

Displaying and Summarizing Quantitative Data



Tsunamis are potentially destructive waves that can occur when the sea floor is suddenly and abruptly deformed. They are most often caused by earthquakes beneath the sea that shift the earth's crust, displacing a large mass of water.

The tsunami of December 26, 2004, with epicenter off the west coast of Sumatra, was caused by an earthquake of magnitude 9.0 on the Richter scale. It killed an estimated 297,248 people, making it the most disastrous tsunami on record. But was the earthquake that caused it truly extraordinary, or did it just happen at an unlucky place and time? The U.S. National Geophysical Data Center¹ has information on more than 2400 tsunamis dating back to 2000 B.C.E., and we have estimates of the magnitude of the underlying earthquake for 1240 of them. What can we learn from these data?

Histograms

WHO 1240 earthquakes known to have caused tsunamis for which we have data or good estimates

WHAT Magnitude (Richter scale ²), depth (m), date, location, and other variables

WHEN From 2000 B.C.E. to the present

WHERE All over the earth

Let's start with a picture. For categorical variables, it is easy to draw the distribution because each category is a natural "pile." But for quantitative variables, there's no obvious way to choose piles. So, usually, we slice up all the possible values into equal-width bins. We then count the number of cases that fall into each bin. The bins, together with these counts, give the **distribution** of the quantitative variable and provide the building blocks for the histogram. By representing the counts as bars and plotting them against the bin values, the **histogram** displays the distribution at a glance.

¹ www.ngdc.noaa.gov

² Technically, Richter scale values are in units of log dyne-cm. But the Richter scale is so common now that usually the units are assumed. The U.S. Geological Survey gives the background details of Richter scale measurements on its Web site www.usgs.gov/.

For example, here are the *Magnitudes* (on the Richter scale) of the 1240 earthquakes in the NGDC data:

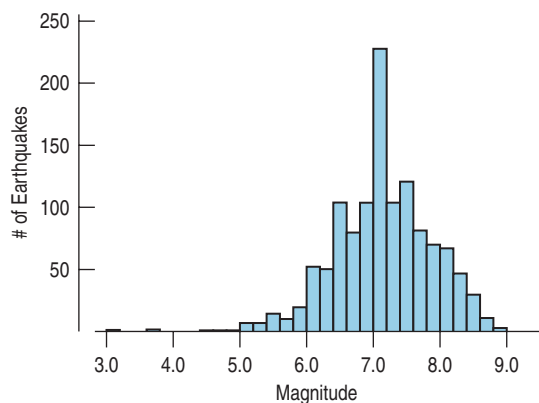


FIGURE 4.1

A histogram of earthquake magnitudes shows the number of earthquakes with magnitudes (in Richter scale units) in each bin.

One surprising feature of the earthquake magnitudes is the spike around magnitude 7.0. Only one other bin holds even half that many earthquakes. These values include historical data for which the magnitudes were estimated by experts and not measured by modern seismographs. Perhaps the experts thought 7 was a typical and reasonable value for a tsunami-causing earthquake when they lacked detailed information. That would explain the overabundance of magnitudes right at 7.0 rather than spread out near that value.

Like a bar chart, a histogram plots the bin counts as the heights of bars. In this histogram of earthquake magnitudes, each bin has a width of 0.2, so, for example, the height of the tallest bar says that there were about 230 earthquakes with magnitudes between 7.0 and 7.2. In this way, the histogram displays the entire distribution of earthquake magnitudes.

Does the distribution look as you expected? It is often a good idea to *imagine* what the distribution might look like before you make the display. That way you'll be less likely to be fooled by errors in the data or when you accidentally graph the wrong variable.

From the histogram, we can see that these earthquakes typically have magnitudes around 7. Most are between 5.5 and 8.5, and some are as small as 3 and as big as 9. Now we can answer the question about the Sumatra tsunami. With a value of 9.0 it's clear that the earthquake that caused it was an extraordinarily powerful earthquake—one of the largest on record.³

The bar charts of categorical variables we saw in Chapter 3 had spaces between the bars to separate the counts of different categories. But in a histogram, the bins slice up *all the values* of the quantitative variable, so any spaces in a histogram are actual **gaps** in the data, indicating a region where there are no values.

Sometimes it is useful to make a **relative frequency histogram**, replacing the counts on the vertical axis with the *percentage* of the total number of cases falling in each bin. Of course, the shape of the histogram is exactly the same; only the vertical scale is different.

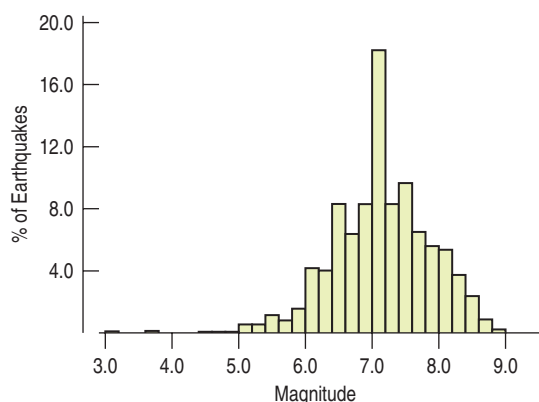


FIGURE 4.2

A relative frequency histogram looks just like a frequency histogram except for the labels on the y-axis, which now show the percentage of earthquakes in each bin.

³ Some experts now estimate the magnitude at between 9.1 and 9.3.

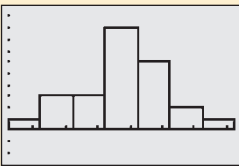
T1 Tips

Making a histogram

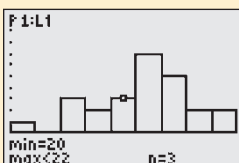
L1	L2	L3	1
22			
17			
18			
29			
22			
22			
23			
24			
23			
17			
21			
25			
20			
L1 = {22, 17, 18, 29...			

STAT PLOT
1: Plot1...On
2: Plot2...Off
3: Plot3...Off
4: PlotsOff

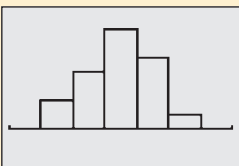
2nd	STAT	Plot1	Plot2	Plot3
On	Off	Off		
Type:				
Xlist:	L1			
Freq:	1			



WINDOW
Xmin=12
Xmax=30
Xscl=2
Ymin=-2.70621
Ymax=10.53
Yscl=1
Xres=3



L1	L2	L3	3
22	60		
17	70		
18	80		
29	90		
22	100		
22			
23			
L3{6} =			



Your calculator can create histograms. First you need some data. For an agility test, fourth-grade children jump from side to side across a set of parallel lines, counting the number of lines they clear in 30 seconds. Here are their scores:

22, 17, 18, 29, 22, 22, 23, 24, 23, 17, 21, 25, 20
12, 19, 28, 24, 22, 21, 25, 26, 25, 16, 27, 22

Enter these data into **L1**.

Now set up the calculator's plot:

- Go to **2nd** **STATPLOT**, choose **Plot1**, then **ENTER**.
- In the **Plot1** screen choose **On**, select the little histogram icon, then specify **Xlist:L1** and **Freq:1**.
- Be sure to turn off any other graphs the calculator may be set up for. Just hit the **Y=** button, and deactivate any functions seen there.

All set? To create your preliminary plot go to **ZOOM**, select **9:ZoomStat**, and then **ENTER**.

You now see the calculator's initial attempt to create a histogram of these data. Not bad. We can see that the distribution is roughly symmetric. But it's hard to tell exactly what this histogram shows, right? Let's fix it up a bit.

- Under **WINDOW**, let's reset the bins to convenient, sensible values. Try **Xmin=12**, **Xmax=30** and **Xscl=2**. That specifies the range of values along the *x*-axis and makes each bar span two lines.
- Hit **GRAPH** (not **ZoomStat**—this time we want control of the scale!).

There. We still see rough symmetry, but also see that one of the scores was much lower than the others. Note that you can now find out exactly what the bars indicate by activating **TRACE** and then moving across the histogram using the arrow keys. For each bar the calculator will indicate the interval of values and the number of data values in that bin. We see that 3 kids had agility scores of 20 or 21.

Play around with the **WINDOW** settings. A different **Ymax** will make the bars appear shorter or taller. What happens if you set the bar width (**Xscl**) smaller? Or larger? You don't want to lump lots of values into just a few bins or make so many bins that the overall shape of the histogram is not clear. Choosing the best bar width takes practice.

Finally, suppose the data are given as a frequency table. Consider a set of test scores, with two grades in the 60s, four in the 70s, seven in the 80s, five in the 90s, and one 100. Enter the group cutoffs 60, 70, 80, 90, 100 in **L2** and the corresponding frequencies 2, 4, 7, 5, 1 in **L3**. When you set up the histogram **STATPLOT**, specify **Xlist:L2** and **Freq:L3**. Can you specify the **WINDOW** settings to make this histogram look the way you want it? (By the way, if you get a **DIM MISMATCH** error, it means you can't count. Look at **L2** and **L3**; you'll see the two lists don't have the same number of entries. Fix the problem by correcting the data you entered.)

Stem-and-Leaf Displays

Histograms provide an easy-to-understand summary of the distribution of a quantitative variable, but they don't show the data values themselves. Here's a histogram of the pulse rates of 24 women, taken by a researcher at a health clinic:

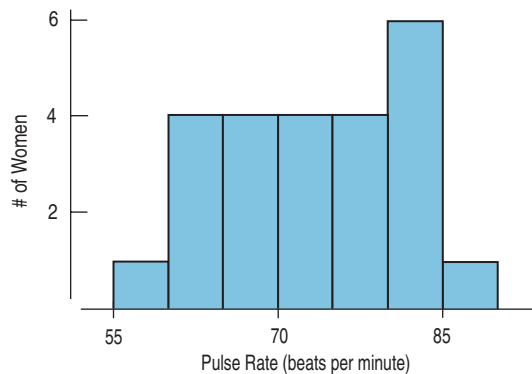
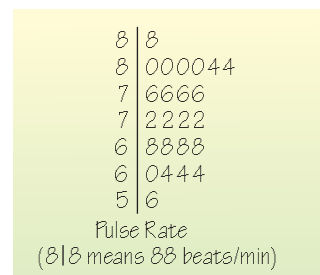


FIGURE 4.3
The pulse rates of 24 women at a health clinic

The Stem-and-Leaf display was devised by John W. Tukey, one of the greatest statisticians of the 20th century. It is called a "Stemplot" in some texts and computer programs, but we prefer Tukey's original name for it.

