. Your graph in that exercise shows that the distribution is strongly right-skewed.

**(a)** 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.

**(b)** Give the five-number summary and explain briefly how it reflects the shape of the distribution.

**1.49** Here are the scores on the Survey of Study Habits and Attitudes (SSHA) for 18 first-year college women:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 154 | 109 | 137 | 115 | 152 | 140 | 154 | 178 | 101 |
| 103 | 126 | 126 | 137 | 165 | 165 | 129 | 200 | 148 |

and for 20 first-year college men:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 108 | 140 | 114 | 91 | 180 | 115 | 126 | 92 | 169 | 146 |
| 109 | 132 | 75 | 88 | 113 | 151 | 70 | 115 | 187 | 104 |

**(a)** Make a back-to-back stemplot of these data, or use your result from Exercise 1.23.

**(b)** Find the mean $\overline{x}$ and the median $M$ for both sets of SSHA scores. What feature of each distribution explains the fact that $\overline{x} > M$?

**(c)** Find the five-number summaries for both sets of SSHA scores. Your plot in (a) suggests that there is an outlier among the women's scores. Does the $1.5 \times IQR$ criterion flag this observation? Make side-by-side modified boxplots for the two distributions.

**(d)** Use your results to write a brief comparison of the two groups. Do women as a group score higher than men? Which of your descriptions (stemplots, boxplots, numerical measures) show this? Which group of scores is more spread out when we ignore outliers? Which of your descriptions shows this most clearly?

**1.50** The SSHA data for women given in the previous exercise contain one high outlier. Calculate the mean $\overline{x}$ and the median $M$ for these data with and without the outlier. How does removing the outlier affect $\overline{x}$? How does it affect $M$? Your results illustrate the greater resistance of the median.


**1.51** The *Mean and Median* applet allows you to place observations on a line and see their mean and median visually. Place two observations on the line. Why does only one arrow appear?


**1.52** In the *Mean and Median* applet, place three observations on the line, two close together near the center of the line, and one somewhat to the right of these two.

**(a)** 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.

**(b)** Now drag the rightmost 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)?

**1.53** Table 1.3 (page 31) gives the number of medical doctors per 100,000 people in each state. Your graph of the distribution in Exercise 1.28 shows that the District of Columbia (D.C.) is a high outlier. Because D.C. is a city rather than a state, we will omit it here.

**(a)** Calculate both the five-number summary and $\overline{x}$ and $s$ for the number of doctors per 100,000 people in the 50 states. Based on your graph, which description do you prefer?

**(b)** 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.

**1.54** Here are the percents of the popular vote won by the successful candidate in each U.S. presidential election from 1948 to 2000:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Year | 1948 | 1952 | 1956 | 1960 | 1964 | 1968 | 1972 |
| Percent | 49.6 | 55.1 | 57.4 | 49.7 | 61.1 | 43.4 | 60.7 |
| Year | 1976 | 1980 | 1984 | 1988 | 1992 | 1996 | 2000 |
| Percent | 50.1 | 50.7 | 58.8 | 53.9 | 43.2 | 49.2 | 47.9 |

**(a)** Make a graph to display the distribution of winners' percents. What are the main features of this distribution?

**(b)** What is the median percent of the vote won by the successful candidate in presidential elections?

**(c)** Call an election a landslide if the winner's percent falls at or above the third quartile. Which elections were landslides?

*Table 1.9 shows* Consumer Reports *magazine's laboratory test results for calories and milligrams of sodium (mostly due to salt) in a number of major brands of hot dogs.*[24] *There are three types: all beef, "meat" (mainly pork and beef, but government regulations allow up to 15% poultry meat), and poultry. Exercises 1.55 to 1.57 analyze these data.*

**1.55** Find the five-number summaries of the calorie content of the three types of hot dogs. Then use the $1.5 \times IQR$ criterion to check for suspected outliers.

TABLE 1.9 Calories and sodium in three types of hot dogs

| Beef hot dogs | | Meat hot dogs | | Poultry hot dogs | |
|---|---|---|---|---|---|
| Calories | Sodium | Calories | Sodium | Calories | Sodium |
| 186 | 495 | 173 | 458 | 129 | 430 |
| 181 | 477 | 191 | 506 | 132 | 375 |
| 176 | 425 | 182 | 473 | 102 | 396 |
| 149 | 322 | 190 | 545 | 106 | 383 |
| 184 | 482 | 172 | 496 | 94 | 387 |
| 190 | 587 | 147 | 360 | 102 | 542 |
| 158 | 370 | 146 | 387 | 87 | 359 |
| 139 | 322 | 139 | 386 | 99 | 357 |
| 175 | 479 | 175 | 507 | 170 | 528 |
| 148 | 375 | 136 | 393 | 113 | 513 |
| 152 | 330 | 179 | 405 | 135 | 426 |
| 111 | 300 | 153 | 372 | 142 | 513 |
| 141 | 386 | 107 | 144 | 86 | 358 |
| 153 | 401 | 195 | 511 | 143 | 581 |
| 190 | 645 | 135 | 405 | 152 | 588 |
| 157 | 440 | 140 | 428 | 146 | 522 |
| 131 | 317 | 138 | 339 | 144 | 545 |
| 149 | 319 | | | | |
| 135 | 298 | | | | |
| 132 | 253 | | | | |

Make modified boxplots to compare the three distributions. Write a brief discussion of your findings.

**1.56** Make a stemplot of the calorie content of the 17 brands of meat hot dogs. What is the most important feature of the overall pattern of the distribution? Are there any outliers? Note that the five-number summary misses the big feature of this distribution. Routine numerical summaries are never a substitute for looking at the data.

**1.57** Use graphs and numerical summaries to compare the sodium content of the three types of hot dogs. Write a summary of your findings suitable for readers who know no statistics. Can we hold down our sodium intake by buying poultry hot dogs?

**1.58** Exercise 1.25 (page 29) presented data on the nightly study time claimed by first-year college men and women. The most common methods for formal comparison of two groups use $\bar{x}$ and $s$ to summarize the data. We wonder if this is appropriate here. Look at your back-to-back stemplot from Exercise 1.25, or make one now if you have not done so.

**(a)** What kinds of distributions are best summarized by $\bar{x}$ and $s$? It isn't easy to decide whether small data sets with irregular distributions fit the criteria. We will learn a better tool for making this decision in the next section.

**(b)** Each set of study times appears to contain a high outlier. Are these points flagged as suspicious by the $1.5 \times IQR$ criterion? How much does removing the outlier change $\bar{x}$ and $s$ for each group? The presence of outliers makes us reluctant to use the mean and standard deviation for these data unless we remove the outliers on the grounds that these students were exaggerating.

**1.59** 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. What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary?

**1.60** In computing the median income of any group, some federal agencies omit all members of the group who had no income. Give an example to show that the reported median income of a group can go down even though the group becomes economically better off. Is this also true of the mean income?

**1.61** This year, the firm in Exercise 1.59 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?

**1.62** Calculate the mean and standard deviation of the metabolic rates in Example 1.18 (page 48), showing each step in detail. First find the mean $\bar{x}$ by summing the 7 observations and dividing by 7. Then find each of the deviations $x_i - \bar{x}$ and their squares. Check that the deviations have sum 0. Calculate the variance as an average of the squared deviations (remember to divide by $n - 1$). Finally, obtain $s$ as the square root of the variance.

**1.63** Give an example of a small set of data for which the mean is larger than the third quartile.

**1.64** Create a set of 5 positive numbers (repeats allowed) that have median 10 and mean 7. What thought process did you use to create your numbers?

**1.65** This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 10, with repeats allowed.

**(a)** Choose four numbers that have the smallest possible standard deviation.

**(b)** Choose four numbers that have the largest possible standard deviation.

**(c)** Is more than one choice possible in either (a) or (b)? Explain.

**1.66** Use the definition of the mean $\bar{x}$ to show that the sum of the deviations $x_i - \bar{x}$ of the observations from their mean is always zero. This is one reason why the variance and standard deviation use squared deviations.

**1.67** This exercise requires a calculator with a standard deviation button, or statistical software on a computer. The observations

$$10{,}001 \quad 10{,}002 \quad 10{,}003$$

have mean $\bar{x} = 10{,}002$ and standard deviation $s = 1$. Adding a 0 in the center of each number, the next set becomes

$$100{,}001 \quad 100{,}002 \quad 100{,}003$$

The standard deviation remains $s = 1$ as more 0s are added. Use your calculator or computer to calculate the standard deviation of these numbers, adding extra 0s until you get an incorrect answer. How soon did you go wrong? This demonstrates that calculators and computers cannot handle an arbitrary number of digits correctly.

**1.68** We saw in Example 1.19 that it is usual in the study of investments to use the mean and standard deviation to summarize and compare investment returns. Table 1.4 (page 31) gives the monthly returns on one company's stock for 83 consecutive months.

**(a)** Find the mean monthly return and the standard deviation of the returns. 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?

**(b)** The distribution can be described as "symmetric and unimodal, with one low outlier." If you invested \$100 in this stock at the beginning of the worst month in the data (the outlier), how much would you have at the end of the month? Find the mean and standard deviation again, this time leaving out the low outlier. How much did this one observation affect the summary measures? Would leaving out this one observation 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

do. So use of the mean and standard deviation is better justified for such data.)

**1.69** Table 1.5 (page 32) gives the survival times of 72 guinea pigs in a medical study. Survival times—whether of cancer patients after treatment or of car batteries in everyday use—are almost always right-skewed. Make a graph to verify that this is true of these survival times. Then give a numerical summary that is appropriate for such data. Explain in simple language, to someone who knows no statistics, what your summary tells us about the guinea pigs.

**1.70** Find the 10th and 90th percentiles of the distribution of doctors per 100,000 population in the states, from Table 1.3 (page 31). Which states are in the top 10%? In the bottom 10%?

quintiles

**1.71** Find the **quintiles** (the 20th, 40th, 60th, and 80th percentiles) of the IQ scores in Table 1.6 (page 33). For quite large sets of data, the quintiles or the **deciles** (10th, 20th, 30th, etc. percentiles) give a more detailed summary than the quartiles.

deciles

**1.72** In each of the following settings, give the values of $a$ and $b$ for the linear transformation $x_{\text{new}} = a + bx$ that expresses the change in units of measurement.

**(a)** Change a speed $x$ measured in miles per hour into the metric system value $x_{\text{new}}$ in kilometers per hour. (A kilometer is 0.62 mile.) What is 65 miles per hour in metric units?

**(b)** You are writing a report on the power of car engines. Your sources use horsepower $x$. Reexpress power in watts $x_{\text{new}}$. (One horsepower is 746 watts.) What is the power in watts of a 140-horsepower engine?

**1.73** In each of the following settings, give the values of $a$ and $b$ for the linear transformation $x_{\text{new}} = a + bx$ that expresses the change in units of measurement.

**(a)** You want to restate water temperature $x$ in a swimming pool, measured in degrees Fahrenheit, as the difference $x_{\text{new}}$ between $x$ and the "normal" body temperature of 98.6 degrees.

**(b)** The recommended daily allowance (RDA) for vitamin C was recently increased to 120 milligrams. You measure milligrams of vitamin C in foods and want to convert your results to percent of RDA.

**1.74** Henry Cavendish (see Exercise 1.27 on page 30) used $\bar{x}$ to summarize his 29 measurements of the density of the earth.

**(a)** Find $\bar{x}$ and $s$ for his data.

**(b)** Cavendish recorded the density of the earth as a multiple of the density of water. The density of water is almost exactly 1 gram per cubic centimeter, so his measurements have these units. In American units, the density of water is 62.43 pounds per cubic foot. This is the weight of a cube of water measuring 1 foot (that is, 30.48 cm) on each side. Express Cavendish's first result for the earth (5.50 g/cm$^3$) in pounds per cubic foot. Then find $\bar{x}$ and $s$ in pounds per cubic foot.

