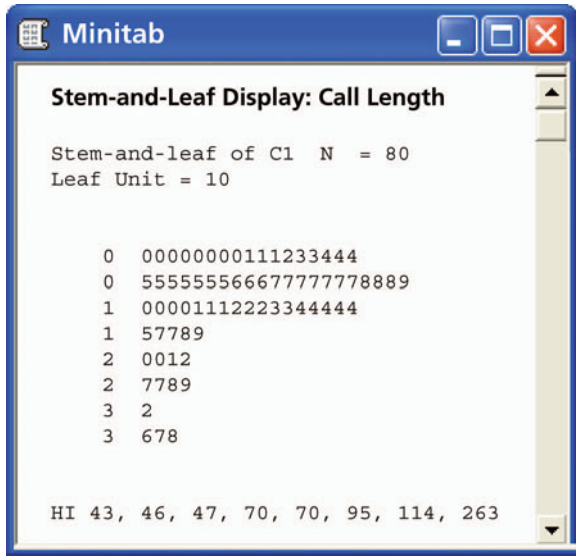



**FIGURE 1.6** Stemplot from Minitab of the 80 call lengths in Table 1.1, for Example 1.8. The software has trimmed the data by removing the last digit. It has also split stems and listed the highest observations apart from the plot.

0 to 4 go on the first 0 stem. Figure 1.6 is a stemplot of these data made by software. The software automatically did what we suggest: trimmed to tens of seconds and split stems. To save space, the software also listed the largest values as "HI" rather than create stems all the way up to 26. The stemplot shows the overall pattern of the distribution, with many short to moderate lengths and some very long calls.

## 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.

histogram

A **histogram** breaks the range of values of a variable into classes and displays only the count or percent of the observations that fall into each class. You can choose any convenient number of classes, but you should always choose classes 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.

EXAMPLE

**1.9 Distribution of IQ scores.** You have probably heard that the distribution of scores on IQ tests is supposed to be roughly "bell-shaped." Let's look at some actual IQ scores. Table 1.3 displays the IQ scores of 60 fifth-grade students chosen at random from one school.[6]

1. Divide the range of the data into classes of equal width. The scores in Table 1.3 range from 81 to 145, so we choose as our classes

**TABLE 1.3**

IQ test scores for 60 randomly chosen fifth-grade students

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 145 | 139 | 126 | 122 | 125 | 130 | 96 | 110 | 118 | 118 |
| 101 | 142 | 134 | 124 | 112 | 109 | 134 | 113 | 81 | 113 |
| 123 | 94 | 100 | 136 | 109 | 131 | 117 | 110 | 127 | 124 |
| 106 | 124 | 115 | 133 | 116 | 102 | 127 | 117 | 109 | 137 |
| 117 | 90 | 103 | 114 | 139 | 101 | 122 | 105 | 97 | 89 |
| 102 | 108 | 110 | 128 | 114 | 112 | 114 | 102 | 82 | 101 |

$$75 \le \text{IQ score} < 85$$

$$85 \le \text{IQ score} < 95$$

$$\vdots$$

$$145 \le \text{IQ score} < 155$$

Be sure to specify the classes precisely so that each individual falls into exactly one class. A student with IQ 84 would fall into the first class, but IQ 85 falls into the second.

frequency
frequency table

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 | Class | Count |
|---|---|---|---|
| 75 to 84 | 2 | 115 to 124 | 13 |
| 85 to 94 | 3 | 125 to 134 | 10 |
| 95 to 104 | 10 | 135 to 144 | 5 |
| 105 to 114 | 16 | 145 to 154 | 1 |

3. Draw the histogram. First, on the horizontal axis mark the scale for the variable whose distribution you are displaying. That's IQ score. The scale runs from 75 to 155 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 1.7 is our histogram. It does look roughly "bell-shaped."

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, we may be interested in the fraction or percent of the observations that fall in each class. A histogram of percents looks just like a frequency histogram such as Figure 1.7. Simply relabel the vertical scale to read in percents. Use histograms of percents for comparing several distributions that have different numbers of observations.