The story seems pretty clear. We can see the entire span of the data and can easily see what a typical pulse rate might be. But is that all there is to these data?

A **stem-and-leaf display** is like a histogram, but it shows the individual values. It's also easier to make by hand. Here's a stem-and-leaf display of the same data:



AS **Activity: Stem-and-Leaf Displays.** As you might expect of something called "stem-and-leaf," these displays grow as you consider each data value.

Turn the stem-and-leaf on its side (or turn your head to the right) and squint at it. It should look roughly like the histogram of the same data. Does it? Well, it's backwards because now the higher values are on the left, but other than that, it has the same shape.⁴

What does the line at the top of the display that says 8 | 8 mean? It stands for a pulse of 88 beats per minute (bpm). We've taken the tens place of the number and made that the "stem." Then we sliced off the ones place and made it a "leaf." The next line down is 8 | 000044. That shows that there were four pulse rates of 80 and two of 84 bpm.

Stem-and-leaf displays are especially useful when you make them by hand for batches of fewer than a few hundred data values. They are a quick way to display—and even to record—numbers. Because the leaves show the individual values, we can sometimes see even more in the data than the distribution's shape. Take another look at all the leaves of the pulse data. See anything

⁴ You could make the stem-and-leaf with the higher values on the bottom. Usually, though, higher on the top makes sense.

unusual? At a glance you can see that they are all even. With a bit more thought you can see that they are all multiples of 4—something you couldn't possibly see from a histogram. How do you think the nurse took these pulses? Counting beats for a full minute or counting for only 15 seconds and multiplying by 4?

How do stem-and-leaf displays work? Stem-and-leaf displays work like histograms, but they show more information. They use part of the number itself (called the stem) to name the bins. To make the “bars,” they use the next digit of the number. For example, if we had a test score of 83, we could write it 8|3, where 8 serves as the stem and 3 as the leaf. Then, to display the scores 83, 76, and 88 together, we would write

$$\begin{array}{r|l} 8 & 38 \\ 7 & 6 \end{array}$$

For the pulse data, we have

$$\begin{array}{r|l} 8 & 0000448 \\ 7 & 22226666 \\ 6 & 04448888 \\ 5 & 6 \\ \hline \end{array}$$

Pulse Rate
(5|6 means 56 beats/min)

This display is OK, but a little crowded. A histogram might split each line into two bars. With a stem-and-leaf, we can do the same by putting the leaves 0–4 on one line and 5–9 on another, as we saw above:

$$\begin{array}{r|l} 8 & 8 \\ 8 & 000044 \\ 7 & 6666 \\ 7 & 2222 \\ 6 & 8888 \\ 6 & 0444 \\ 5 & 6 \\ \hline \end{array}$$

Pulse Rate
(8|8 means 88 beats/min)

For numbers with three or more digits, you'll often decide to truncate (or round) the number to two places, using the first digit as the stem and the second as the leaf. So, if you had 432, 540, 571, and 638, you might display them as shown below with an indication that 6|3 means 630–639.

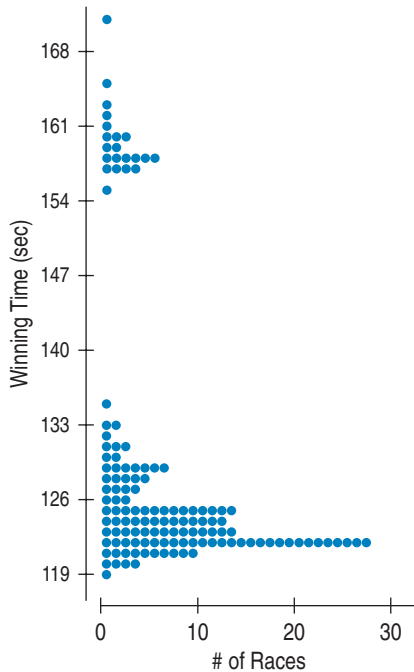
$$\begin{array}{r|l} 6 & 3 \\ 5 & 47 \\ 4 & 3 \end{array}$$

When you make a stem-and-leaf by hand, make sure to give each digit the same width, in order to preserve the area principle. (That can lead to some fat 1's and thin 8's—but it makes the display honest.)

Dotplots

AS

Activity: Dotplots. Click on points to see their values and even drag them around.



A **dotplot** is a simple display. It just places a dot along an axis for each case in the data. It's like a stem-and-leaf display, but with dots instead of digits for all the leaves. Dotplots are a great way to display a small data set (especially if you forget how to write the digits from 0 to 9). Here's a dotplot of the time (in seconds) that the winning horse took to win the Kentucky Derby in each race between the first Derby in 1875 and the 2008 Derby.

Dotplots show basic facts about the distribution. We can find the slowest and quickest races by finding times for the topmost and bottommost dots. It's also clear that there are two clusters of points, one just below 160 seconds and the other at about 122 seconds. Something strange happened to the Derby times. Once we know to look for it, we can find out that in 1896 the distance of the Derby race was changed from 1.5 miles to the current 1.25 miles. That explains the two clusters of winning times.

Some dotplots stretch out horizontally, with the counts on the vertical axis, like a histogram. Others, such as the one shown here, run vertically, like a stem-and-leaf display. Some dotplots place points next to each other when they would otherwise overlap. Others just place them on top of one another. Newspapers sometimes offer dotplots with the dots made up of little pictures.

FIGURE 4.4

A dotplot of Kentucky Derby winning times plots each race as its own dot, showing the bimodal distribution.

Think Before You Draw, Again

Suddenly, we face a lot more options when it's time to invoke our first rule of data analysis and make a picture. You'll need to *Think* carefully to decide which type of graph to make. In the previous chapter you learned to check the Categorical Data Condition before making a pie chart or a bar chart. Now, before making a stem-and-leaf display, a histogram, or a dotplot, you need to check the

Quantitative Data Condition: The data are values of a quantitative variable whose units are known.

Although a bar chart and a histogram may look somewhat similar, they're not the same display. You can't display categorical data in a histogram or quantitative data in a bar chart. Always check the condition that confirms what type of data you have before proceeding with your display.

Step back from a histogram or stem-and-leaf display. What can you say about the distribution? When you describe a distribution, you should always tell about three things: its **shape, center, and spread.**

The Shape of a Distribution

1. Does the histogram have a single, central hump or several separated humps? These humps are called **modes**.⁵ The earthquake magnitudes have a single mode

⁵ Well, technically, it's the value on the horizontal axis of the histogram that is the mode, but anyone asked to point to the mode would point to the hump.

The **mode** is sometimes defined as the single value that appears most often. That definition is fine for categorical variables because all we need to do is count the number of cases for each category. For quantitative variables, the mode is more ambiguous. What is the mode of the Kentucky Derby times? Well, seven races were timed at 122.2 seconds—more than any other race time. Should that be the mode? Probably not. For quantitative data, it makes more sense to use the term “mode” in the more general sense of the peak of the histogram rather than as a single summary value. In this sense, the important feature of the Kentucky Derby races is that there are two distinct modes, representing the two different versions of the race and warning us to consider those two versions separately.

at just about 7. A histogram with one peak, such as the earthquake magnitudes, is dubbed **unimodal**; histograms with two peaks are **bimodal**, and those with three or more are called **multimodal**.⁶ For example, here’s a bimodal histogram.

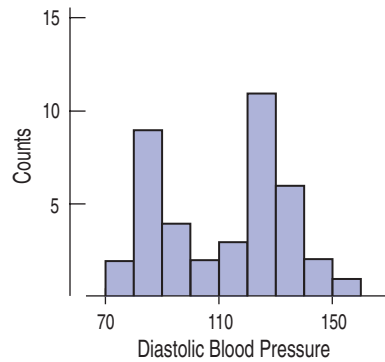


FIGURE 4.5
A bimodal histogram has two apparent peaks.

A histogram that doesn’t appear to have any mode and in which all the bars are approximately the same height is called **uniform**.

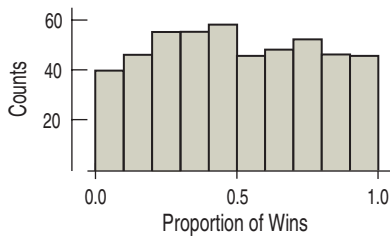


FIGURE 4.6
In a uniform histogram, the bars are all about the same height. The histogram doesn’t appear to have a mode.

You’ve heard of pie à la mode. Is there a connection between pie and the mode of a distribution? Actually, there is! The mode of a distribution is a *popular* value near which a lot of the data values gather. And “à la mode” means “in style”—not “with ice cream.” That just happened to be a *popular* way to have pie in Paris around 1900.

2. *Is the histogram symmetric?* Can you fold it along a vertical line through the middle and have the edges match pretty closely, or are more of the values on one side?

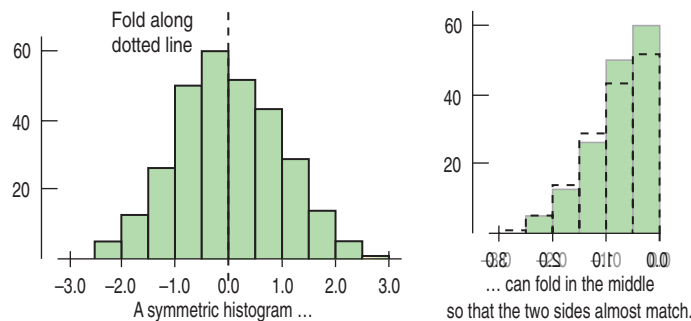


FIGURE 4.7

The (usually) thinner ends of a distribution are called the **tails**. If one tail stretches out farther than the other, the histogram is said to be **skewed** to the side of the longer tail.