**1.75** A change of units that multiplies each unit by $b$, such as the change $x_{\text{new}} = 2.54x$ from inches $x$ to centimeters $x_{\text{new}}$, multiplies our usual measures of spread by $b$. This is true of *IQR* and the standard deviation. What happens to the variance when we change units in this way?

trimmed mean **1.76** The **trimmed mean** is a measure of center that is more resistant than the mean but uses more of the available information than the median. To compute the 10% trimmed mean, discard the highest 10% and the lowest 10% of the observations and compute the mean of the remaining 80%. Trimming eliminates the effect of a small number of outliers. Compute the 10% trimmed mean of the guinea pig survival time data in Table 1.5 (page 32). Then compute the 20% trimmed mean. Compare the values of these measures with the median and the ordinary untrimmed mean.

## 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 stemplot or a histogram.
2. Look for the overall pattern and for striking deviations such as outliers.
3. Calculate an appropriate 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.

Figure 1.20 is a histogram of the scores of all 947 seventh-grade students in Gary, Indiana, on the vocabulary part of the Iowa Test of Basic Skills. Scores of many students on this national test have a very regular distribution. The histogram is symmetric, and both tails fall off quite smoothly from a single center peak. There are no large gaps or obvious outliers. The smooth curve drawn through the tops of the histogram bars in Figure 1.20 is a good description of the overall pattern of the data. The curve 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.

mathematical model

The use of mathematical models is a common and powerful tool in statistics, as in other sciences. Suppose, for example, that we are watching a caterpillar crawl forward. Once each minute we measure the distance it has moved. When we plot the observed values against time, we see that the points fall close to a straight line. It would be natural to draw a straight line on our graph as a compact description of these data. The straight line, with its equation of the form $y = a + bt$ (where $y$ is distance and $t$ is time), is a mathematical model

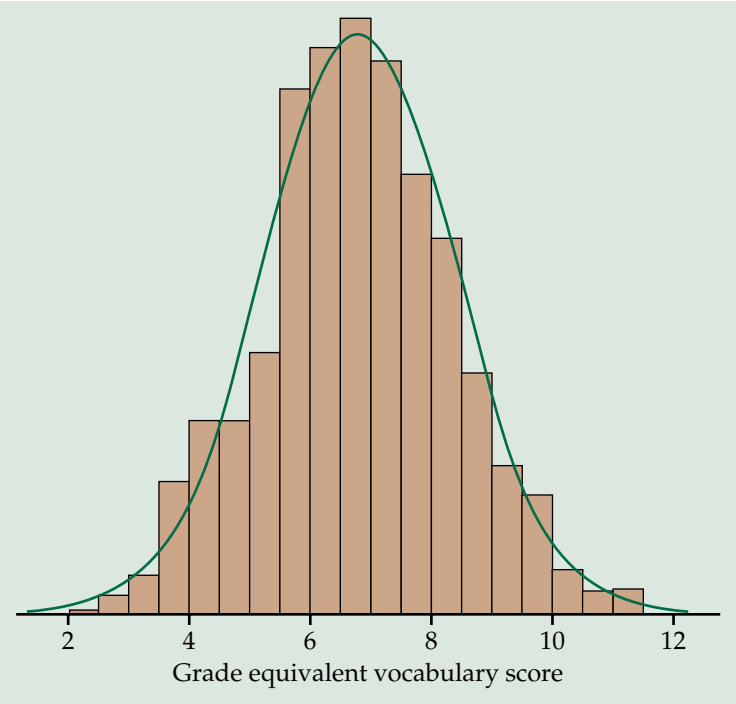



FIGURE 1.20 Histogram of the Iowa Test vocabulary scores for Gary, Indiana, seventh graders, showing the approximation of the distribution by a normal curve.

of the caterpillar's progress. It is an idealized description because the points on the graph do not fall *exactly* on the line. We want to give a similarly compact description of the distribution of a quantitative variable, to replace stemplots and histograms. Just as a straight line is one of many types of curves that can be used to describe a plot of distance traveled against time, there are many types of mathematical distributions that can be used to describe a set of single-variable data. This section will concentrate on one type, the normal distributions.

## Density curves

We will see that it is easier to work with the smooth curve in Figure 1.20 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 the idea.

EXAMPLE 1.22

In a histogram, the *areas* of the bars represent either counts or proportions of the observations. Figure 1.21(a) is a copy of Figure 1.20 with the leftmost bars shaded. The area of the shaded bars in Figure 1.21(a) represents the students with vocabulary scores 6.0 or lower. There are 287 such students, who make up the proportion $287/947 = 0.303$ of all Gary seventh graders.

In Figure 1.21(b), the area under the smooth curve to the left of 6.0 is shaded. Adjust the scale so that the total area under the curve is exactly 1. Areas under the curve

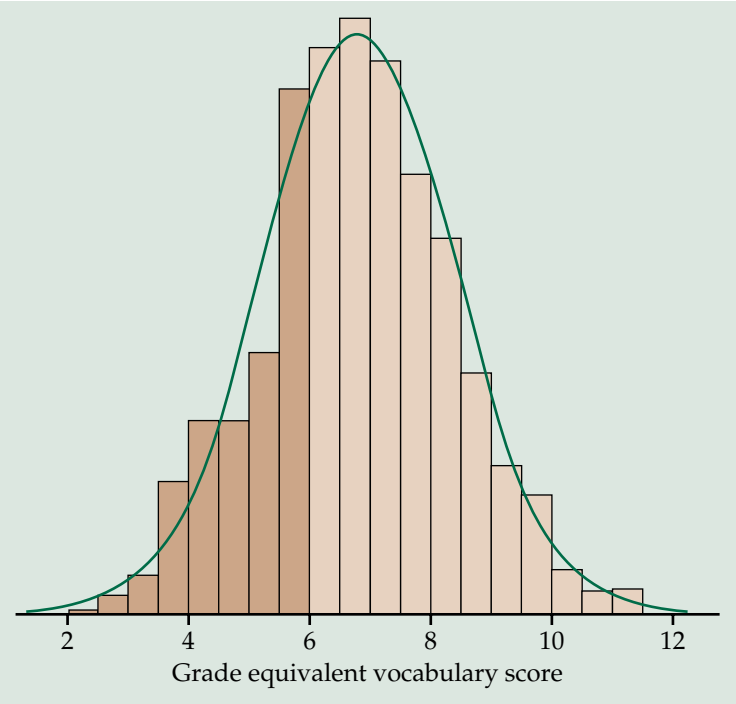



FIGURE 1.21(a) The relative frequency of scores less than or equal to 6.0 from the histogram is 0.303.

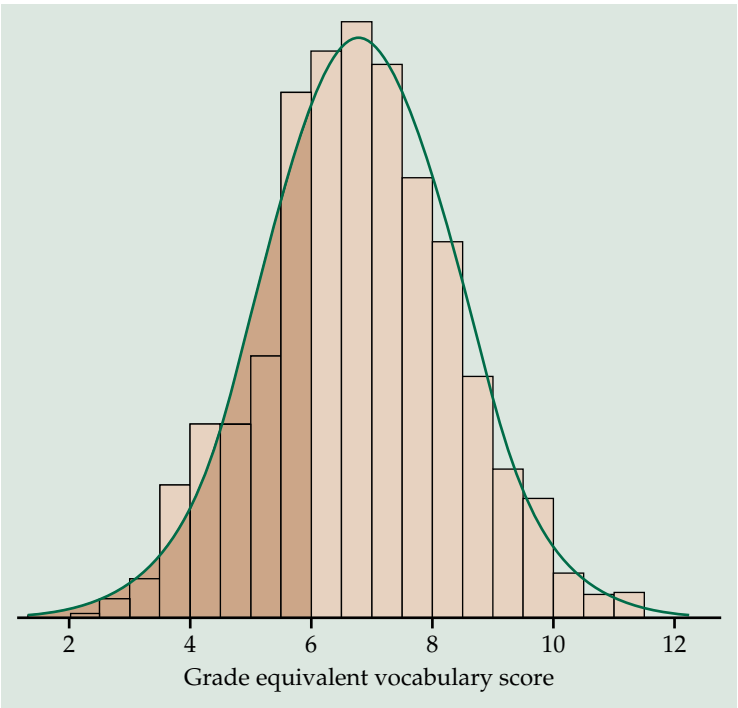



FIGURE 1.21(b) The relative frequency of scores less than or equal to 6.0 from the density curve is 0.293.

then represent proportions of the observations. That is, *area = relative frequency*. The curve is then a *density curve*. The shaded area under the density curve in Figure 1.21(b) represents the proportion of students with score 6.0 or lower. This area is 0.293, only 0.010 away from the histogram result. You can see that areas under the density curve give quite good approximations of areas given by the histogram.

### 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 relative frequency of all observations that fall in that range.

The density curve in Figures 1.20 and 1.21 is a *normal curve*. Density curves, like distributions, come in many shapes. Figure 1.22 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.

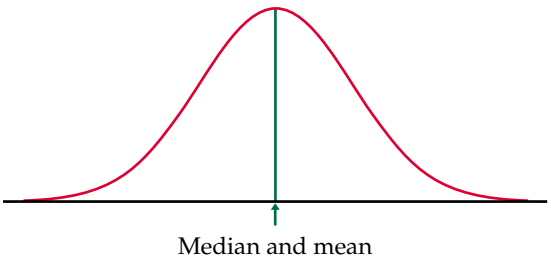



FIGURE 1.22(a) A symmetric density curve with its mean and median marked.

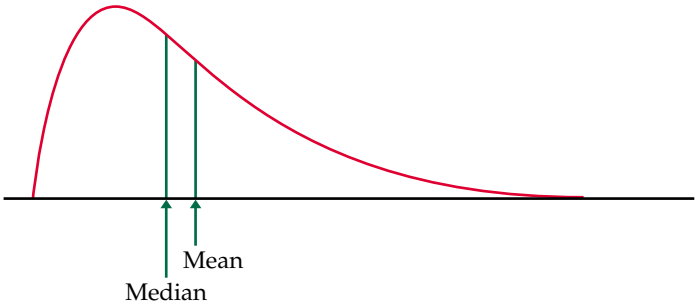



FIGURE 1.22(b) A right-skewed density curve with its mean and median marked.

## Measuring center and spread for density curves

Our measures of center and spread apply to density curves as well as to actual sets of observations, but only some of these measures are easily seen from the curve. A **mode** of a distribution described by a density curve is a peak point of the curve, the location where the curve is highest. Because areas under a density curve represent proportions of the observations, the **median** is the point with half the total area on each side. You can roughly locate the **quartiles** by dividing the area under the curve into quarters as accurately as possible by eye. The *IQR* is then the distance between the first and third quartiles. There are mathematical ways of calculating areas under curves. These allow us to locate the median and quartiles exactly on any density curve.

What about the mean and standard deviation? 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 it were made out of solid material. Figure 1.23 illustrates this interpretation of the mean. We have marked the mean and median on the density curves in Figure 1.22. A symmetric curve, such as the normal curve in Figure 1.22(a), balances at its center of symmetry. Half the area under a symmetric curve lies on either side of its center, so this is also the median. For a right-skewed curve, such as that shown in Figure 1.22(b), the small area in the long right tail tips the curve more than the same area near the center. The mean (the balance point) therefore lies to the right of the median. It is 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 1.22(b). The standard deviation can also be calculated mathematically, but it can't be located by eye on most density curves.

### 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.

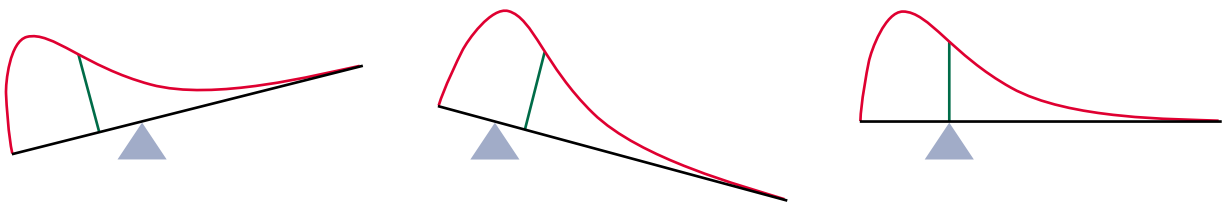


FIGURE 1.23 The mean of a density curve is the point at which it would balance.