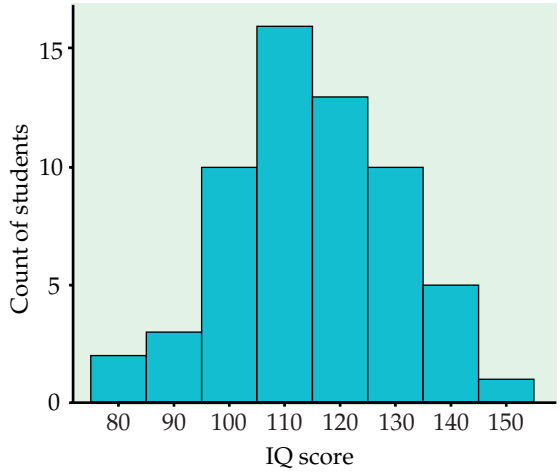



**FIGURE 1.7** Histogram of the IQ scores of 60 fifth-grade students, for Example 1.9.

## USE YOUR KNOWLEDGE

**1.6** **Make a histogram.** Refer to the first-exam scores from Exercise 1.5. Use these data to make a histogram using classes 50–59, 60–69, etc. Compare the histogram with the stemplot as a way of describing this distribution. Which do you prefer for these data?

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 often a good one, but you can change it if you want.


*You should be aware that the appearance of a histogram can change when you change the classes.* Figure 1.8 is a histogram of the customer service call lengths

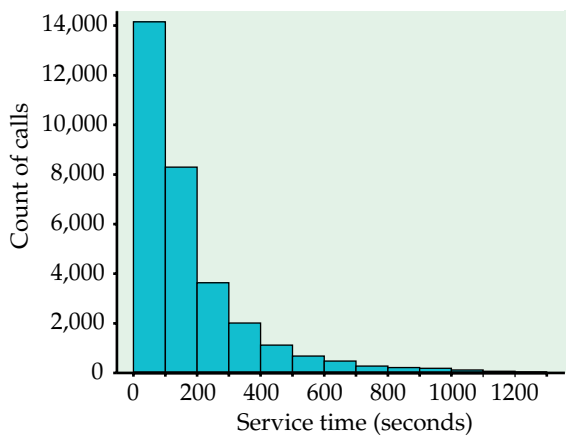



**FIGURE 1.8** The "default" histogram produced by software for the call lengths in Example 1.6. This choice of classes hides the large number of very short calls that is revealed by the histogram of the same data in Figure 1.4.

that are also displayed in Figure 1.4. It was produced by software with no special instructions from the user. The software's "default" histogram shows the overall shape of the distribution, but it hides the spike of very short calls by lumping all calls of less than 100 seconds into the first class. We produced Figure 1.4 by asking for smaller classes after Table 1.1 suggested that very short calls might be a problem. Software automates making graphs, but it can't replace thinking about your data. The histogram function in the *One-Variable Statistical Calculator* applet on the text CD and Web site allows you to change the number of classes by dragging with the mouse, so that it is easy to see how the choice of classes affects the histogram.


## USE YOUR KNOWLEDGE

**1.7** **Change the classes in the histogram.** Refer to the first-exam scores from Exercise 1.5 and the histogram you produced in Exercise 1.6. Now make a histogram for these data using classes 40–59, 60–79, and 80–100. Compare this histogram with the one that you produced in Exercise 1.6.

**1.8** **Use smaller classes.** Repeat the previous exercise using classes 55–59, 60–64, 65–69, etc.

Although histograms resemble bar graphs, their details and uses are distinct. A histogram shows the distribution of counts or percents 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 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 spreadsheet programs, which are not primarily intended for statistics, will draw histograms as if they were bar graphs, with space between the bars. Often, you can tell the software to eliminate the space to produce a proper histogram.*


## 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.

In Section 1.2, we will learn how to describe center and spread numerically. 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 *smallest and largest values.* Stemplots and histograms display the shape of a distribution in the same way. Just imagine a stemplot turned 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).

Some variables commonly have distributions with predictable shapes. Many biological measurements on specimens from the same species and sex—lengths of bird bills, heights of young women—have symmetric distributions. Money amounts, on the other hand, usually have right-skewed distributions. There are many moderately priced houses, for example, but the few very expensive mansions give the distribution of house prices a strong right-skew.