⁶ Apparently, statisticians don’t like to count past two.

AS **Activity: Attributes of Distribution Shape.** This activity and the others on this page show off aspects of distribution shape through animation and example, then let you make and interpret histograms with your statistics package.

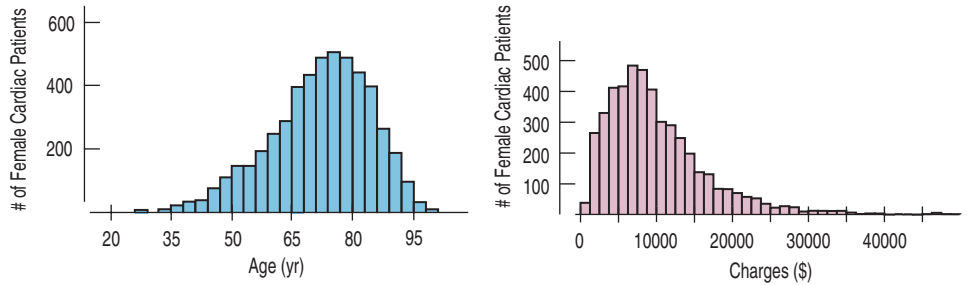
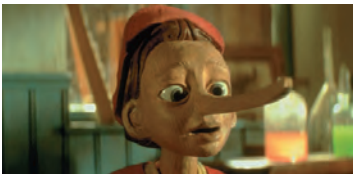


FIGURE 4.8

Two skewed histograms showing data on two variables for all female heart attack patients in New York state in one year. The blue one (age in years) is skewed to the left. The purple one (charges in \$) is skewed to the right.



3. *Do any unusual features stick out?* Often such features tell us something interesting or exciting about the data. You should always mention any stragglers, or outliers, that stand off away from the body of the distribution. If you're collecting data on nose lengths and Pinocchio is in the group, you'd probably notice him, and you'd certainly want to mention it.

Outliers can affect almost every method we discuss in this course. So we'll always be on the lookout for them. An outlier can be the most informative part of your data. Or it might just be an error. But don't throw it away without comment. Treat it specially and discuss it when you tell about your data. Or find the error and fix it if you can. Be sure to look for outliers. Always.

In the next chapter you'll learn a handy rule of thumb for deciding when a point might be considered an outlier.

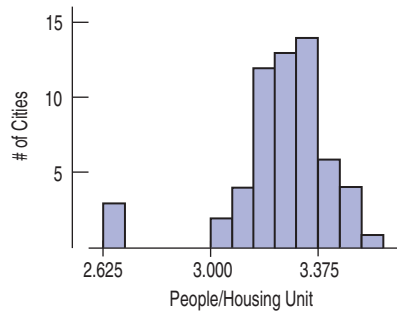


FIGURE 4.9

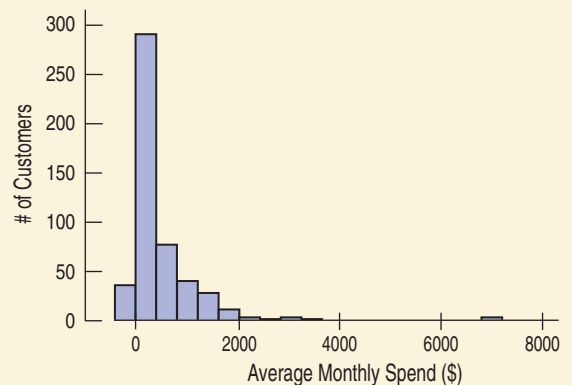
A histogram with outliers. There are three cities in the leftmost bar.

FOR EXAMPLE Describing histograms

A credit card company wants to see how much customers in a particular segment of their market use their credit card. They have provided you with data⁷ on the amount spent by 500 selected customers during a 3-month period and have asked you to summarize the expenditures. Of course, you begin by making a histogram.

Question: Describe the shape of this distribution.

The distribution of expenditures is unimodal and skewed to the high end. There is an extraordinarily large value at about \$7000, and some of the expenditures are negative.



⁷These data are real, but cannot be further identified for obvious privacy reasons.

Are there any gaps in the distribution? The Kentucky Derby data that we saw in the dotplot on page 49 has a large gap between two groups of times, one near 120 seconds and one near 160. Gaps help us see multiple modes and encourage us to notice when the data may come from different sources or contain more than one group.



Toto, I've a feeling we're not in math class anymore . . . When Dorothy and her dog Toto land in Oz, everything is more vivid and colorful, but also more dangerous and exciting. Dorothy has new choices to make. She can't always rely on the old definitions, and the yellow brick road has many branches. You may be coming to a similar realization about Statistics.

When we summarize data, our goal is usually more than just developing a detailed knowledge of the data we have at hand. Scientists generally don't care about the particular guinea pigs they've treated, but rather about what their reactions say about how animals (and, perhaps, humans) would respond.

When you look at data, you want to know what the data say about the world, so you'd like to know whether the patterns you see in histograms and summary statistics generalize to other individuals and situations. You'll want to calculate summary statistics accurately, but then you'll also want to think about what they may say beyond just describing the data. And your knowledge about the world matters when you think about the overall meaning of your analysis.

It may surprise you that many of the most important concepts in Statistics are not defined as precisely as most concepts in mathematics. That's done on purpose, to leave room for judgment.

Because we want to see broader patterns rather than focus on the details of the data set we're looking at, we deliberately leave some statistical concepts a bit vague. Whether a histogram is symmetric or skewed, whether it has one or more modes, whether a point is far enough from the rest of the data to be considered an outlier—these are all somewhat vague concepts. And they all require judgment. You may be used to finding a single correct and precise answer, but in Statistics, there may be more than one interpretation. That may make you a little uncomfortable at first, but soon you'll see that this room for judgment brings you enormous power and responsibility. It means that using your own knowledge and judgment and supporting your findings with statistical evidence and justifications entitles you to your own opinions about what you see.



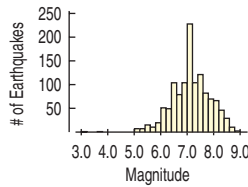
JUST CHECKING

It's often a good idea to think about what the distribution of a data set might look like before we collect the data. What do you think the distribution of each of the following data sets will look like? Be sure to discuss its shape. Where do you think the center might be? How spread out do you think the values will be?

1. Number of miles run by Saturday morning joggers at a park.
2. Hours spent by U.S. adults watching football on Thanksgiving Day.
3. Amount of winnings of all people playing a particular state's lottery last week.
4. Ages of the faculty members at your school.
5. Last digit of phone numbers on your campus.

The Center of the Distribution: The Median

Let's return to the tsunami earthquakes. But this time, let's look at just 25 years of data: 176 earthquakes that occurred from 1981 through 2005. These should be more accurately measured than prehistoric quakes because seismographs were in wide use. Try to put your finger on the histogram at the value you think is



typical. (Read the value from the horizontal axis and remember it.) When we think of a typical value, we usually look for the **center** of the distribution. Where do you think the center of this distribution is? For a unimodal, symmetric distribution such as these earthquake data, it's easy. We'd all agree on the center of symmetry, where we would fold the histogram to match the two sides. But when the distribution is skewed or possibly multimodal, it's not immediately clear what we even mean by the center.

One reasonable choice of typical value is the value that is literally in the middle, with half the values below it and half above it.

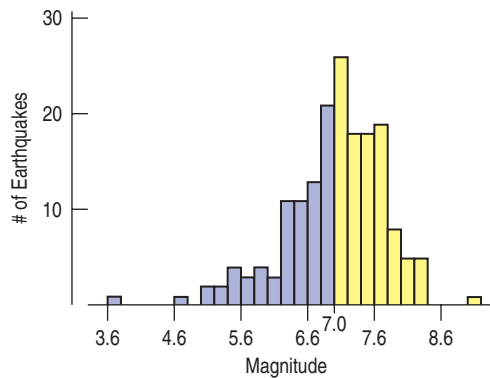


FIGURE 4.10 *Tsunami-causing earthquakes (1981–2005)*

The median splits the histogram into two halves of equal area.

Histograms follow the area principle, and each half of the data has about 88 earthquakes, so each colored region has the same area in the display. The middle value that divides the histogram into two equal areas is called the **median**.

The median has the same units as the data. Be sure to include the units whenever you discuss the median.

For the recent tsunamis, there are 176 earthquakes, so the median is found at the $(176 + 1)/2 = 88.5$ th place in the sorted data. That “.5” just says to average the two values on either side: the 88th and the 89th. The median earthquake magnitude is 7.0.

NOTATION ALERT:

We always use n to indicate the number of values. Some people even say, “How big is the n ?” when they mean the number of data values.

How do medians work? Finding the median of a batch of n numbers is easy as long as you remember to order the values first. If n is odd, the median is the middle value. Counting in from the ends, we find this value in the $\frac{n+1}{2}$ position.

When n is even, there are two middle values. So, in this case, the median is the average of the two values in positions $\frac{n}{2}$ and $\frac{n}{2} + 1$.

Here are two examples:

Suppose the batch has these values: 14.1, 3.2, 25.3, 2.8, -17.5 , 13.9, 45.8.

First we order the values: -17.5 , 2.8, 3.2, 13.9, 14.1, 25.3, 45.8.

Since there are 7 values, the median is the $(7 + 1)/2 = 4$ th value, counting from the top or bottom: 13.9. Notice that 3 values are lower, 3 higher.

Suppose we had the same batch with another value at 35.7. Then the ordered values are -17.5 , 2.8, 3.2, 13.9, 14.1, 25.3, 35.7, 45.8.