A density curve is an idealized mathematical model for a distribution of data. For example, the symmetric density curve in Figures 1.20 and 1.21 is exactly symmetric, but the histogram of vocabulary scores is only approximately symmetric. We therefore need to distinguish between the mean and standard deviation of the density curve and the numbers $\bar{x}$ and $s$ computed from the actual observations. The usual notation for the mean of an idealized distribution is $\mu$ (the Greek letter mu). We write the standard deviation of a density curve as $\sigma$ (the Greek letter sigma).

mean $\mu$
standard deviation $\sigma$

## Normal distributions

One particularly important class of density curves has already appeared in Figures 1.20 and 1.22(a). These density curves are symmetric, unimodal, 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 specified by giving its mean $\mu$ and its standard deviation $\sigma$. The mean is located at the center of the symmetric curve and is the same as the median. Changing $\mu$ without changing $\sigma$ moves the normal curve along the horizontal axis without changing its spread. The standard deviation $\sigma$ controls the spread of a normal curve. Figure 1.24 shows two normal curves with different values of $\sigma$. The curve with the larger standard deviation is more spread out.

normal curves

The standard deviation $\sigma$ is the natural measure of spread for normal distributions. Not only do $\mu$ and $\sigma$ completely determine the shape of a normal curve, but we can locate $\sigma$ by eye on the curve. Here's how. As we move out in either direction from the center $\mu$, the curve changes from falling ever more steeply

to falling ever less steeply

*The points at which this change of curvature takes place are located at distance $\sigma$ on either side of the mean $\mu$.* You can feel the change as you run a pencil along a normal curve, and so find the standard deviation. Remember that $\mu$

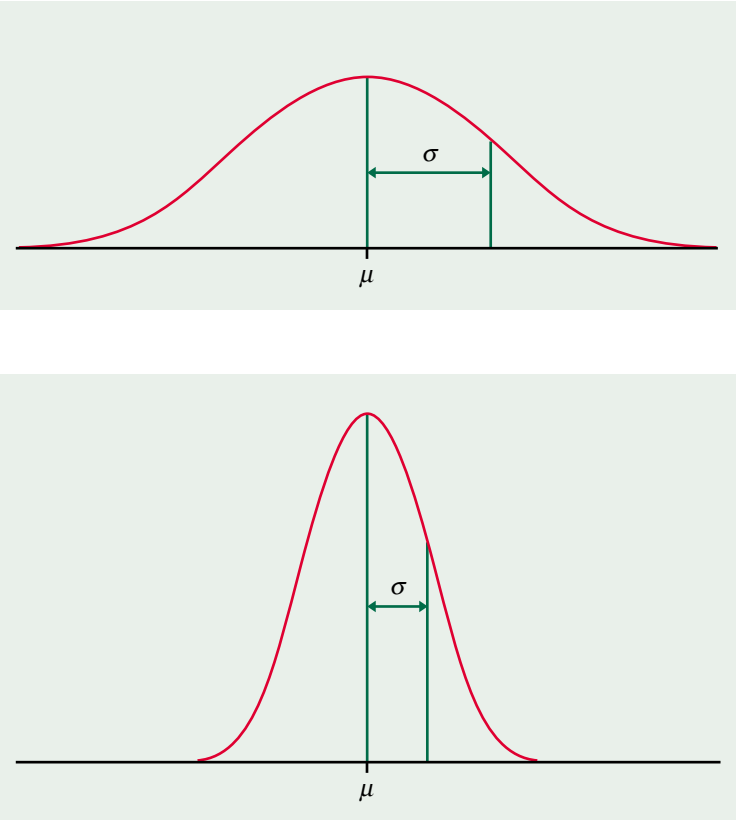



FIGURE 1.24 Two normal curves, showing the mean $\mu$ and standard deviation $\sigma$.

and $\sigma$ alone do not specify the shape of most distributions, and that the shape of density curves in general does not reveal $\sigma$. These are special properties of normal distributions.

There are other symmetric bell-shaped density curves that are not normal. The normal density curves are specified by a particular equation. The height of the density curve at any point $x$ is given by

$$\frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

We will not make direct use of this fact, although it is the basis of mathematical work with normal distributions. Note that the equation of the curve is completely determined by the mean $\mu$ and the standard deviation $\sigma$.

Why are the normal distributions important in statistics? Here are three reasons. First, normal distributions are good descriptions for some distributions of *real data*. Distributions that are often close to normal include scores on tests taken by many people (such as SAT exams and many psychological tests), repeated careful measurements of the same quantity, and characteristics

of biological populations (such as lengths of baby pythons and 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 income distributions, for example, are skewed to the right and so are not normal. Nonnormal 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. Here are some of the most important.

### The 68–95–99.7 Rule

In the normal distribution with mean $\mu$ and standard deviation $\sigma$:

- Approximately **68%** of the observations fall within $\sigma$ of the mean $\mu$.
- Approximately **95%** of the observations fall within $2\sigma$ of $\mu$.
- Approximately **99.7%** of the observations fall within $3\sigma$ of $\mu$.

Figure 1.25 illustrates the 68–95–99.7 rule. By remembering these three numbers, you can think about normal distributions without constantly making detailed calculations.

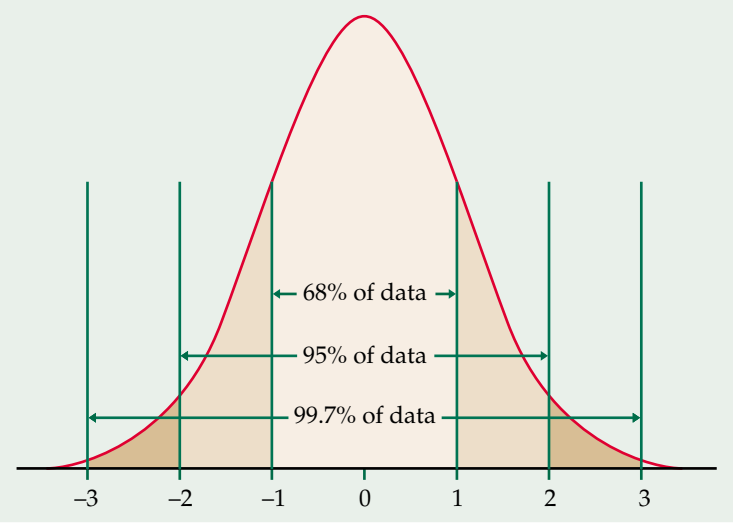



FIGURE 1.25 The 68–95–99.7 rule for normal distributions.

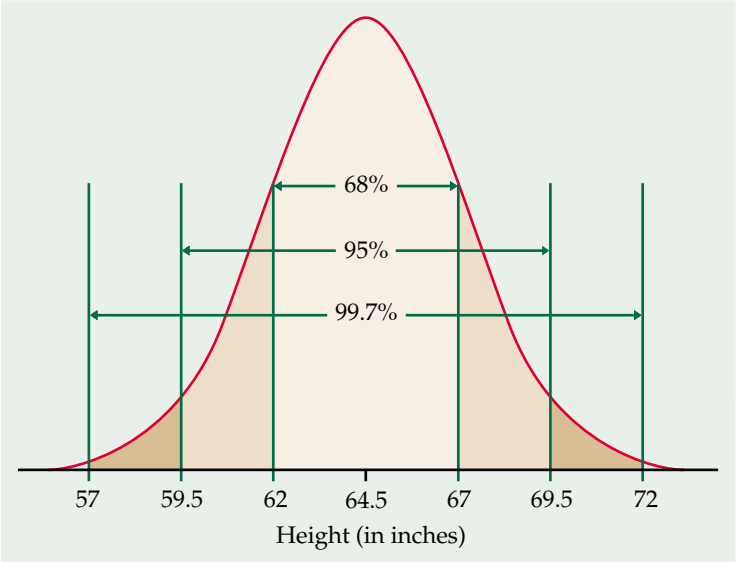



FIGURE 1.26 The 68–95–99.7 rule applied to the heights of young women.

EXAMPLE 1.23

The distribution of heights of young women aged 18 to 24 is approximately normal with mean $\mu = 64.5$ inches and standard deviation $\sigma = 2.5$ inches. Figure 1.26 shows what the 68–95–99.7 rule says about this distribution.

Two standard deviations is 5 inches for this distribution. The 95 part of the 68–95–99.7 rule says that the middle 95% of young women are between $64.5 - 5$ and $64.5 + 5$ inches tall, that is, between 59.5 inches and 69.5 inches. This fact is exactly true for an exactly normal distribution. It is approximately true for the heights of young women because the distribution of heights is approximately normal.

The other 5% of young women have heights outside the range from 59.5 to 69.5 inches. Because the normal distributions are symmetric, half of these women are on the tall side. So the tallest 2.5% of young women are taller than 69.5 inches.

$N(\mu, \sigma)$

Because we will mention normal distributions often, a short notation is helpful. We abbreviate the normal distribution with mean $\mu$ and standard deviation $\sigma$ as $N(\mu, \sigma)$. For example, the distribution of young women's heights is $N(64.5, 2.5)$.

## Standardizing observations

As the 68–95–99.7 rule suggests, all normal distributions share many properties. In fact, all normal distributions are the same if we measure in units of size $\sigma$ about the mean $\mu$ 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 $\mu$ and standard deviation $\sigma$, 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.

EXAMPLE 1.24

The heights of young women are approximately normal with $\mu = 64.5$ inches and $\sigma = 2.5$ inches. The standardized height is

$$z = \frac{\text{height} - 64.5}{2.5}$$

A woman's standardized height is the number of standard deviations by which her height differs from the mean height of all young women. A woman 68 inches tall, for example, has standardized height

$$z = \frac{68 - 64.5}{2.5} = 1.4$$

or 1.4 standard deviations above the mean. Similarly, a woman 5 feet (60 inches) tall has standardized height

$$z = \frac{60 - 64.5}{2.5} = -1.8$$

or 1.8 standard deviations less than the mean height.

We need a way to write variables, such as "height" in Example 1.24, that follow a theoretical distribution such as a normal distribution. We use capital letters near the end of the alphabet for such variables. If $X$ is the height of a young woman, we can then shorten "the height of a young woman is less than 68 inches" to "$X < 68$." We will use lowercase $x$ to stand for any specific value of the variable $X$.

We often standardize observations from symmetric distributions to express them in a common scale. We might, for example, compare the height of two children of different ages by calculating their $z$-scores. The standardized heights tell us where each child stands in the distribution for his or her age group.

Standardizing is a linear transformation that transforms the data into the standard scale of $z$-scores. We know that a linear transformation does not change the shape of a distribution, and that the mean and standard deviation change in a simple manner. In fact, *any variable obtained from a normal variable by a linear transformation remains normal*. Of course, the mean and

standard deviation will change in the usual way. If $X$ has the normal distribution with mean $\mu$ and standard deviation $\sigma$, then $X_{\text{new}} = a + bX$ for a positive $b$ has the normal distribution with mean $a + b\mu$ and standard deviation $b\sigma$. In particular, *the standardized values for any distribution always have mean 0 and standard deviation 1*. If the original distribution was normal, the standardized values have the normal distribution with mean 0 and standard deviation 1.

## The standard normal distribution

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*.

### The 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 $\mu$ and standard deviation $\sigma$, then the standardized variable

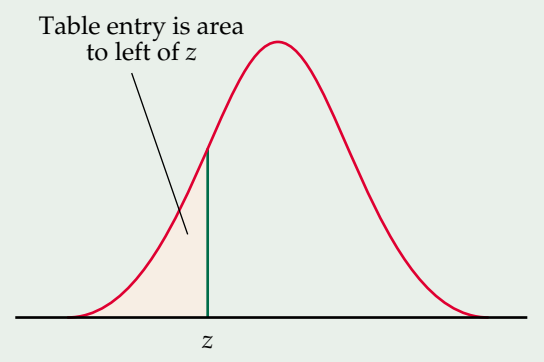
$$Z = \frac{X - \mu}{\sigma}$$

has the standard normal distribution.