EXAMPLE

**1.10 Examine the histogram.** What does the histogram of IQ scores (Figure 1.7) tell us? **Shape:** The distribution is *roughly symmetric* with a *single peak* in the center. We don't expect real data to be perfectly symmetric, so we are satisfied if the two sides of the histogram are roughly similar in shape and extent. **Center:** You can see from the histogram that the midpoint is not far from 110. Looking at the actual data shows that the midpoint is 114. **Spread:** The spread is from 81 to 145. There are no outliers or other strong deviations from the symmetric, unimodal pattern.

The distribution of call lengths in Figure 1.8, on the other hand, is strongly *skewed to the right.* The midpoint, the length of a typical call, is about 115 seconds, or just under 2 minutes. The spread is very large, from 1 second to 28,739 seconds.

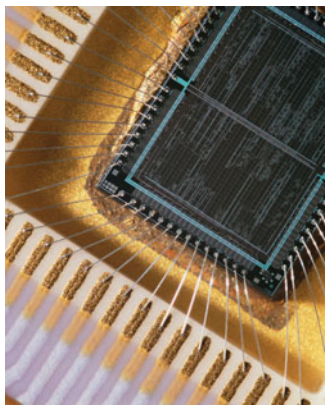
The longest few calls are *outliers.* They stand apart from the long right tail of the distribution, though we can't see this from Figure 1.8, which omits the largest observations. The longest call lasted almost 8 hours—that may well be due to equipment failure rather than an actual customer call.

## USE YOUR KNOWLEDGE

**1.9 Describe the first-exam scores.** Refer to the first-exam scores from Exercise 1.5. Use your favorite graphical display to describe the shape, the center, and the spread of these data. Are there any outliers?

## Dealing with outliers


In data sets smaller than the service call data, you can spot outliers by looking for observations that stand apart (either high or low) from the overall pattern of a histogram or stemplot. *Identifying outliers is a matter for judgment. Look for points that are clearly apart from the body of the data, not just the most extreme observations in a distribution.* 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.



EXAMPLE

**1.11 Semiconductor wires.** Manufacturing an electronic component requires attaching very fine wires to a semiconductor wafer. If the strength of the bond is weak, the component may fail. Here are measurements on the breaking strength (in pounds) of 23 connections:[7]

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 550 | 750 | 950 | 950 | 1150 | 1150 |
| 1150 | 1150 | 1150 | 1250 | 1250 | 1350 | 1450 | 1450 |
| 1450 | 1550 | 1550 | 1550 | 1850 | 2050 | 3150 | |

Figure 1.9 is a histogram of these data. We expect the breaking strengths of supposedly identical connections to have a roughly symmetric overall pattern, showing chance variation among the connections. Figure 1.9 does show a symmetric pattern centered at about 1250 pounds—but it also shows three *outliers* that stand apart from this pattern, two low and one high.

The engineers were able to explain all three outliers. The two low outliers had strength 0 because the bonds between the wire and the wafer were not made. The high outlier at 3150 pounds was a measurement error. Further study of the data can simply omit the three outliers. One immediate finding is that the variation in breaking strength is too large—550 pounds to 2050 pounds when we ignore the outliers. The process of bonding wire to wafer must be improved to give more consistent results.

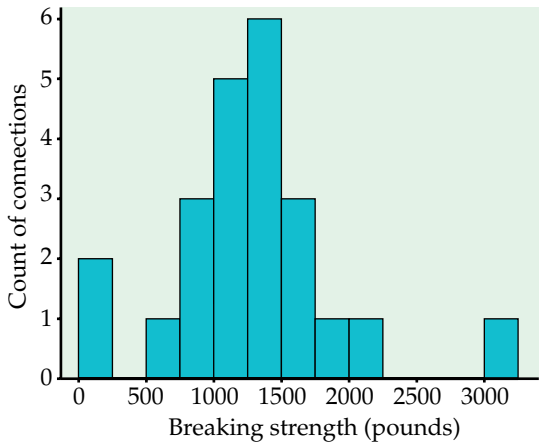



**FIGURE 1.9** Histogram of a distribution with both low and high outliers, for Example 1.11.