The median is the average of the $8/2$ or 4th, and the $(8/2) + 1$, or 5th, values. So the median is $(13.9 + 14.1)/2 = 14.0$. Four data values are lower, and four higher.

The median is one way to find the center of the data. But there are many others. We'll look at an even more important measure later in this chapter.

Knowing the median, we could say that a typical tsunami-causing earthquake, worldwide, was about 7.0 on the Richter scale. How much does that really say? How well does the median describe the data? After all, not every earthquake has a Richter scale value of 7.0. Whenever we find the center of data, the next step is always to ask how well it actually summarizes the data.

Spread: Home on the Range

Statistics pays close attention to what we *don't* know as well as what we do know. Understanding how spread out the data are is a first step in understanding what a summary *cannot* tell us about the data. It's the beginning of telling us what we don't know.

If every earthquake that caused a tsunami registered 7.0 on the Richter scale, then knowing the median would tell us everything about the distribution of earthquake magnitudes. The more the data vary, however, the less the median alone can tell us. So we need to measure how much the data values vary around the center. In other words, how spread out are they? **When we describe a distribution numerically, we always report a measure of its spread along with its center.**

How should we measure the spread? We could simply look at the extent of the data. How far apart are the two extremes? **The range of the data is defined as the difference between the maximum and minimum values:**

$$\text{Range} = \text{max} - \text{min}.$$

Notice that the range is a *single number*, not an interval of values, as you might think from its use in common speech. The maximum magnitude of these earthquakes is 9.0 and the minimum is 3.7, so the *range* is $9.0 - 3.7 = 5.3$.

The range has the disadvantage that a single extreme value can make it very large, giving a value that doesn't really represent the data overall.

Spread: The Interquartile Range

A better way to describe the spread of a variable might be to ignore the extremes and concentrate on the middle of the data. We could, for example, find the range of just the middle half of the data. What do we mean by the middle half? Divide the data in half at the median. Now divide both halves in half again, cutting the data into four quarters. We call these new dividing points **quartiles**. **One quarter of the data lies below the lower quartile, and one quarter of the data lies above the upper quartile, so half the data lies between them. The quartiles border the middle half of the data.**

How do quartiles work? A simple way to find the quartiles is to start by splitting the batch into two halves at the median. (When n is odd, some statisticians include the median in both halves; others omit it.) The lower quartile is the median of the lower half, and the upper quartile is the median of the upper half.

Here are our two examples again.

The ordered values of the first batch were $-17.5, 2.8, 3.2, 13.9, 14.1, 25.3,$ and 45.8 , with a median of 13.9 . Excluding the median, the two halves of the list are $-17.5, 2.8, 3.2$ and $14.1, 25.3, 45.8$.

Each half has 3 values, so the median of each is the middle one. The lower quartile is 2.8 , and the upper quartile is 25.3 .

The second batch of data had the ordered values $-17.5, 2.8, 3.2, 13.9, 14.1, 25.3, 35.7,$ and 45.8 .

Here n is even, so the two halves of 4 values are $-17.5, 2.8, 3.2, 13.9$ and $14.1, 25.3, 35.7, 45.8$.

Now the lower quartile is $(2.8 + 3.2)/2 = 3.0$, and the upper quartile is $(25.3 + 35.7)/2 = 30.5$.

The difference between the quartiles tells us how much territory the middle half of the data covers and is called the **interquartile range**. It's commonly abbreviated IQR (and pronounced "eye-cue-are," not "ikker"):

$$IQR = \text{upper quartile} - \text{lower quartile}.$$

For the earthquakes, there are 88 values below the median and 88 values above the median. The midpoint of the lower half is the average of the 44th and 45th values in the ordered data; that turns out to be 6.6. In the upper half we average the 132nd and 133rd values, finding a magnitude of 7.6 as the third quartile. The *difference* between the quartiles gives the IQR:

$$IQR = 7.6 - 6.6 = 1.0.$$

Now we know that the middle half of the earthquake magnitudes extends across a (interquartile) range of 1.0 Richter scale units. This seems like a reasonable summary of the spread of the distribution, as we can see from this histogram:

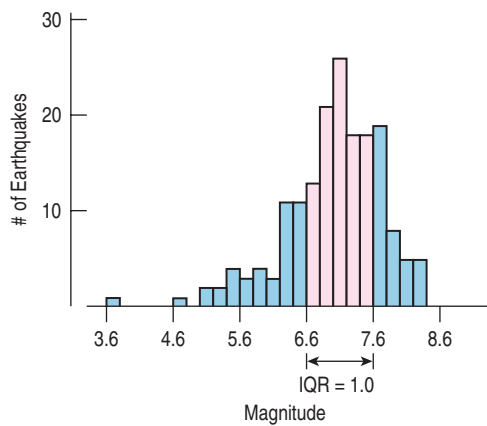


FIGURE 4.11

The quartiles bound the middle 50% of the values of the distribution. This gives a visual indication of the spread of the data. Here we see that the IQR is 1.0 Richter scale units.

The IQR is almost always a reasonable summary of the spread of a distribution. Even if the distribution itself is skewed or has some outliers, the IQR should provide useful information. The one exception is when the data are strongly bimodal. For example, remember the dotplot of winning times in the Kentucky Derby (page 49)? Because the race distance was changed, we really have data on two different races, and they shouldn't be summarized together.

So, what is a quartile anyway? Finding the quartiles sounds easy, but surprisingly, the quartiles are not well-defined. It's not always clear how to find a value such that exactly one quarter of the data lies above or below that value. We offered a simple rule for Finding Quartiles in the box on page 54: Find the median of each half of the data split by the median. When n is odd, we (and your TI calculator) omit the median from each of the halves. Some other texts include the median in both halves before finding the quartiles. Both methods are commonly used. If you are willing to do a bit more calculating, there are several other methods that locate a quartile somewhere between adjacent data values. We know of at least six different rules for finding quartiles. Remarkably, each one is in use in some software package or calculator.

So don't worry too much about getting the "exact" value for a quartile. All of the methods agree pretty closely when the data set is large. When the data set is small, different rules will disagree more, but in that case there's little need to summarize the data anyway.

Remember, Statistics is about understanding the world, not about calculating the right number. The "answer" to a statistical question is a sentence about the issue raised in the question.

The lower and upper quartiles are also known as the 25th and 75th percentiles of the data, respectively, since the lower quartile falls above 25% of the data and the upper quartile falls above 75% of the data. If we count this way, the median is the 50th percentile. We could, of course, define and calculate any percentile that we want. For example, the 10th percentile would be the number that falls above the lowest 10% of the data values.

5-Number Summary

NOTATION ALERT:

We always use Q1 to label the lower (25%) quartile and Q3 to label the upper (75%) quartile. We skip the number 2 because the median would, by this system, naturally be labeled Q2—but we don't usually call it that.

The **5-number summary** of a distribution reports its median, quartiles, and extremes (maximum and minimum). The 5-number summary for the recent tsunami earthquake *Magnitudes* looks like this:

Max	9.0
Q3	7.6
Median	7.0
Q1	6.6
Min	3.7

It's good idea to report the number of data values and the identity of the cases (the *Who*). Here there are 176 earthquakes.

The 5-number summary provides a good overview of the distribution of magnitudes of these tsunami-causing earthquakes. For a start, we can see that the median magnitude is 7.0. Because the IQR is only $7.6 - 6.6 = 1$, we see that many quakes are close to the median magnitude. Indeed, the quartiles show us that the middle half of these earthquakes had magnitudes between 6.6 and 7.6. One quarter of the earthquakes had magnitudes above 7.6, although one tsunami was caused by a quake measuring only 3.7 on the Richter scale.

STEP-BY-STEP EXAMPLE

Shape, Center, and Spread: Flight Cancellations



The U.S. Bureau of Transportation Statistics (www.bts.gov) reports data on airline flights. Let's look at data giving the percentage of flights cancelled each month between 1995 and 2005.

Question: How often are flights cancelled?

WHO	Months
WHAT	Percentage of flights cancelled at U.S. airports
WHEN	1995–2005
WHERE	United States



Variable: Identify the *variable*, and decide how you wish to display it.

To identify a variable, report the W's.

Select an appropriate display based on the nature of the data and what you want to know.

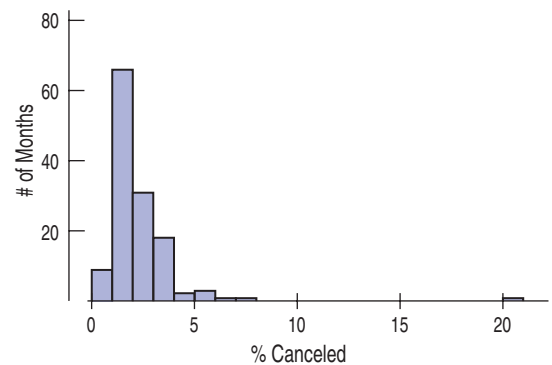
I want to learn about the monthly percentage of flight cancellations at U.S. airports.

I have data from the U.S. Bureau of Transportation Statistics giving the percentage of flights cancelled at U.S. airports each month between 1995 and 2005.

✓ **Quantitative Data Condition:** Percentages are quantitative. A histogram and numerical summaries would be appropriate.



Mechanics: We usually make histograms with a computer or graphing calculator.



The histogram shows a distribution skewed to the high end and one extreme outlier, a month in which more than 20% of flights were cancelled.

In most months, fewer than 5% of flights are cancelled and usually only about 2% or 3%. That seems reasonable.

REALITY CHECK

It's always a good idea to think about what you expect to see so that you can check whether the histogram looks like what you expected.

With 132 cases, we probably have more data than you'd choose to work with by hand. The results given here are from technology.

Count	132
Max	20.240
Q3	2.615
Median	1.755
Q1	1.445
Min	0.770
IQR	1.170

TELL

Interpretation: Describe the shape, center, and spread of the distribution. Report on the symmetry, number of modes, and any gaps or outliers. You should also mention any concerns you may have about the data.