Table A in the back of the book gives areas under the standard normal curve. Table A also appears on the inside front cover. You can use Table A to do normal calculations, although software or a statistical calculator is usually more efficient.

### 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 $z$.



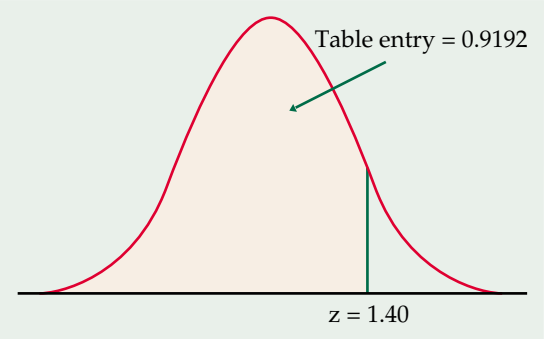



FIGURE 1.27(a) The area under a standard normal curve to the left of the point $z = 1.40$ is 0.9192. Table A gives areas under the standard normal curve.

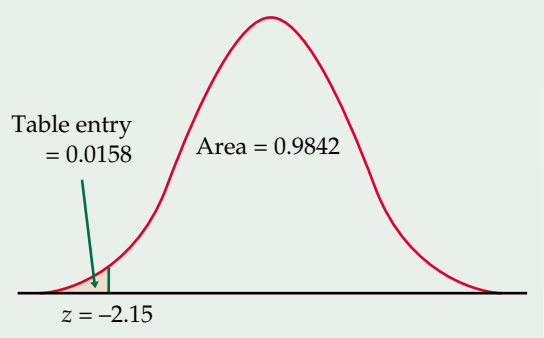



FIGURE 1.27(b) Areas under the standard normal curve to the right and left of $z = -2.15$. Table A gives the area to the left.

EXAMPLE 1.25

*Problem:* What proportion of observations on a standard normal variable $Z$ take values less than 1.4?

*Solution:* To find the area to the left of 1.40, locate 1.4 in the left-hand column of Table A, then locate the remaining digit 0 as .00 in the top row. The entry opposite 1.4 and under .00 is 0.9192. This is the area we seek. Figure 1.27(a) illustrates this area.

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

*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 1.27(b) shows these areas.

## Normal distribution calculations

We saw in Example 1.24 that standardizing the height $x = 68$ inches for a young woman gives the $z$-score $z = 1.4$. Example 1.25 then showed that the area under the standard normal curve to the left of $z = 1.4$ is 0.9192. This

is the relative frequency of $z$-scores less than 1.4. But the $z$-score is less than 1.4 whenever the height is less than 68 inches. So the proportion of all young women who are less than 68 inches tall is 0.9192, or about 92%. We can find relative frequencies for *any* normal distribution by standardizing and using Table A. Of course, software or a statistical calculator often automates the calculation. Here is an example.

EXAMPLE 1.26

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 in order to compete in their first college year. (Higher scores are required for students with poor high school grades.) In 2000, the scores of the more than one million students taking the SATs were approximately normal with mean 1019 and standard deviation 209. What percent of all students had SAT scores of at least 820?

**1.** *State the problem*. Call the SAT score of a randomly chosen student $X$. The variable $X$ has the $N(1019, 209)$ distribution. We want the proportion of students with $X > 820$.

**2.** *Standardize*. Subtract the mean, then divide by the standard deviation, to transform the problem about $X$ into a problem about a standard normal $Z$:

$$X > 820$$

$$\frac{X - 1019}{209} > \frac{820 - 1019}{209}$$

$$Z > -0.95$$

Figure 1.28 shows the standard normal curve with the area of interest shaded.

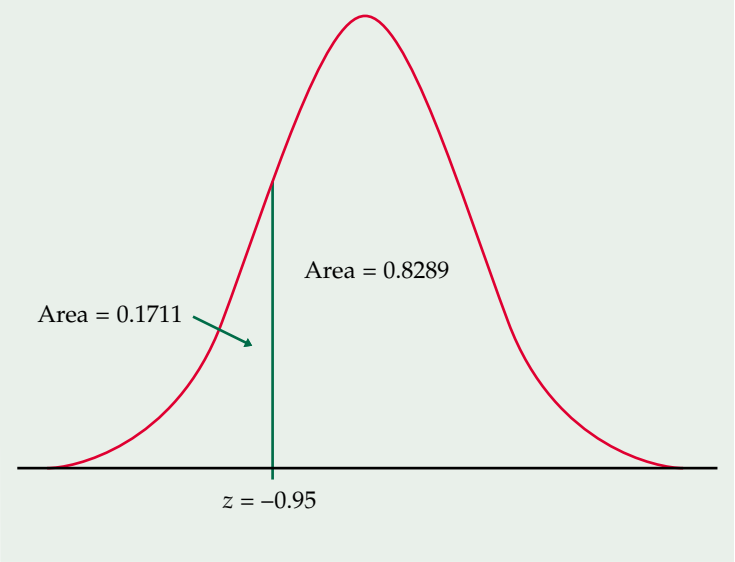



FIGURE 1.28 Areas under the standard normal curve for Example 1.26.

**3.** *Use the table*. From Table A, we see that the proportion of observations less than $-0.95$ is 0.1711. About 17% of students who take the SAT score less than 820. The area to the right of $-0.95$ is therefore $1 - 0.1711 = 0.8289$. This is about 0.83, or 83%. About 83% of all students taking the SAT would be academically qualified for Division I college athletics.

There is *no* area under a smooth curve and exactly over the point 820. Consequently, the area to the right of 820 (the relative frequency of $X > 820$) is the same as the area at or to the right of this point (the relative frequency of $X \geq 820$). The actual data may contain a student who scored exactly 820 on the SAT. The fact that it is not possible for a normal variable to have $X = 820$ exactly is a consequence of the idealized smoothing of normal models for data.

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.

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 area that the table or software gives you. Here is another example.

EXAMPLE 1.27

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. What percent of all students who take the SAT would be partial qualifiers?

**1.** *State the problem*. We want the proportion of SAT scores in the interval $720 \leq X < 820$.

**2.** *Standardize*.

$$\begin{gathered} 720 \leq X < 820 \\ \frac{720 - 1019}{209} \leq \frac{X - 1019}{209} < \frac{820 - 1019}{209} \\ -1.43 \leq Z < -0.95 \end{gathered}$$

Figure 1.29 shows the area under the standard normal curve. This picture guides our use of the table.

**3.** *Use the table*. The area between $-1.43$ and $-0.95$ is the area to the left of $-0.95$ *minus* the area to the left of $-1.43$. Look at Figure 1.29 to visualize this. From Table A,

$$\begin{aligned} \text{area between } -1.43 \text{ and } -0.95 &= (\text{area left of } -0.95) - (\text{area left of } -1.43) \\ &= 0.1711 - 0.0764 = 0.0947 \end{aligned}$$

About 9.5% of students taking the SAT would be partial qualifiers in the eyes of the NCAA.

Examples 1.26 and 1.27 illustrate the use of Table A to find the relative frequency of a given event, such as "SAT score between 720 and 820." We may instead want to find the observed value corresponding to a given relative

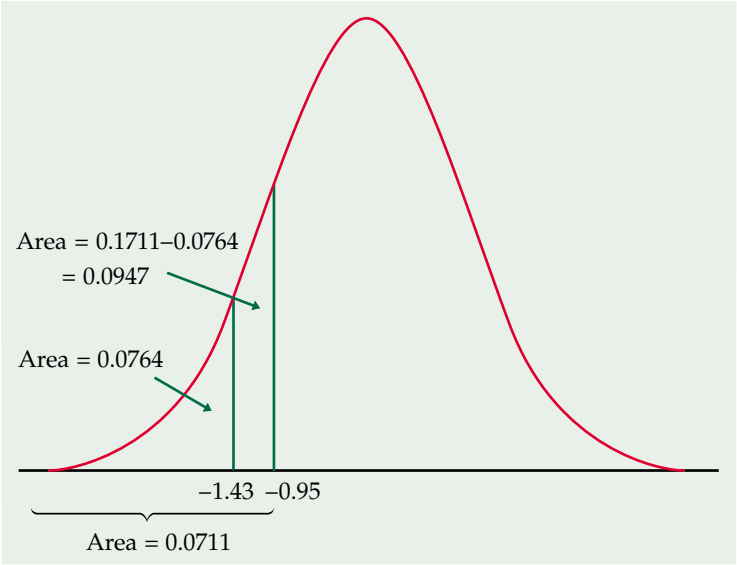



FIGURE 1.29 Areas under the standard normal curve for Example 1.27.

frequency. To do this, we use Table A backward, finding the desired relative frequency in the body of the table and then reading the corresponding $z$ from the left column and top row.

EXAMPLE 1.28

Scores on the SAT verbal test in recent years follow approximately the $N(505, 110)$ distribution. How high must a student score in order to place in the top 10% of all students taking the SAT?

1. *State the problem*. We want to find the SAT score $x$ with area 0.1 to its *right* under the normal curve with mean $\mu = 505$ and standard deviation $\sigma = 110$. That's the same as finding the SAT score $x$ with area 0.9 to its *left*. Figure 1.30 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 $x$.
2. *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.
3. *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

$$\frac{x - 505}{110} = 1.28$$

Solving this equation for $x$ gives

$$x = 505 + (1.28)(110) = 645.8$$

This equation should make sense: it finds the $x$ that lies 1.28 standard deviations above the mean on this particular normal curve. That is the

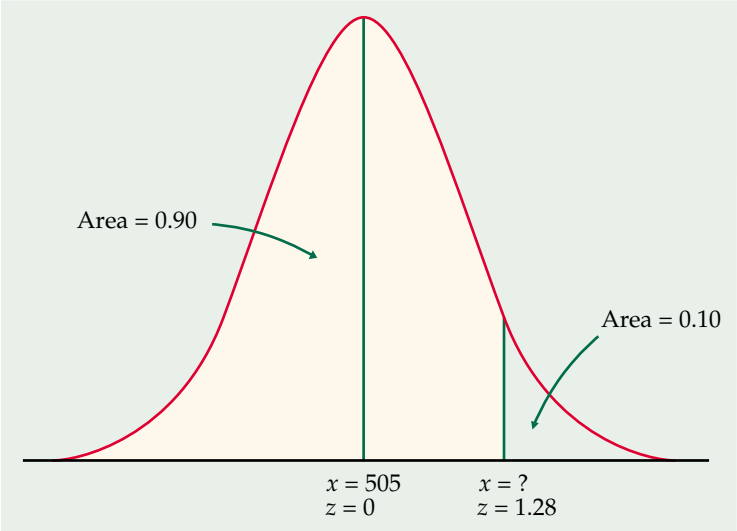



FIGURE 1.30 Locating the point on a normal curve with area 0.10 to its right.

"unstandardized" meaning of $z = 1.28$. We see that a student must score at least 646 to place in the highest 10%. The general rule for unstandardizing a $z$-score is

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

## Normal quantile plots

The normal distributions provide good models for some distributions of real data. Examples include the Gary vocabulary scores and Newcomb's measurements of the passage time of light (after dropping the outliers). The distributions of some other common variables are usually skewed and therefore distinctly nonnormal. Examples include economic variables such as personal income and gross sales of business firms, the survival times of cancer patients after treatment, 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 relative frequency calculations 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 (Newcomb's histogram, Figure 1.5 on page 17), pronounced skewness (the supermarket purchase stemplot, Figure 1.3 on page 12), or gaps and clusters (the histogram of Massachusetts college tuitions in Figure 1.12, page 27). If the stemplot or histogram appears

roughly symmetric and unimodal, 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**.*

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 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 nonnormal distribution. Outliers appear as points that are far away from the overall pattern of the plot.