The distribution of cancellations is skewed to the right, and this makes sense: The values can't fall below 0%, but can increase almost arbitrarily due to bad weather or other events.

The median is 1.76% and the IQR is 1.17%. The low IQR indicates that in most months the cancellation rate is close to the median. In fact, it's between 1.4% and 2.6% in the middle 50% of all months, and in only 1/4 of the months were more than 2.6% of flights cancelled.

There is one extraordinary value: 20.2%. Looking it up, I find that the extraordinary month was September 2001. The attacks of September 11 shut down air travel for several days, accounting for this outlier.

Summarizing Symmetric Distributions: The Mean

NOTATION ALERT:

In Algebra you used letters to represent values in a problem, but it didn't matter what letter you picked. You could call the width of a rectangle X or you could call it w (or *Fred*, for that matter). But in Statistics, the notation is part of the vocabulary. For example, in Statistics n is always the number of data values. Always.

We have already begun to point out such special notation conventions: n , $Q1$, and $Q3$. Think of them as part of the terminology you need to learn in this course.

Here's another one: Whenever we put a bar over a symbol, it means "find the mean."

Medians do a good job of summarizing the center of a distribution, even when the shape is skewed or when there is an outlier, as with the flight cancellations. But when we have symmetric data, there's another alternative. You probably already know how to average values. In fact, to find the median when n is even, we said you should average the two middle values, and you didn't even flinch.

The earthquake magnitudes are pretty close to symmetric, so we can also summarize their center with a mean. The mean tsunami earthquake magnitude is 6.96—about what we might expect from the histogram. You already know how to average values, but this is a good place to introduce notation that we'll use throughout the book. We use the Greek capital letter sigma, Σ , to mean "sum" (sigma is "S" in Greek), and we'll write:

$$\bar{y} = \frac{\text{Total}}{n} = \frac{\sum y}{n}.$$

The formula says to add up all the values of the variable and divide that sum by the number of data values, n —just as you've always done.⁸

Once we've averaged the data, you'd expect the result to be called the *average*, but that would be too easy. Informally, we speak of the "average person" but we don't add up people and divide by the number of people. A median is also a kind of average. To make this distinction, the value we calculated is called the mean, \bar{y} , and pronounced "y-bar."

⁸ You may also see the variable called x and the equation written $\bar{x} = \frac{\text{Total}}{n} = \frac{\sum x}{n}$. Don't let that throw you. You are free to name the variable anything you want, but we'll generally use y for variables like this that we want to summarize, model, or predict. (Later we'll talk about variables that are used to explain, model, or predict y . We'll call them x .)

The **mean** feels like the center because it is the point where the histogram balances:

In everyday language, sometimes “average” does mean what we want it to mean. We don’t talk about your grade point mean or a baseball player’s batting mean or the Dow Jones Industrial mean. So we’ll continue to say “average” when that seems most natural. When we do, though, you may assume that what we mean is the mean.

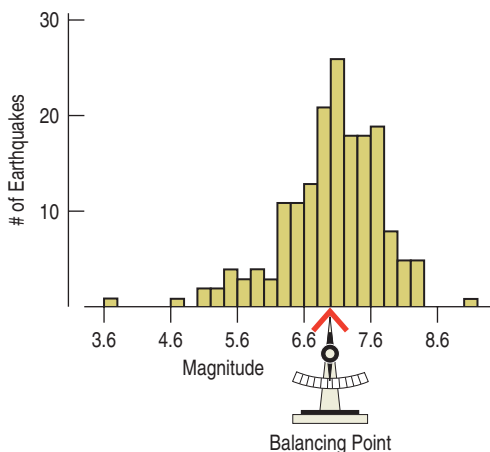


FIGURE 4.12
The mean is located at the balancing point of the histogram.

Mean or Median?

Using the center of balance makes sense when the data are symmetric. But data are not always this well behaved. If the distribution is skewed or has outliers, the center is not so well defined and the mean may not be what we want. For example, the mean of the flight cancellations doesn’t give a very good idea of the typical percentage of cancellations.

TI-*nspire*
Mean, median, and outliers.
Drag data points around to explore how outliers affect the mean and median.

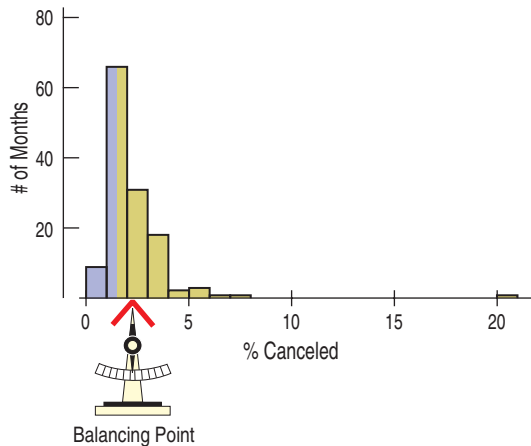


FIGURE 4.13
The median splits the area of the histogram in half at 1.75%. Because the distribution is skewed to the right, the mean (2.28%) is higher than the median. The points at the right have pulled the mean toward them away from the median.

A S **Activity: The Center of a Distribution.** Compare measures of center by dragging points up and down and seeing the consequences. Another activity shows how to find summaries with your statistics package.

The mean is 2.28%, but nearly 70% of months had cancellation rates below that, so the mean doesn’t feel like a good overall summary. Why is the balancing point so high? The large outlying value pulls it to the right. For data like these, the median is a better summary of the center.

Because the median considers only the order of the values, it is **resistant** to values that are extraordinarily large or small; it simply notes that they are one of the “big ones” or the “small ones” and ignores their distance from the center.

For the tsunami earthquake magnitudes, it doesn’t seem to make much difference—the mean is 6.96; the median is 7.0. When the data are symmetric, the mean and median will be close, but when the data are skewed, the median is likely to be a better choice. So, why not just use the median? Well, for one, the median can go overboard. It’s not just resistant to occasional outliers, but can be unaffected by changes in up to half the data values. By contrast, the mean includes input from

each data value and gives each one equal weight. It's also easier to work with, so when the distribution is unimodal and symmetric, we'll use the mean.

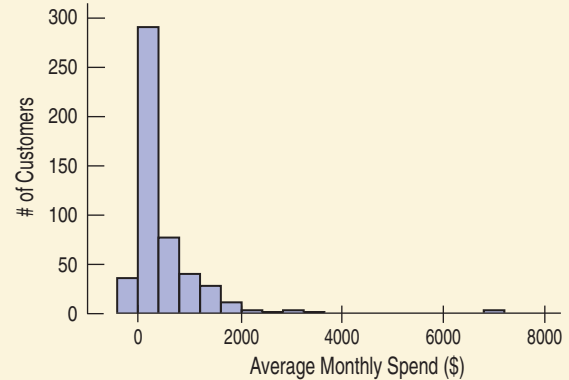
Of course, to choose between mean and median, we'll start by looking at the data. If the histogram is symmetric and there are no outliers, we'll prefer the mean. However, if the histogram is skewed or has outliers, we're usually better off with the median. If you're not sure, report both and discuss why they might differ.

FOR EXAMPLE

Describing center

Recap: You want to summarize the expenditures of 500 credit card company customers, and have looked at a histogram.

Question: You have found the mean expenditure to be \$478.19 and the median to be \$216.28. Which is the more appropriate measure of center, and why?



Because the distribution of expenditures is skewed, the median is the more appropriate measure of center. Unlike the mean, it's not affected by the large outlying value or by the skewness. Half of these credit card customers had average monthly expenditures less than \$216.28 and half more.

When to expect skewness Even without making a histogram, we can expect some variables to be skewed. When values of a quantitative variable are bounded on one side but not the other, the distribution may be skewed. For example, incomes and waiting times can't be less than zero, so they are often skewed to the right. Amounts of things (dollars, employees) are often skewed to the right for the same reason. If a test is too easy, the distribution will be skewed to the left because many scores will bump against 100%. And combinations of things are often skewed. In the case of the cancelled flights, flights are more likely to be cancelled in January (due to snowstorms) and in August (thunderstorms). Combining values across months leads to a skewed distribution.

What About Spread? The Standard Deviation

AS **Activity: The Spread of a Distribution.** What happens to measures of spread when data values change may not be quite what you expect.

The IQR is always a reasonable summary of spread, but because it uses only the two quartiles of the data, it ignores much of the information about how individual values vary. A more powerful approach uses the **standard deviation**, which takes into account how far *each* value is from the mean. Like the mean, the standard deviation is appropriate only for symmetric data.

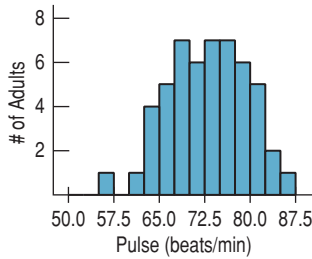
One way to think about spread is to examine how far each data value is from the mean. This difference is called a *deviation*. We could just average the deviations, but the positive and negative differences always cancel each other out. So the average deviation is always zero—not very helpful.

To keep them from canceling out, we *square* each deviation. Squaring always gives a positive value, so the sum won't be zero. That's great. Squaring also emphasizes larger differences—a feature that turns out to be both good and bad.

NOTATION ALERT:

s^2 always means the variance of a set of data, and s always denotes the standard deviation.

WHO 52 adults
WHAT Resting heart rates
UNITS Beats per minute



When we add up these squared deviations and find their average (almost), we call the result the **variance**:

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

Why almost? It *would* be a mean if we divided the sum by n . Instead, we divide by $n - 1$. Why? The simplest explanation is “to drive you crazy.” But there are good technical reasons, some of which we’ll see later.

The variance will play an important role later in this book, but it has a problem as a measure of spread. Whatever the units of the original data are, the variance is in *squared* units. We want measures of spread to have the same units as the data. And we probably don’t want to talk about squared dollars or *mpg*². So, to get back to the original units, we take the square root of s^2 . The result, s , is the **standard deviation**.

Putting it all together, the standard deviation of the data is found by the following formula:

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

You will almost always rely on a calculator or computer to do the calculating.

Understanding what the standard deviation really means will take some time, and we’ll revisit the concept in later chapters. For now, have a look at this histogram of resting pulse rates. The distribution is roughly symmetric, so it’s okay to choose the mean and standard deviation as our summaries of center and spread. The mean pulse rate is 72.7 beats per minute, and we can see that’s a typical heart rate. We also see that some heart rates are higher and some lower—but how much? Well, the standard deviation of 6.5 beats per minute indicates that, on average, we might expect people’s heart rates to differ from the mean rate by about 6.5 beats per minute. Looking at the histogram, we can see that 6.5 beats above or below the mean appears to be a typical deviation.

How does standard deviation work? To find the standard deviation, start with the mean, \bar{y} . Then find the *deviations* by taking \bar{y} from each value: $(y - \bar{y})$. Square each deviation: $(y - \bar{y})^2$.

Now you’re nearly home. Just add these up and divide by $n - 1$. That gives you the variance, s^2 . To find the standard deviation, s , take the square root. Here we go:

Suppose the batch of values is 14, 13, 20, 22, 18, 19, and 13.

The mean is $\bar{y} = 17$. So the deviations are found by subtracting 17 from each value:

Original Values	Deviations	Squared Deviations
14	$14 - 17 = -3$	$(-3)^2 = 9$
13	$13 - 17 = -4$	$(-4)^2 = 16$
20	$20 - 17 = 3$	9
22	$22 - 17 = 5$	25
18	$18 - 17 = 1$	1
19	$19 - 17 = 2$	4
13	$13 - 17 = -4$	16

Add up the squared deviations: $9 + 16 + 9 + 25 + 1 + 4 + 16 = 80$.

Now divide by $n - 1$: $80/6 = 13.33$.

Finally, take the square root: $s = \sqrt{13.33} = 3.65$

Thinking About Variation

AS **Activity: Displaying Spread.** What does the standard deviation look like on a histogram? How about the IQR?

Why do banks favor a single line that feeds several teller windows rather than separate lines for each teller? The average waiting time is the same. But the time you can expect to wait is less variable when there is a single line, and people prefer consistency.

Statistics is about variation, so spread is an important fundamental concept in Statistics. Measures of spread help us to be precise about what we *don't* know. If many data values are scattered far from the center, the IQR and the standard deviation will be large. If the data values are close to the center, then these measures of spread will be small. If all our data values were exactly the same, we'd have no question about summarizing the center, and all measures of spread would be zero—and we wouldn't need Statistics. You might think this would be a big plus, but it would make for a boring world. Fortunately (at least for Statistics), data do vary.

Measures of spread tell how well other summaries describe the data. That's why we always (always!) report a spread along with any summary of the center.



JUST CHECKING

- The U.S. Census Bureau reports the median family income in its summary of census data. Why do you suppose they use the median instead of the mean? What might be the disadvantages of reporting the mean?
- You've just bought a new car that claims to get a highway fuel efficiency of 31 miles per gallon. Of course, your mileage will "vary." If you had to guess, would you expect the IQR of gas mileage attained by all cars like yours to be 30 mpg, 3 mpg, or 0.3 mpg? Why?
- A company selling a new MP3 player advertises that the player has a mean lifetime of 5 years. If you were in charge of quality control at the factory, would you prefer that the standard deviation of lifespans of the players you produce be 2 years or 2 months? Why?

What to Tell About a Quantitative Variable

AS **Activity: Playing with Summaries.** Here's a Statistics game about summaries that even some experienced statisticians find . . . well, challenging. Your intuition may be better. Give it a try!

TI-*n*spire
Standard deviation, IQR, and outliers. Drag data points around to explore how outliers affect measures of spread.


What should you *Tell* about a quantitative variable?

- ▶ Start by making a histogram or stem-and-leaf display, and discuss the shape of the distribution.
- ▶ Next, discuss the center *and* spread.
 - ▶ We always pair the median with the IQR and the mean with the standard deviation. It's not useful to report one without the other. Reporting a center without a spread is dangerous. You may think you know more than you do about the distribution. Reporting only the spread leaves us wondering where we are.
 - ▶ If the shape is skewed, report the median and IQR. You may want to include the mean and standard deviation as well, but you should point out why the mean and median differ.
 - ▶ If the shape is symmetric, report the mean and standard deviation and possibly the median and IQR as well. For unimodal symmetric data, the IQR is usually a bit larger than the standard deviation. If that's not true of your data set, look again to make sure that the distribution isn't skewed and there are no outliers.

How “Accurate” Should We Be?

Don’t think you should report means and standard deviations to a zillion decimal places; such implied accuracy is really meaningless. Although there is no ironclad rule, statisticians commonly report summary statistics to one or two decimal places more than the original data have.

- ▶ Also, discuss any unusual features.
 - ▶ If there are multiple modes, try to understand why. If you can identify a reason for separate modes (for example, women and men typically have heart attacks at different ages), it may be a good idea to split the data into separate groups.
 - ▶ If there are any clear outliers, you should point them out. If you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be revealing. (Of course, the median and IQR won’t be affected very much by the outliers.)

STEP-BY-STEP EXAMPLE		Summarizing a distribution
<p>One of the authors owned a 1989 Nissan Maxima for 8 years. Being a statistician, he recorded the car’s fuel efficiency (in mpg) each time he filled the tank. He wanted to know what fuel efficiency to expect as “ordinary” for his car. (Hey, he’s a statistician. What would you expect?⁹) Knowing this, he was able to predict when he’d need to fill the tank again and to notice if the fuel efficiency suddenly got worse, which could be a sign of trouble.</p> <p>Question: How would you describe the distribution of <i>Fuel efficiency</i> for this car?</p>		
	<p>Plan State what you want to find out.</p> <p>Variable Identify the variable and report the W’s.</p> <p>Be sure to check the appropriate condition.</p>	<p>I want to summarize the distribution of Nissan Maxima fuel efficiency.</p> <p>The data are the fuel efficiency values in miles per gallon for the first 100 fill-ups of a 1989 Nissan Maxima between 1989 and 1992.</p> <p>✓ Quantitative Data Condition: The fuel efficiencies are quantitative with units of miles per gallon. Histograms and boxplots are appropriate displays for displaying the distribution. Numerical summaries are appropriate as well.</p>

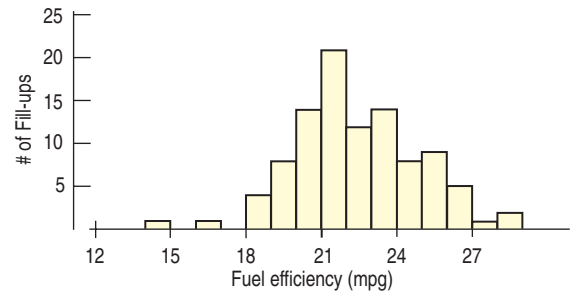
⁹ He also recorded the time of day, temperature, price of gas, and phase of the moon. (OK, maybe not phase of the moon.) His data are on the DVD.

SHOW

Mechanics Make a histogram and boxplot. Based on the shape, choose appropriate numerical summaries.

REALITY CHECK

A value of 22 mpg seems reasonable for such a car. The spread is reasonable, although the range looks a bit large.



A histogram of the data shows a fairly symmetric distribution with a low outlier.

Count	100
Mean	22.4 mpg
StdDev	2.45
Q1	20.8
Median	22.0
Q3	24.0
IQR	3.2

The mean and median are close, so the outlier doesn't seem to be a problem. I can use the mean and standard deviation.

TELL

Conclusion Summarize and interpret your findings in context. Be sure to discuss the distribution's shape, center, spread, and unusual features (if any).

The distribution of mileage is unimodal and roughly symmetric with a mean of 22.4 mpg. There is a low outlier that should be investigated, but it does not influence the mean very much. The standard deviation suggests that from tankful to tankful, I can expect the car's fuel economy to differ from the mean by an average of about 2.45 mpg.

Are my statistics “right”? When you calculate a mean, the computation is clear: You sum all the values and divide by the sample size. You may round your answer less or more than someone else (we recommend one more decimal place than the data), but all books and technologies agree on how to find the mean. Some statistics, however, are more problematic. For example we've already pointed out that methods of finding quartiles differ.

Differences in numeric results can also arise from decisions in the middle of calculations. For example, if you round off your value for the mean before you calculate the sum of squared deviations, your standard deviation probably won't agree with a computer program that calculates using many decimal places. (We do recommend that you do calculations using as many digits as you can to minimize this effect.)

Don't be overly concerned with these discrepancies, especially if the differences are small. They don't mean that your answer is “wrong,” and they usually won't change any conclusion you might draw about the data. Sometimes (in footnotes and in the answers in the back of the book) we'll note alternative results, but we could never list all the possible values, so we'll rely on your common sense to focus on the meaning rather than on the digits. Remember: Answers are sentences!

TI Tips

Calculating the statistics

```

EDIT [2nd] [MODE] TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
    
```

```

1-Var Stats L1
    
```

```

1-Var Stats
x=22
Σx=550
Σx²=12480
Sx=3.979112129
σx=3.898717738
n=25
    
```

```

1-Var Stats
n=25
minX=12
Q1=19.5
Med=22
Q3=25
maxX=29
    
```

Your calculator can easily find all the numerical summaries of data. To try it out, you simply need a set of values in one of your datalists. We'll illustrate using the boys' agility test results from this chapter's earlier TI Tips (still in L1), but you can use any data currently stored in your calculator.

- Under the **STAT** **CALC** menu, select **1-Var Stats** and hit **ENTER**.
- Specify the location of your data, creating a command like **1-Var Stats L1**.
- Hit **ENTER** again.