Figures 1.31 to 1.34 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 scale extends 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 1.29

Figure 1.31 is a normal quantile plot of Newcomb's passage time data. Most of the points lie close to a straight line, indicating that a normal model fits well. The two outliers deviate from the line and show how the plot responds to low outliers. These two points are below the line formed by the rest of the data—they are farther out

*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.

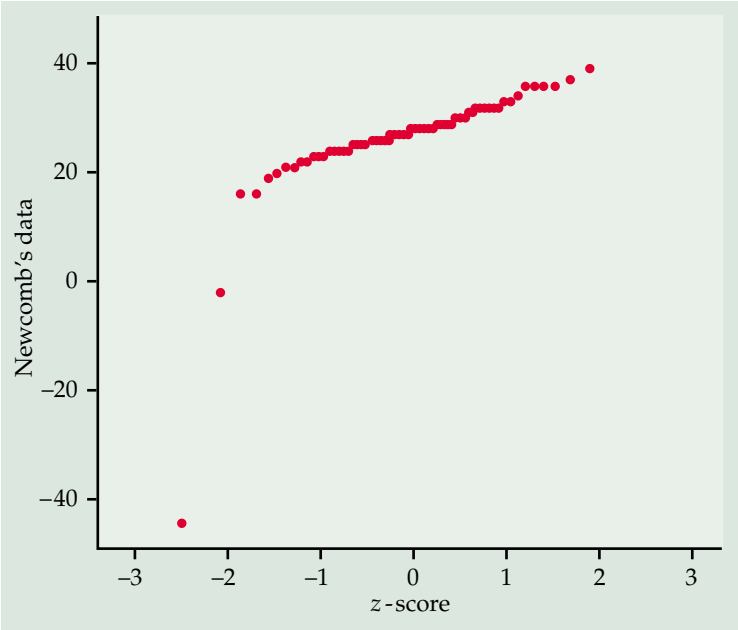



FIGURE 1.31 Normal quantile plot for Newcomb's 66 measurements of the passage time of light. There are two low outliers.

in the low direction than we expect normal data to be. Compare the histogram of these data (Figure 1.5 on page 17).

Figure 1.32 is a normal quantile plot of Newcomb's measurements of the passage time of light with the two outliers omitted. The effect of omitting the outliers is to magnify the plot of the remaining data. As we saw from Figure 1.31, a normal distribution fits quite well. The only important deviation from normality is the "stair-step" appearance caused by numerous short horizontal runs of points. Each run represents repeated observations having the same value—there are six measurements at 27 and seven at 28, for example. This phenomenon is called **granularity**. It is caused by the limited precision of the measurements. The granularity would disappear if Newcomb had been able to measure one more decimal place to spread out the "steps." This minor granularity does not prevent us from adopting a normal distribution as a model.

granularity

EXAMPLE 1.30

Figure 1.33 is a normal quantile plot of the supermarket spending data. The stemplots in Figure 1.3 (page 12) show that the distribution is right-skewed. To see the right skewness in the normal quantile plot, draw a line through the leftmost points, 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 1.31, there are no individual outliers.

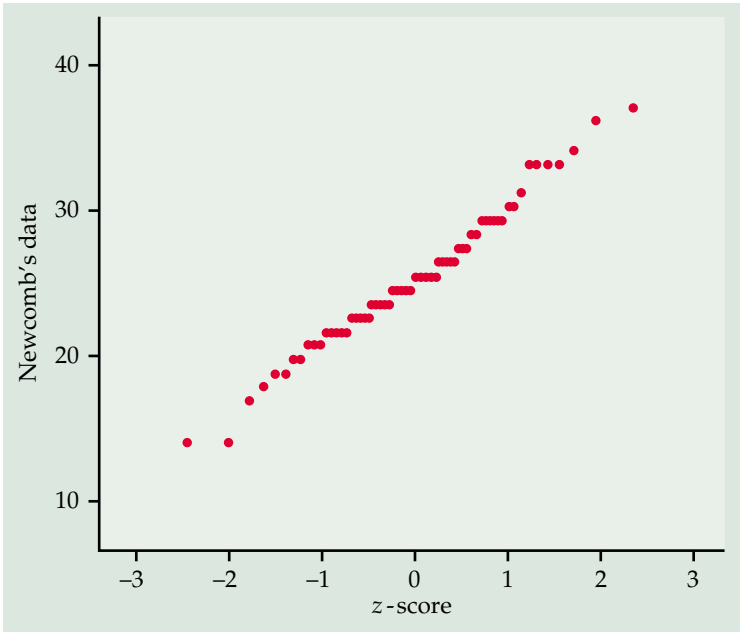



FIGURE 1.32 Newcomb's data without the two outliers. The data are close to normal except for slight granularity.

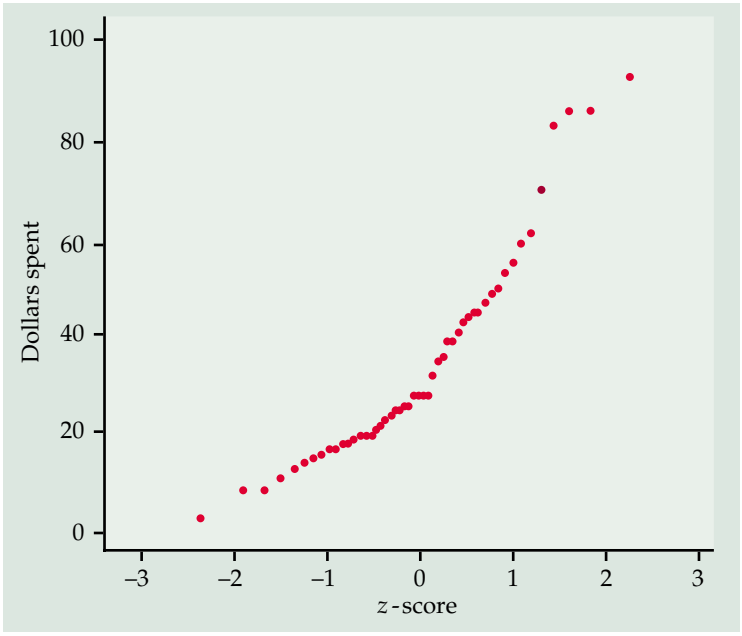



FIGURE 1.33 Normal quantile plot of the supermarket spending data. The pattern bends up at the right, showing right skewness.

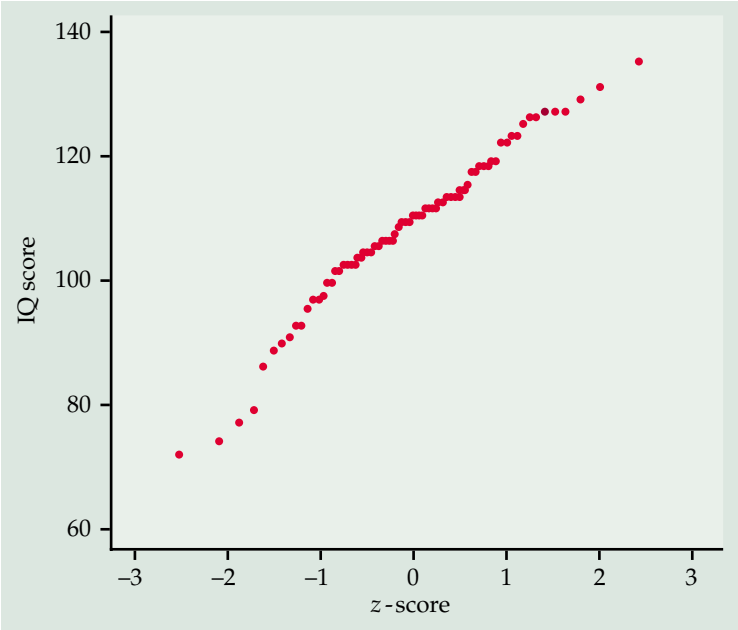



FIGURE 1.34 Normal quantile plot of the IQ scores of 78 seventh-grade students. The straight pattern shows that this distribution is close to normal.

EXAMPLE 1.31

Figure 1.34 is a normal quantile plot of the 78 IQ scores from Table 1.6. We expect IQ scores to be roughly normal. The plot is roughly straight, showing good fit to a normal model except for some granularity due to the fact that all IQ scores are whole numbers.

As Figures 1.32 and 1.34 illustrate, 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 approximately normal and outliers are not present.

## Beyond the basics ▸ density estimation

A density curve gives a compact summary of the overall shape of a distribution. Figure 1.20 (page 64) shows a normal density curve that provides a good summary of the distribution of Gary vocabulary scores. 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.

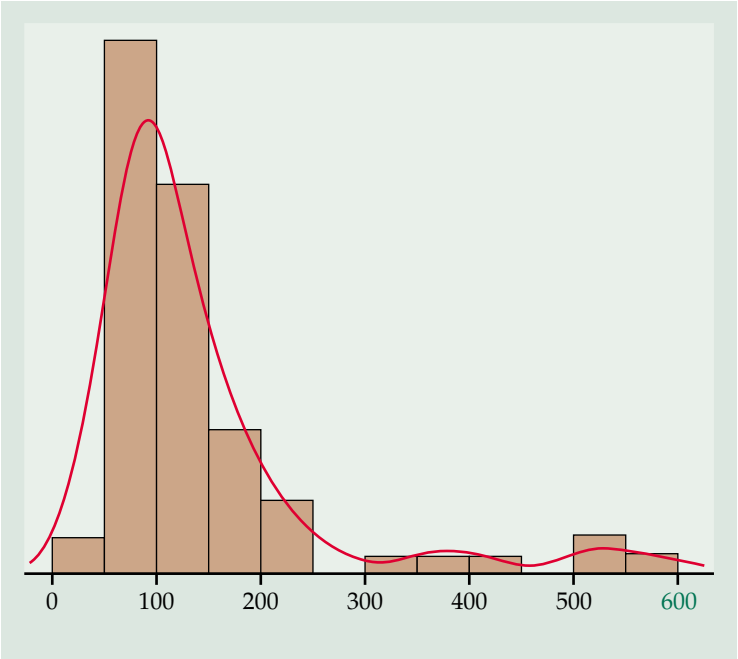



FIGURE 1.35 A density curve for the guinea pig survival data, calculated from the data by density estimation software.

It looks at the data and draws a density curve that describes the overall shape of the data. Figure 1.35 shows a histogram of the strongly right-skewed guinea pig survival data from Table 1.5. 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 100 days, and the two smaller clusters in the right tail. You can also see that the density estimator works blindly: in order to get a smooth curve, it starts below 0, even though survival times less than 0 days are not possible. Nonetheless, density estimation offers a quick way to picture the overall shapes of distributions. It is another useful tool for exploratory data analysis.[25]

## SUMMARY

The overall pattern of a distribution can often be described compactly by a **density curve**. A density curve has total area 1 underneath it. Areas under a density curve give **relative frequencies** for the distribution.

The **mean** $\mu$ (balance point), the **median** (equal-areas point), and the **quartiles** can be approximately located by eye on a density curve. The **standard deviation** $\sigma$ cannot be located by eye on most density curves. The mean and median are equal for symmetric density curves, but the mean of a skewed curve is located farther toward the long tail than is the median.

The **normal distributions** are described by bell-shaped, symmetric, unimodal density curves. The mean $\mu$ and standard deviation $\sigma$ completely specify the normal distribution $N(\mu, \sigma)$. The mean is the center of symmetry, and $\sigma$ is the distance from $\mu$ 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** $z = (x - \mu)/\sigma$ 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**.

If $X$ has the $N(\mu, \sigma)$ distribution, then the standardized variable $Z = (X - \mu)/\sigma$ has the **standard normal** distribution $N(0, 1)$. Relative frequencies for any normal distribution can be calculated from the **standard normal table** (Table A), which gives relative frequencies for the events $Z < z$ for many values of $z$.

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

**1.77** **(a)** Sketch a density curve that is symmetric but has a shape different from that of the normal curves.

**(b)** Sketch a density curve that is strongly skewed to the left.

uniform distribution

**1.78** If you ask a computer to generate "random numbers" between 0 and 1, you will get observations from a **uniform distribution**. Figure 1.36 graphs the density curve for a uniform distribution.