Voilà! Everything you wanted to know, and more. Among all of the information shown, you are primarily interested in these statistics: \bar{x} (the mean), Sx (the standard deviation), n (the count), and—scrolling down— $\min X$ (the smallest datum), Q_1 (the first quartile), Med (the median), Q_3 (the third quartile), and $\max X$ (the largest datum).

Sorry, but the TI doesn't explicitly tell you the range or the IQR. Just subtract: $\text{IQR} = Q_3 - Q_1 = 25 - 19.5 = 5.5$. What's the range?

By the way, if the data come as a frequency table with the values stored in, say, **L4** and the corresponding frequencies in **L5**, all you have to do is ask for **1-Var Stats L4,L5**.

WHAT CAN GO WRONG?

A data display should tell a story about the data. To do that, it must speak in a clear language, making plain what variable is displayed, what any axis shows, and what the values of the data are. And it must be consistent in those decisions.

A display of quantitative data can go wrong in many ways. The most common failures arise from only a few basic errors:

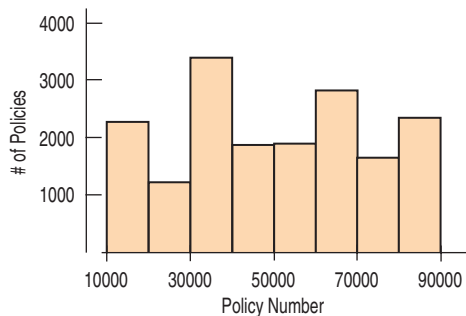


FIGURE 4.14
It's not appropriate to display these data with a histogram.

▶ **Don't make a histogram of a categorical variable.** Just because the variable contains numbers doesn't mean that it's quantitative. Here's a histogram of the insurance policy numbers of some workers. It's not very informative because the policy numbers are just labels. A histogram or stem-and-leaf display of a categorical variable makes no sense. A bar chart or pie chart would be more appropriate.

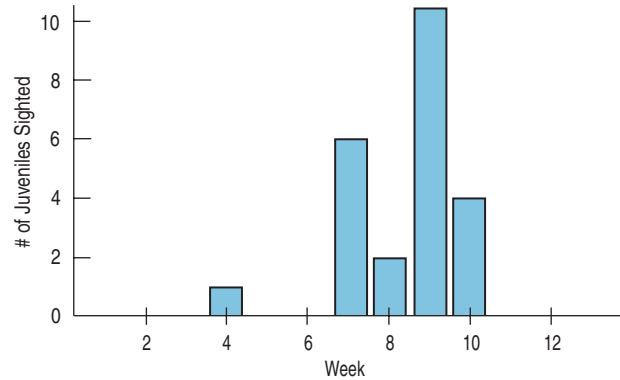
▶ **Don't look for shape, center, and spread of a bar chart.** A bar chart showing the sizes of the piles displays the distribution of a categorical variable, but the bars could be arranged in any order left to right. Concepts like symmetry, center, and spread make sense only for quantitative variables.

(continued)

- **Don't use bars in every display—save them for histograms and bar charts.** In a bar chart, the bars indicate how many cases of a categorical variable are piled in each category. Bars in a histogram indicate the number of cases piled in each interval of a quantitative variable. In both bar charts and histograms, the bars represent counts of data values. Some people create other displays that use bars to represent individual data values. Beware: Such graphs are neither bar charts nor histograms. For example, a student was asked to make a histogram from data showing the number of juvenile bald eagles seen during each of the 13 weeks in the winter of 2003–2004 at a site in Rock Island, IL. Instead, he made this plot:

FIGURE 4.15

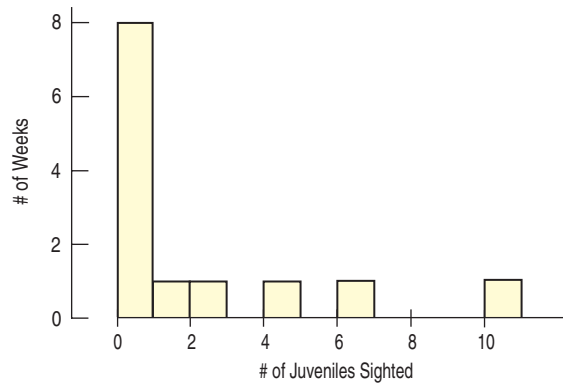
This isn't a histogram or a bar chart. It's an ill-conceived graph that uses bars to represent individual data values (number of eagles sighted) week by week.



Look carefully. That's not a histogram. A histogram shows *What* we've measured along the horizontal axis and counts of the associated *Who*'s represented as bar heights. This student has it backwards: He used bars to show counts of birds for each week.¹⁰ We need counts of weeks. A correct histogram should have a tall bar at "0" to show there were many weeks when no eagles were seen, like this:

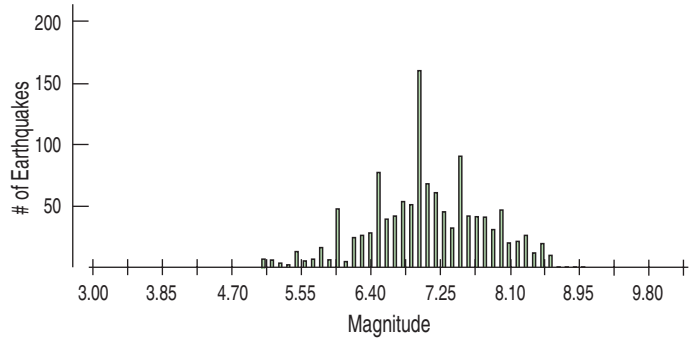
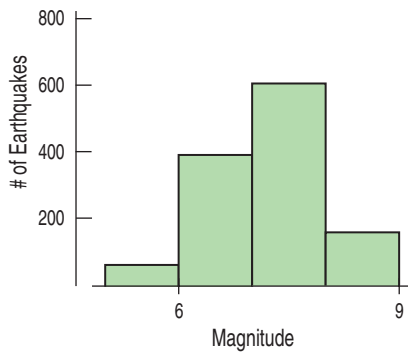
FIGURE 4.16

A histogram of the eagle-sighting data shows the number of weeks in which different counts of eagles occurred. This display shows the distribution of juvenile-eagle sightings.



- **Choose a bin width appropriate to the data.** Computer programs usually do a pretty good job of choosing histogram bin widths. Often there's an easy way to adjust the width, sometimes interactively. Here are the tsunami earthquakes with two (rather extreme) choices for the bin size:

¹⁰ Edward Tufte, in his book *The Visual Display of Quantitative Information*, proposes that graphs should have a high data-to-ink ratio. That is, we shouldn't waste a lot of ink to display a single number when a dot would do the job.



The task of summarizing a quantitative variable is relatively simple, and there is a simple path to follow. However, you need to watch out for certain features of the data that make summarizing them with a number dangerous. Here's some advice:

- ▶ **Don't forget to do a reality check.** Don't let the computer or calculator do your thinking for you. Make sure the calculated summaries make sense. For example, does the mean look like it is in the center of the histogram? Think about the spread: An IQR of 50 mpg would clearly be wrong for gas mileage. And no measure of spread can be negative. The standard deviation can take the value 0, but only in the very unusual case that all the data values equal the same number. If you see an IQR or standard deviation equal to 0, it's probably a sign that something's wrong with the data.
- ▶ **Don't forget to sort the values before finding the median or percentiles.** It seems obvious, but when you work by hand, it's easy to forget to sort the data first before counting in to find medians, quartiles, or other percentiles. Don't report that the median of the five values 194, 5, 1, 17, and 893 is 1 just because 1 is the middle number.
- ▶ **Don't worry about small differences when using different methods.** Finding the 10th percentile or the lower quartile in a data set sounds easy enough. But it turns out that the definitions are not exactly clear. If you compare different statistics packages or calculators, you may find that they give slightly different answers for the same data. These differences, though, are unlikely to be important in interpreting the data, the quartiles, or the IQR, so don't let them worry you.

Gold Card Customers—Regions National Banks

Month	April 2007	May 2007
Average Zip Code	45,034.34	38,743.34

- ▶ **Don't compute numerical summaries of a categorical variable.** Neither the mean zip code nor the standard deviation of social security numbers is meaningful. If the variable is categorical, you should instead report summaries such as percentages of individuals in each category. It is easy to make this mistake when using technology to do the summaries for you. After all, the computer doesn't care what the numbers mean.
- ▶ **Don't report too many decimal places.** Statistical programs and calculators often report a ridiculous number of digits. A general rule for numerical summaries is to report one or two more digits than the number of digits in the data. For example, earlier we saw a dotplot of the Kentucky Derby race times. The mean and standard deviation of those times could be reported as:

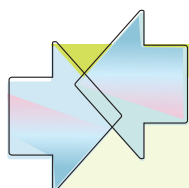
$$\bar{y} = 130.63401639344262 \text{ sec} \quad s = 13.66448201942662 \text{ sec}$$

But we knew the race times only to the nearest quarter second, so the extra digits are meaningless.

- ▶ **Don't round in the middle of a calculation.** Don't report too many decimal places, but it's best not to do any rounding until the end of your calculations. Even though you might report the mean of the earthquakes as 7.08, it's really 7.08339. Use the more precise number in your calculations if you're finding the standard deviation by hand—or be prepared to see small differences in your final result.

(continued)

- ▶ **Watch out for multiple modes.** The summaries of the Kentucky Derby times are meaningless for another reason. As we saw in the dotplot, the Derby was initially a longer race. It would make much more sense to report that the old 1.5 mile Derby had a mean time of 159.6 seconds, while the current Derby has a mean time of 124.6 seconds. If the distribution has multiple modes, consider separating the data into different groups and summarizing each group separately.
- ▶ **Beware of outliers.** The median and IQR are resistant to outliers, but the mean and standard deviation are not. To help spot outliers . . .
- ▶ **Don't forget to: Make a picture (make a picture, make a picture).** The sensitivity of the mean and standard deviation to outliers is one reason you should always make a picture of the data. Summarizing a variable with its mean and standard deviation when you have not looked at a histogram or dotplot to check for outliers or skewness invites disaster. You may find yourself drawing absurd or dangerously wrong conclusions about the data. And, of course, you should demand no less of others. Don't accept a mean and standard deviation blindly without some evidence that the variable they summarize is unimodal, symmetric, and free of outliers.



CONNECTIONS

Distributions of quantitative variables, like those of categorical variables, show the possible values and their relative frequencies. A histogram shows the distribution of values in a quantitative variable with adjacent bars. Don't confuse histograms with bar charts, which display categorical variables. For categorical data, the mode is the category with the biggest count. For quantitative data, modes are peaks in the histogram.

The shape of the distribution of a quantitative variable is an important concept in most of the subsequent chapters. We will be especially interested in distributions that are unimodal and symmetric.

In addition to their shape, we summarize distributions with center and spread, usually pairing a measure of center with a measure of spread: median with IQR and mean with standard deviation. We favor the mean and standard deviation when the shape is unimodal and symmetric, but choose the median and IQR for skewed distributions or when there are outliers we can't otherwise set aside.

WHAT HAVE WE LEARNED?



We've learned how to make a picture of quantitative data to help us see the story the data have to *Tell*.

- ▶ We can display the distribution of quantitative data with a *histogram*, a *stem-and-leaf* display, or a *dotplot*.
- ▶ We *Tell* what we see about the distribution by talking about *shape*, *center*, *spread*, and any *unusual features*.

We've learned how to summarize distributions of quantitative variables numerically.

- ▶ Measures of center for a distribution include the median and the mean.

We write the formula for the mean as $\bar{y} = \frac{\sum y}{n}$.

- ▶ Measures of spread include the range, IQR, and standard deviation.

The standard deviation is computed as $s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$.

The median and IQR are not usually given as formulas.

- ▶ We'll report the median and IQR when the distribution is skewed. If it's symmetric, we'll summarize the distribution with the mean and standard deviation (and possibly the median and IQR as well). Always pair the median with the IQR and the mean with the standard deviation.

We've learned to *Think* about the type of variable we're summarizing.

- ▶ All the methods of this chapter assume that the data are quantitative.
- ▶ The **Quantitative Data Condition** serves as a check that the data are, in fact, quantitative. One good way to be sure is to know the measurement units. You'll want those as part of the *Think* step of your answers.

Terms

Distribution	44. The distribution of a quantitative variable slices up all the possible values of the variable into equal-width bins and gives the number of values (or counts) falling into each bin.
Histogram (relative frequency histogram)	45. A histogram uses adjacent bars to show the distribution of a quantitative variable. Each bar represents the frequency (or relative frequency) of values falling in each bin.
Gap	45. A region of the distribution where there are no values.
Stem-and-leaf display	47. A stem-and-leaf display shows quantitative data values in a way that sketches the distribution of the data. It's best described in detail by example.
Dotplot	49. A dotplot graphs a dot for each case against a single axis.
Shape	49. To describe the shape of a distribution, look for <ul style="list-style-type: none"> ▶ single vs. multiple modes. ▶ symmetry vs. skewness. ▶ outliers and gaps.
Center	52, 58. The place in the distribution of a variable that you'd point to if you wanted to attempt the impossible by summarizing the entire distribution with a single number. Measures of center include the mean and median.
Spread	54, 61. A numerical summary of how tightly the values are clustered around the center. Measures of spread include the IQR and standard deviation.
Mode	49. A hump or local high point in the shape of the distribution of a variable. The apparent location of modes can change as the scale of a histogram is changed.
Unimodal (Bimodal)	50. Having one mode. This is a useful term for describing the shape of a histogram when it's generally mound-shaped. Distributions with two modes are called bimodal . Those with more than two are multimodal .
Uniform	50. A distribution that's roughly flat is said to be uniform.
Symmetric	50. A distribution is symmetric if the two halves on either side of the center look approximately like mirror images of each other.
Tails	50. The tails of a distribution are the parts that typically trail off on either side. Distributions can be characterized as having long tails (if they straggle off for some distance) or short tails (if they don't).
Skewed	50. A distribution is skewed if it's not symmetric and one tail stretches out farther than the other. Distributions are said to be skewed left when the longer tail stretches to the left, and skewed right when it goes to the right.
Outliers	51. Outliers are extreme values that don't appear to belong with the rest of the data. They may be unusual values that deserve further investigation, or they may be just mistakes; there's no obvious way to tell. Don't delete outliers automatically—you have to think about them. Outliers can affect many statistical analyses, so you should always be alert for them.
Median	52. The median is the middle value, with half of the data above and half below it. If n is even, it is the average of the two middle values. It is usually paired with the IQR.
Range	54. The difference between the lowest and highest values in a data set. $Range = max - min$.
Quartile	54. The lower quartile (Q1) is the value with a quarter of the data below it. The upper quartile (Q3) has three quarters of the data below it. The median and quartiles divide data into four parts with equal numbers of data values.

Interquartile range (IQR)	55. The IQR is the difference between the first and third quartiles. $IQR = Q3 - Q1$. It is usually reported along with the median.
Percentile	55. The i th percentile is the number that falls above $i\%$ of the data.
5-Number Summary	56. The 5-number summary of a distribution reports the minimum value, $Q1$, the median, $Q3$, and the maximum value.
Mean	58. The mean is found by summing all the data values and dividing by the count: $\bar{y} = \frac{\text{Total}}{n} = \frac{\sum y}{n}.$ <p>It is usually paired with the standard deviation.</p>
Resistant	59. A calculated summary is said to be resistant if outliers have only a small effect on it.
Variance	61. The variance is the sum of squared deviations from the mean, divided by the count minus 1: $s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}.$ <p>It is useful in calculations later in the book.</p>
Standard deviation	61. The standard deviation is the square root of the variance: $s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$ <p>It is usually reported along with the mean.</p>

Skills

THINK

- ▶ Be able to identify an appropriate display for any quantitative variable.
- ▶ Be able to guess the shape of the distribution of a variable by knowing something about the data.
- ▶ Be able to select a suitable measure of center and a suitable measure of spread for a variable based on information about its distribution.
- ▶ Know the basic properties of the median: The median divides the data into the half of the data values that are below the median and the half that are above.
- ▶ Know the basic properties of the mean: The mean is the point at which the histogram balances.
- ▶ Know that the standard deviation summarizes how spread out all the data are around the mean.
- ▶ Understand that the median and IQR resist the effects of outliers, while the mean and standard deviation do not.
- ▶ Understand that in a skewed distribution, the mean is pulled in the direction of the skewness (toward the longer tail) relative to the median.

SHOW

- ▶ Know how to display the distribution of a quantitative variable with a stem-and-leaf display (drawn by hand for smaller data sets), a dotplot, or a histogram (made by computer for larger data sets).
- ▶ Know how to compute the mean and median of a set of data.
- ▶ Know how to compute the standard deviation and IQR of a set of data.

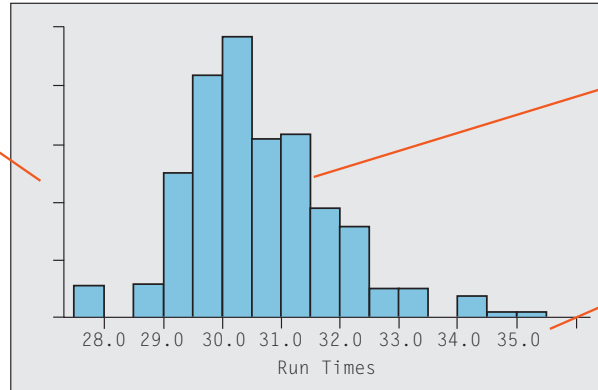
TELL

- ▶ Be able to describe the distribution of a quantitative variable in terms of its shape, center, and spread.
- ▶ Be able to describe any anomalies or extraordinary features revealed by the display of a variable.
- ▶ Know how to describe summary measures in a sentence. In particular, know that the common measures of center and spread have the same units as the variable that they summarize, and should be described in those units.
- ▶ Be able to describe the distribution of a quantitative variable with a description of the shape of the distribution, a numerical measure of center, and a numerical measure of spread. Be sure to note any unusual features, such as outliers, too.

DISPLAYING AND SUMMARIZING QUANTITATIVE VARIABLES ON THE COMPUTER

Almost any program that displays data can make a histogram, but some will do a better job of determining where the bars should start and how they should partition the span of the data.

The vertical scale may be counts or proportions. Sometimes it isn't clear which. But the shape of the histogram is the same either way.



Most packages choose the number of bars for you automatically. Often you can adjust that choice.

The axis should be clearly labeled so you can tell what "pile" each bar represents. You should be able to tell the lower and upper bounds of each bar.

Many statistics packages offer a prepackaged collection of summary measures. The result might look like this:

Variable: W eight
 N = 234
 Mean = 143.3 Median = 139
 St. Dev = 11.1 IQR = 14

Alternatively, a package might make a table for several variables and summary measures:

A S

Case Study: Describing Distribution Shapes. Who's safer in a crash—passengers or the driver? Investigate with your statistics package.

Variable	N	mean	median	stdev	IQR
Weight	234	143.3	139	11.1	14
Height	234	68.3	68.1	4.3	5
Score	234	86	88	9	5

It is usually easy to read the results and identify each computed summary. You should be able to read the summary statistics produced by any computer package.

Packages often provide many more summary statistics than you need. Of course, some of these may not be appropriate when the data are skewed or have outliers. It is your responsibility to check a histogram or stem-and-leaf display and decide which summary statistics to use.

It is common for packages to report summary statistics to many decimal places of "accuracy." Of course, it is rare data that have such accuracy in the original measurements. The ability to calculate to six or seven digits beyond the decimal point doesn't mean that those digits have any meaning. Generally it's a good idea to round these values, allowing perhaps one more digit of precision than was given in the original data.

Displays and summaries of quantitative variables are among the simplest things you can do in most statistics packages.