**(a)** Check that the area under this curve is 1.

**(b)** What proportion of the outcomes lie between 0.1 and 0.9? (Sketch the density curve, shade the area that represents the proportion, then find that area.)

**1.79** Many random number generators allow users to specify the range of the random numbers to be produced. Suppose that you specify that the


FIGURE 1.36 The density curve of a uniform distribution, for Exercise 1.78.

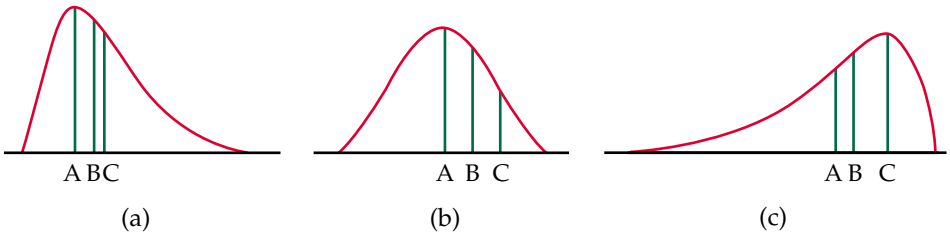



FIGURE 1.37 Three density curves, for Exercise 1.81.

outcomes are to be distributed uniformly 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 areas under the curve are relative frequencies of outcomes to find the proportion of outcomes that are less than 1.

**(c)** Find the proportion of outcomes that lie between 0.5 and 1.3.

**1.80** What are the mean and the median of the uniform distribution in Figure 1.36? What are the quartiles?

**1.81** Figure 1.37 displays three density curves, each with three points marked on it. At which of these points on each curve do the mean and the median fall?

**1.82** 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.

**(a)** Between what values do the lengths of the middle 95% of all pregnancies fall?

**(b)** How short are the shortest 2.5% of all pregnancies? How long do the longest 2.5% last?

**1.83** Bigger animals tend to carry their young longer before birth. The length of horse pregnancies from conception to birth varies according to a roughly normal distribution with mean 336 days and standard deviation 3 days. Use the 68–95–99.7 rule to answer the following questions.

**(a)** Almost all (99.7%) of horse pregnancies fall in what range of lengths?

**(b)** What percent of horse pregnancies are longer than 339 days?


**1.84** The 68–95–99.7 rule for normal distributions is a useful approximation. You can use the *Normal Curve* applet on the Web site for this book to see how accurate the rule is. Drag one flag across the other so that the applet shows the area under the curve between the two flags.

**(a)** Place the flags one standard deviation on either side of the mean. What is the area between these two values? What does the 68–95–99.7 rule say this area is?

**(b)** Repeat for locations two and three standard deviations on either side of the mean. Again compare the 68–95–99.7 rule with the area given by the applet.

**1.85** The normal quantile plot in Figure 1.34 shows that the IQ scores of the 78 seventh-grade students reported in Table 1.6 are approximately normal. How well do these scores satisfy the 68–95–99.7 rule? To find out, calculate the mean $\overline{x}$ and standard deviation $s$ of the scores. Then calculate the percent of the 78 scores that fall between $\overline{x} - s$ and $\overline{x} + s$ and compare your result with 68%. Do the same for the intervals covering two and three standard deviations on either side of the mean. (The 68–95–99.7 rule is exact for any theoretical normal distribution. It will hold only approximately for actual data.)

**1.86** Eleanor scores 680 on the mathematics part of the SAT examination. The distribution of SAT scores in a reference population is normal with mean 500 and standard deviation 100. Gerald takes the ACT mathematics test and scores 27. ACT scores are normally distributed with mean 18 and standard deviation 6. Find the $z$-scores for both students. Assuming that both tests measure the same kind of ability, who has the higher score?

**1.87** Three landmarks of baseball achievement are Ty Cobb's batting average of .420 in 1911, Ted Williams's .406 in 1941, and George Brett's .390 in 1980. These batting averages cannot be compared directly because the distribution of major league batting averages has changed over the decades. The distributions are quite symmetric and (except for outliers such as Cobb, Williams, and Brett) reasonably normal. While the mean batting average has been held roughly constant by rule changes and the balance between hitting and pitching, the standard deviation has dropped over time. Here are the facts:[26]

| Decade | Mean | Std. dev. |
|---|---|---|
| 1910s | .266 | .0371 |
| 1940s | .267 | .0326 |
| 1970s | .261 | .0317 |

Compute the standardized batting averages for Cobb, Williams, and Brett to compare how far each stood above his peers.

**1.88** Using either Table A or your calculator or software, 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.

**(a)** $Z < 2.85$

**(b)** $Z > 2.85$

**(c)** $Z > -1.66$

**(d)** $-1.66 < Z < 2.85$

**1.89** Using either Table A or your calculator or software, find the relative frequency of each of the following events in a standard normal distribution. In each case, sketch a standard normal curve with the area representing the relative frequency shaded.

**(a)** $Z \leq -2.25$

**(b)** $Z \geq -2.25$

**(c)** $Z > 1.77$

**(d)** $-2.25 < Z < 1.77$

**1.90** Find the value $z$ of a standard normal variable $Z$ that satisfies each of the following conditions. (If you use Table A, report the value of $z$ 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.

**1.91** The variable $Z$ has a standard normal distribution.

**(a)** Find the number $z$ such that the event $Z < z$ has relative frequency 0.8.

**(b)** Find the number $z$ such that the event $Z > z$ has relative frequency 0.35.

**1.92** In 2000, the scores of men on the math part of the SAT approximately followed a normal distribution with mean 533 and standard deviation 115.

**(a)** What proportion of men scored above 500?

**(b)** What proportion scored between 400 and 600?

**1.93** Too much cholesterol in the blood increases the risk of heart disease. Young women are generally less afflicted with high cholesterol than other groups. The cholesterol levels for women aged 20 to 34 follow an approximately normal distribution with mean 185 milligrams per deciliter (mg/dL) and standard deviation 39 mg/dL.[27]

**(a)** Cholesterol levels above 240 mg/dL demand medical attention. What percent of young women have levels above 240 mg/dL?

**(b)** Levels above 200 mg/dL are considered borderline high. What percent of young women have blood cholesterol between 200 and 240 mg/dL?

**1.94** It is possible to score higher than 800 on the SAT, but scores above 800 are reported as 800. (That is, a student can get a reported score of 800 without a perfect paper.) In 2000, the scores of men on the math part of the SAT approximately followed a normal distribution with mean 533 and standard deviation 115. What percent of scores were above 800 (and so reported as 800)?

**1.95** Middle-aged men are more susceptible to high cholesterol than the young women of Exercise 1.93. The blood cholesterol levels of men aged 55 to 64 are approximately normal with mean 222 mg/dL and standard deviation

37 mg/dL. What percent of these men have high cholesterol (levels above 240 mg/dL)? What percent have borderline high cholesterol (between 200 and 240 mg/dL)?

**1.96** Osteoporosis is a condition in which the bones become brittle due to loss of minerals. To diagnose osteoporosis, an elaborate apparatus measures bone mineral density (BMD). BMD is usually reported in standardized form. The standardization is based on a population of healthy young adults. The World Health Organization (WHO) criterion for osteoporosis is a BMD 2.5 standard deviations below the mean for young adults. BMD measurements in a population of people similar in age and sex roughly follow a normal distribution.

**(a)** What percent of healthy young adults have osteoporosis by the WHO criterion?

**(b)** Women aged 70 to 79 are of course not young adults. The mean BMD in this age is about $-2$ on the standard scale for young adults. Suppose that the standard deviation is the same as for young adults. What percent of this older population has osteoporosis?

**1.97** Changing the mean of a normal distribution by a moderate amount can greatly change the percent of observations in the tails. Suppose that a college is looking for applicants with SAT math scores 750 and above.

**(a)** In 2000, the scores of men on the math SAT followed a normal distribution with mean 533 and standard deviation 115. What percent of men scored 750 or better?

**(b)** Women's scores that year had a normal distribution with mean 498 and standard deviation 109. What percent of women scored 750 or better? You see that the percent of men above 750 is almost three times the percent of women with such high scores.

**1.98** 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.

**(a)** In what range do the middle 95% of all yearly returns lie?

**(b)** The market is down for the year if the return is less than zero. In what percent of years is the market down?

**(c)** In what percent of years does the index gain 25% or more?

**1.99** 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.

**(a)** What percent of pregnancies last less than 240 days (that's about 8 months)?

**(b)** What percent of pregnancies last between 240 and 270 days (roughly between 8 months and 9 months)?

**(c)** How long do the longest 20% of pregnancies last?

**1.100** 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.[28] 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. One 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?

**1.101** We have met the Survey of Study Habits and Attitudes (SSHA) as a common psychological instrument to evaluate the attitudes of students. The SSHA is used for subjects from seventh grade through college. Different groups have different distributions. To prepare to use the SSHA to evaluate future teachers, researchers gave the test to 238 college juniors majoring in elementary education. Their scores were roughly normal with mean 114 and standard deviation 30. Take this as the distribution of SSHA scores in the population of future elementary school teachers.

A study of Native American education students in Canada found that this relatively disadvantaged group had mean SSHA score 99.[29] What percentile of the overall distribution is this?

**1.102** Scores on the Wechsler Adult Intelligence Scale for the 20 to 34 age group are approximately normally distributed with mean 110 and standard deviation 25. Scores for the 60 to 64 age group are approximately normally distributed with mean 90 and standard deviation 25.

Sarah, who is 30, scores 135 on this test. Sarah's mother, who is 60, also takes the test and scores 120. Who scored higher relative to her age group, Sarah or her mother? Who has the higher absolute level of the variable measured by the test? At what percentile of their age groups are Sarah and her mother? (That is, what percent of the age group has lower scores?)

**1.103** How high a score on the SSHA test of Exercise 1.101 must an elementary education student obtain to be among the highest-scoring 30% of the population? What scores make up the lowest 30%?

**1.104** The scores of a reference population on the Wechsler Intelligence Scale for Children (WISC) are normally distributed with $\mu = 100$ and $\sigma = 15$.

**(a)** What percent of this population have WISC scores below 100?

**(b)** Below 80?

**(c)** Above 140?

**(d)** Between 100 and 120?

**1.105** The distribution of scores on the WISC is described in the previous exercise. What score will place a child in the top 5% of the population? In the top 1%?

**1.106** The median of any normal distribution is the same as its mean. We can use normal calculations to find the quartiles and related descriptive measures for normal distributions.

**(a)** 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.

**(b)** Your work in (a) gives the $z$-scores for the quartiles of any normal distribution. Scores on the Wechsler Intelligence Scale for Children (WISC) are normally distributed with mean 100 and standard deviation 15. What are the quartiles of WISC scores?

**(c)** What is the value of the *IQR* for the standard normal distribution?

**(d)** What percent of the observations in the standard normal distribution are suspected outliers according to the $1.5 \times IQR$ criterion? (This percent is the same for any normal distribution.)

**1.107** The lower and upper 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 area 0.1 and 0.9 to their left under the curve.

**(a)** What are the lower and upper deciles of the standard normal distribution?

**(b)** The length of human pregnancies is approximately normal with mean 266 days and standard deviation 16 days. What are the lower and upper deciles of this distribution?

**1.108** Figure 1.38 is a normal quantile plot of the DRP scores from Exercise 1.26 (page 30). Are these scores approximately normally distributed? Discuss any major deviations from normality that appear in the plot.

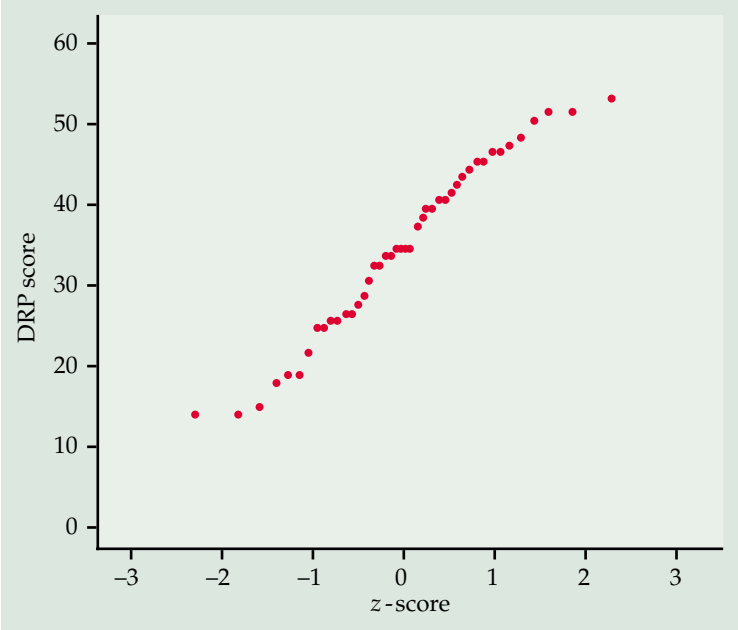



FIGURE 1.38 Normal quantile plot of DRP scores, for Exercise 1.108.

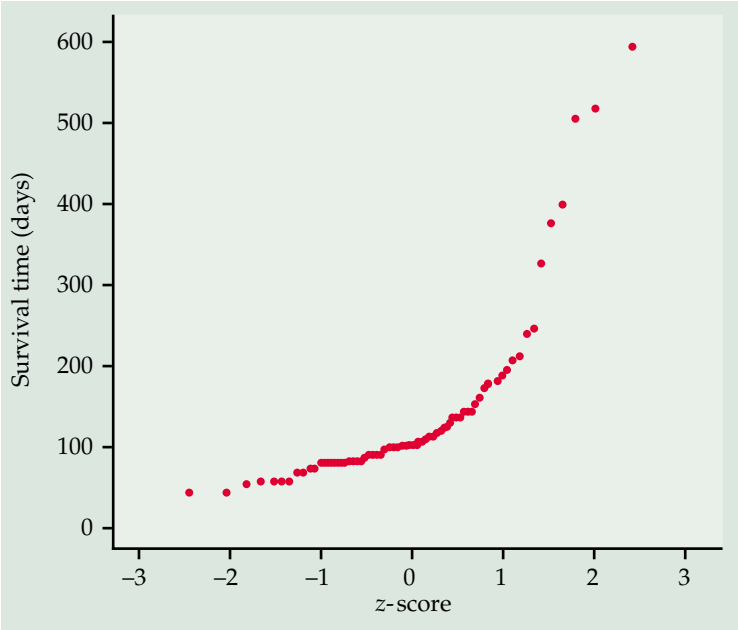



FIGURE 1.39 Normal quantile plot of guinea pig survival times, for Exercise 1.109.

**1.109** Figure 1.39 is a normal quantile plot of the survival times of the guinea pigs in a medical experiment, from Table 1.5. Explain carefully how the strong right skewness of this distribution is seen in the plot.

**1.110** The distance between two mounting holes is important to the performance of an electrical meter. The manufacturer measures this distance regularly for quality control purposes, recording the data as thousandths of an inch more than 0.600 inches. For example, 0.644 is recorded as 44. Figure 1.40 is a normal quantile plot of the distances for the last 90 electrical meters measured.[30] Is the overall shape of the distribution approximately normal? Why does the plot have a "stair-step" appearance?

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

**1.111** We expect repeated careful measurements of the same quantity to be approximately normal. Make a normal quantile plot for Cavendish's measurements in Exercise 1.27 (page 30). Are the data approximately normal? If not, describe any clear deviations from normality.

**1.112** The distribution of Internet access costs in Exercise 1.45 (page 56) has a compact center with a long tail on either side. Make a normal quantile plot of these data. Explain carefully why the pattern of this plot is typical of a "long-tailed" distribution.

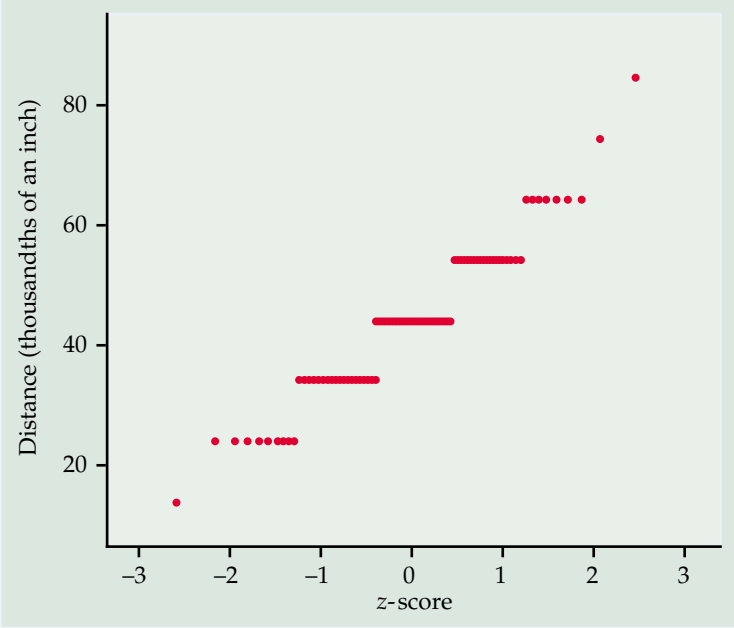



FIGURE 1.40 Normal quantile plot of distance between mounting holes, for Exercise 1.110.

**1.113** Table 1.6 (page 33) reports the scores of 78 seventh-grade students on the Piers-Harris Children's Self-Concept Scale. Give a careful description of the distribution of self-concept scores using both graphs and numbers. Are there outliers? Is the distribution roughly normal if we ignore the outliers?

**1.114** Table 1.6 (page 33) gives the genders and Piers-Harris self-concept scores for 78 seventh-grade students. Gender is coded as F for females and M for males. Compare the distributions of scores for female and male students, using both graphs and numbers. Report your findings. Some people think that female students in mixed classes have lower self-concept than male students. Do you see any evidence that this is true for these students?

**1.115** 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.)

**1.116** Use software to generate 100 observations from the uniform distribution described in Exercise 1.78. Make a histogram of these observations. How does the histogram compare with the density curve in Figure 1.36? Make a normal quantile plot of your data. According to this plot, how does the uniform distribution deviate from normality?

## CHAPTER 1 EXERCISES

**1.117** 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?

**1.118** What type of graph or graphs would you plan to make in a study of each of the following issues?

**(a)** What makes of cars do students drive? How old are their cars?

**(b)** How many hours per week do students study? How does the number of study hours change during a semester?

**(c)** Which radio stations are most popular with students?

**(d)** When many students measure the concentration of the same solution for a chemistry course laboratory assignment, do their measurements follow a normal distribution?

**1.119** Are the rich different? Figure 1.41 shows histograms of the number of people in a household for households in the top 5% of incomes (left) and for the other 95% of households (right).[31]

**(a)** These are relative frequency histograms, in which the vertical scales show the proportion of observations in a class rather than the count of observations. Why are these preferable to frequency histograms to display these data?

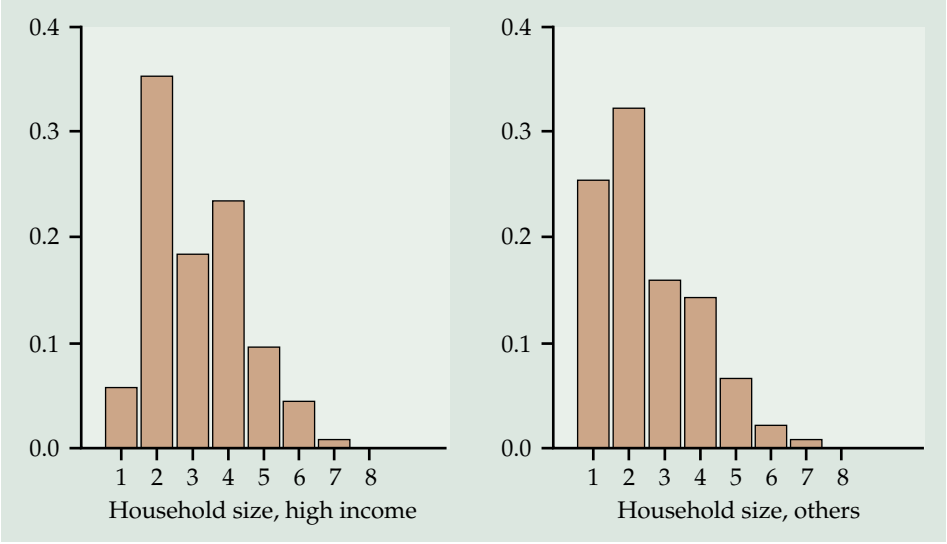



FIGURE 1.41 The distributions of household size for households in the top 5% of income (left) and for other households (right), for Exercise 1.119.

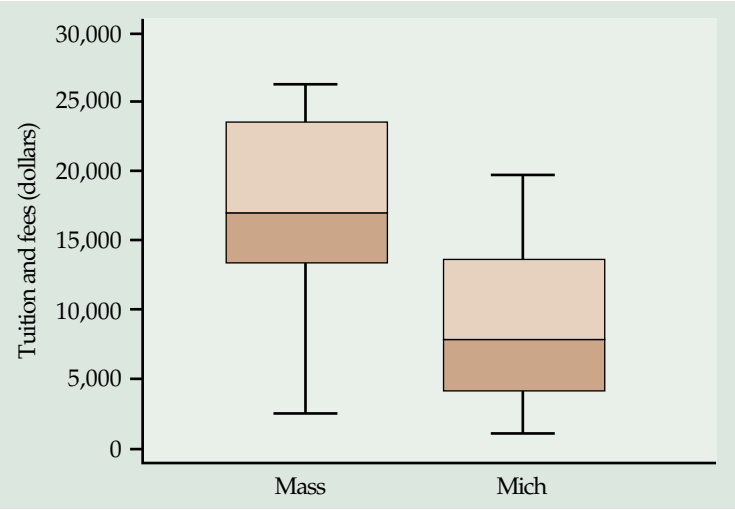



FIGURE 1.42 Boxplots comparing the tuition and fees charged by four-year colleges in Massachusetts and Michigan, for Exercise 1.120.

**(b)** What are the most important ways in which the sizes of high-income households differ from other households?

**1.120** Figure 1.42 displays side-by-side boxplots comparing the 2000–2001 tuition and fees charged by all four-year colleges in Massachusetts and in Michigan. For state schools, we used the in-state tuition.

**(a)** Use the boxplots to write a brief comparison of the two distributions.

**(b)** Eastern states had many private colleges before state-supported higher education became common. Public higher education is more prevalent in the Midwest, which was settled later. How does this historical pattern help explain the differences between the Massachusetts and Michigan distributions of college charges?

**1.121** If two distributions have exactly the same mean and standard deviation, must their histograms have the same shape? If they have the same five-number summary, must their histograms have the same shape? Explain.

**1.122** By-products from the pesticide DDT were major threats to the survival of birds of prey until use of DDT was banned at the end of 1972. Can time plots show the effect of the ban? Here are two sets of data for bald eagles nesting in the forests of northwestern Ontario.[32] The first data set gives the mean number of young per breeding area:

| Year | 1966 | 1967 | 1968 | 1969 | 1970 | 1971 | 1972 | 1973 |
|---|---|---|---|---|---|---|---|---|
| Young | 1.26 | 0.73 | 0.89 | 0.84 | 0.54 | 0.60 | 0.54 | 0.78 |
| Year | 1974 | 1975 | 1976 | 1977 | 1978 | 1979 | 1980 | 1981 |
| Young | 0.46 | 0.77 | 0.86 | 0.96 | 0.82 | 0.98 | 0.93 | 1.12 |

The second set of data are measurements of the chemical DDE, the by-product of DDT that most threatens birds of prey, from bald eagle eggs in the same area of Canada. These are in parts per million (ppm). There are often several measurements per year.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | 1967 | 1967 | 1968 | 1971 | 1971 | 1972 | 1976 | 1976 | 1976 | 1976 |
| DDE | 44 | 95 | 121 | 125 | 95 | 87 | 13.3 | 16.4 | 50.4 | 59.8 |
| Year | 1976 | 1977 | 1977 | 1980 | 1980 | 1980 | 1981 | 1981 | 1981 | |
| DDE | 56.4 | 0.6 | 23.8 | 16.6 | 14.5 | 24.0 | 7.8 | 48.2 | 53.4 | |

Make time plots of eagle young and of mean DDE concentration in eggs. How does the effect of banning DDT appear in your plots?

**1.123** Our direct comparison of Mark McGwire and Babe Ruth (page 11) ignores the historical context. Here are the number of home runs by the major league leader for each year in baseball history, 1876 to 2001, in order from left to right. Make a time plot. (Be sure to add the scale of years.)

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 3 | 4 | 9 | 6 | 7 | 7 | 10 | 27 | 11 | 11 | 17 | 14 | 20 | 14 | 16 | 13 | 19 |
| 18 | 17 | 13 | 11 | 15 | 25 | 12 | 16 | 16 | 13 | 10 | 9 | 12 | 10 | 12 | 9 | 10 | 21 |
| 14 | 19 | 19 | 24 | 12 | 12 | **11** | **29** | **54** | **59** | 42 | **41** | **46** | 39 | **47** | **60** | **54** | **46** |
| 56 | **46** | 58 | 48 | 49 | 36 | 49 | 46 | 58 | 35 | 43 | 37 | 36 | 34 | 33 | 28 | 44 | 51 |
| 40 | 54 | 47 | 42 | 37 | 47 | 49 | 51 | 52 | 44 | 47 | 46 | 41 | 61 | 49 | 45 | 49 | 52 |
| 49 | 44 | 44 | 49 | 45 | 48 | 40 | 44 | 36 | 38 | 38 | 52 | 46 | 48 | 48 | 31 | 39 | 40 |
| 43 | 40 | 40 | **49** | 42 | 47 | 51 | 44 | 43 | 46 | 43 | 50 | **52** | **58** | **70** | **65** | 50 | 73 |

**(a)** Describe the effect of World War II (1942 to 1945 seasons).

**(b)** Ruth led in the 11 years in boldface between 1918 and 1931. McGwire led in the 5 boldface years between 1987 and 1999. Briefly compare the achievements of Ruth and McGwire in the context of their times.

**1.124** The single-season home run record is now held by Barry Bonds of the San Francisco Giants, who hit 73 in 2001. Here are Bonds's home run totals from 1986 (his first year) to 2001:

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 25 | 24 | 19 | 33 | 25 | 34 | 46 | 37 | 33 | 42 | 40 | 37 | 34 | 49 | 73 |

Bonds's record year is a high outlier. How do his career mean and median number of home runs change when we drop the record 73? What general fact about the mean and median does your result illustrate?

**1.125** The Bureau of Justice Statistics says that in 1999, 51% of homicides were committed with handguns, 14% with other firearms, 13% with knives, and 6% with blunt objects. Make a graph to display these data. Do you need an "other methods" category?

**1.126** The Internal Revenue Service reports that in 1998 about 124 million individual income tax returns showed adjusted gross income (AGI) greater than 0. The mean and median AGI on these tax returns were $25,491 and $44,186. Which of these numbers is the mean and which is the median? How do you know?

**1.127** Raw scores on behavioral tests are often transformed for easier comparison. A test of reading ability has mean 75 and standard deviation 10 when given to third graders. Sixth graders have mean score 82 and standard deviation 11 on the same test. To provide separate "norms" for each grade, we want scores in each grade to have mean 100 and standard deviation 20.

**(a)** What linear transformation will change third-grade scores $x$ into new scores $x_{\text{new}} = a + bx$ that have the desired mean and standard deviation? (Use $b > 0$ to preserve the order of the scores.)

**(b)** Do the same for the sixth-grade scores.

**(c)** David is a third-grade student who scores 78 on the test. Find David's transformed score. Nancy is a sixth-grade student who scores 78. What is her transformed score? Who scores higher within his or her grade?

**(d)** Suppose that the distribution of scores in each grade is normal. Then both sets of transformed scores have the $N(100, 20)$ distribution. What percent of third graders have scores less than 78? What percent of sixth graders have scores less than 78?

**1.128** The Chapin Social Insight Test evaluates how accurately the subject appraises other people. In the reference population used to develop the test, scores are approximately normally distributed with mean 25 and standard deviation 5. The range of possible scores is 0 to 41.

**(a)** What proportion of the population has scores below 20 on the Chapin test?

**(b)** What proportion has scores below 10?

**(c)** How high a score must you have in order to be in the top quarter of the population in social insight?

**1.129** Exercise 1.25 (page 29) presents data on the nightly study time claimed by first-year college men and women. We noted in Exercise 1.58 that each group contains an outlier. Make a normal quantile plot for each group, omitting the outlier. How does the fact that most students estimated study time in multiples of 10 minutes affect the appearance of the plots? Are the distributions roughly normal in other respects?

**1.130** The Chapin Social Insight Test described in Exercise 1.128 has a mean of 25 and a standard deviation of 5. You want to rescale the test using a linear

TABLE 1.10 Position and weight (pounds) for a major college football team

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| QB | 235 | QB | 220 | QB | 215 | QB | 228 | K | 175 | K | 205 |
| K | 220 | K | 185 | RB | 200 | RB | 188 | RB | 190 | RB | 215 |
| RB | 190 | RB | 225 | RB | 225 | RB | 240 | RB | 237 | R | 205 |
| R | 185 | R | 185 | R | 190 | R | 195 | R | 201 | R | 190 |
| R | 195 | R | 180 | OL | 291 | OL | 280 | OL | 300 | OL | 320 |
| OL | 325 | OL | 285 | OL | 305 | OL | 305 | OL | 290 | OL | 310 |
| OL | 315 | OL | 285 | OL | 290 | OL | 325 | OL | 310 | OL | 256 |
| OL | 305 | OL | 300 | DB | 170 | DB | 207 | DB | 185 | DB | 175 |
| DB | 180 | DB | 190 | DB | 210 | DB | 200 | DB | 180 | DB | 195 |
| DB | 185 | DB | 170 | DB | 180 | DB | 190 | DB | 190 | LB | 220 |
| LB | 212 | LB | 233 | LB | 190 | LB | 195 | LB | 215 | LB | 220 |
| LB | 240 | LB | 230 | LB | 220 | LB | 225 | LB | 200 | DL | 260 |
| DL | 245 | DL | 264 | DL | 254 | DL | 215 | DL | 250 | DL | 295 |
| DL | 240 | DL | 275 | DL | 255 | DL | 285 | DL | 245 | DL | 270 |
| DL | 250 | DL | 250 | TE | 255 | TE | 245 | TE | 260 | TE | 245 |

transformation so that the mean is 100 and the standard deviation is 20. Let $x$ denote the score in the original scale and $x_{\text{new}}$ be the transformed score.

**(a)** Find the linear transformation required. That is, find the values of $a$ and $b$ in the equation $x_{\text{new}} = a + bx$.

**(b)** Give the rescaled score for someone who scores 30 in the original scale.

**(c)** What are the quartiles of the rescaled scores?

**1.131** The Florida State University Seminoles have been among the more successful teams in college football. Table 1.10 gives the weights in pounds and positions of the players on the 2000–2001 football team, which was defeated in the national title game by the University of Oklahoma.[33] The positions are quarterback (QB), running back (RB), offensive line (OL), receiver (R), tight end (TE), kicker (K), defensive back (DB), linebacker (LB), and defensive line (DL).

**(a)** Make side-by-side modified boxplots of the weights for running backs, receivers, offensive linemen, defensive linemen, linebackers, and defensive backs.

**(b)** Briefly compare the weight distributions. Which position has the heaviest players overall? Which has the lightest?

**(c)** Are any individual players outliers within their position?

**1.132** 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 $\mu$ and $\sigma$ 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.)

**1.133** Table 1.11 shows the salaries paid to the members of the New York Yankees baseball team as of opening day of the 2001 season. Display this distribution with a graph and describe its main features. Find the mean and median salary and explain how the pattern of the distribution explains the relationship between these two measures of center. Find the standard deviation and the quartiles. Do you prefer the five-number summary or $\bar{x}$ and $s$ as a quick description of this distribution?

**1.134** The American Housing Survey provides data on all housing units in the United States—houses, apartments, mobile homes, and so on. Here are the years in which a random sample of 100 housing units were built. The survey does not produce exact dates for years before 1990. Years before 1920 are

TABLE 1.11 2001 salaries for the New York Yankees baseball team

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

given as 1919. Dates between 1920 and 1970 are given in ten-year blocks, so that a unit built in 1956 appears as 1950. Dates between 1970 and 1990 are given in five-year blocks, so that 1987 appears as 1985.[34]

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1960 | 1920 | 1991 | 1919 | 1985 | 1985 | 1975 | 1980 | 1975 | 1985 |
| 1930 | 1993 | 1985 | 1975 | 1970 | 1970 | 1975 | 1980 | 1940 | 1940 |
| 1980 | 1919 | 1980 | 1950 | 1940 | 1950 | 1993 | 1985 | 1975 | 1960 |
| 1919 | 1950 | 1960 | 1975 | 1950 | 1919 | 1920 | 1985 | 1970 | 1975 |
| 1930 | 1975 | 1960 | 1920 | 1940 | 1950 | 1985 | 1990 | 1950 | 1970 |
| 1985 | 1920 | 1950 | 1980 | 1975 | 1950 | 1950 | 1919 | 1919 | 1985 |
| 1985 | 1991 | 1980 | 1960 | 1940 | 1960 | 1930 | 1998 | 1994 | 1960 |
| 1919 | 1975 | 1919 | 1950 | 1975 | 1930 | 1919 | 1970 | 1920 | 1930 |
| 1950 | 1975 | 1970 | 1985 | 1919 | 1960 | 1930 | 1980 | 1960 | 1950 |
| 1996 | 1940 | 1950 | 1998 | 1930 | 1919 | 1930 | 1950 | 1950 | 1920 |

**(a)** Make a histogram of these dates, using classes 10 years wide beginning with 1910 to 1919. The first class will contain all housing units built before 1920. In which decades after 1920 were most housing units that still exist built?

**(b)** Give the five-number summary of these data. Write a brief warning on how to interpret your results. For example, what does the fact that the median is 1960 tell us about the age of American housing?

**1.135** At the time the salaries in Table 1.11 were announced, one dollar was worth 1.72 Swiss francs. Answer these questions without doing any calculations in addition to those you did in Exercise 1.133.

**(a)** What transformation converts a salary in dollars into the same salary in Swiss francs?

**(b)** What are the mean, median, and quartiles of the distribution in francs?

**(c)** What are the standard deviation and interquartile range of the distribution in francs?

**1.136** 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 1.12 gives the populations of California counties from the 2000 census. 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

TABLE 1.12 Population of California counties, 2000 census

| 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 | | |

first group, half of the second, and a smaller fraction of the third in your survey?

**1.137** Each March, the Bureau of Labor Statistics collects detailed information about more than 50,000 randomly selected households. The INDIVIDUALS data set contains data on 55,899 people from the March 2000 survey. The Data Appendix describes this data set in detail. 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.

**1.138** Continue your study of the INDIVIDUALS data set. 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.

**1.139** The CHEESE data set described in the Data Appendix records measurements on 30 specimens of Australian cheddar cheese. Investigate the distributions of the variables H2S and LACTIC using graphical and numerical summaries of your choice. Write a short description of the notable features for each distribution.

**1.140** The CSDATA data set described in the Data Appendix contains information on 234 computer science students. We are interested in comparing the SAT mathematics scores and grade point averages of female students with those of male students. Make two sets of side-by-side boxplots to carry out these comparisons. Write a brief discussion of the male-female comparisons. Then make normal quantile plots of grade point averages and SAT math scores separately for men and women. Which of the four distributions are approximately normal?