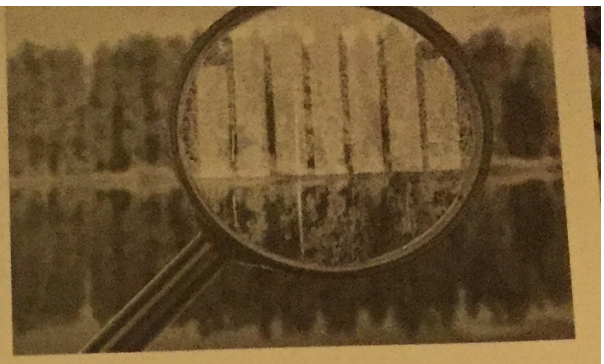
CHAPTER 7



Summarizing and Displaying Measurement Data

Thought Questions

1. If you were to read the results of a study showing that daily use of a certain exercise machine for two months resulted in an average 10-pound weight loss, what more would you want to know about the numbers in addition to the average? (*Hint:* Do you think everyone who used the machine lost 10 pounds?)
2. Suppose you are comparing two job offers, and one of your considerations is the cost of living in each area. You record the price of 50 advertised apartments for each community. What summary measures of the rent values for each community would you need in order to make a useful comparison? For instance, would the lowest rent in the list be enough information?
3. In February 2013, the median sales price of a new home in the United States was $264,900, and the average price was $310,000. How do you think these values are computed? Which do you think is more useful to someone considering the purchase of a home, the median or the average? (*Source:* http://www.investmenttools.com/median_and_average_sales_prices_of_houses_sold_in_the_us.htm, June 3, 2013.)
4. The Stanford-Binet IQ test, 5th edition, is designed to have a mean, or average, of 100 for the entire population. It is also said to have a *standard deviation* of 15. What aspect of the population of IQ scores do you think is described by the "standard deviation"? For instance, do you think it describes something about the average? If not, what might it describe?
5. Students in a statistics class at a large state university were given a survey in which one question asked was age (in years). One student was a retired person, and her age was an "outlier." What do you think is meant by an "outlier"? If the students' heights were measured, would this same retired person necessarily have a value that was an "outlier"? Explain.

7.1 Turning Data into Information

How old is the oldest person you know who is currently alive? That question was posed as part of an insurance company commercial during the 2013 Super Bowl and prompted a statistics student to ask the same question of the members of her class as part of a class project. The 31 responses, in the order collected, were as follows*:

75, 90, 60, 95, 85, 84, 76, 74, 92, 62, 83, 80, 90, 65, 72, 79, 36, 78, 65, 98, 70, 88, 99, 60, 82, 65, 79, 76, 80, 52, 75

Suppose the oldest person you know is your great-grandmother Margaret, who is 86, and you are curious about where she falls relative to the people in this class project. It certainly isn't immediately obvious from the list of ages shown above. In fact, looking at a scrambled list of numbers is about as informative as looking at a scrambled set of letters. To get information out of data, the data have to be organized and summarized.

The first thought that may occur to you is to put the ages into increasing order so you could see where Margaret's age is relative to the ages from this class project. Doing that, you find:

36, 52, 60, 60, 62, 65, 65, 65, 70, 72, 74, 75, 75, 76, 76, 78, 79, 79, 80, 80, 82, 83, 84, 85, 88, 90, 90, 92, 95, 98, 99

Now you can see that Margaret would fall quite a bit above the middle, and you can count to see that there are only seven out of the 31 people in the list who are older than she is. But this list still isn't easy to assimilate into a useful picture. It would help if we could summarize the ages. There are many useful summaries of data, and they are generally categorized into four kinds of information. These are the center (mean or median), unusual values called outliers, the variability, and the shape.

Mean and Median as Measures of Center

The first useful concept is the idea of the "center" or "location" of the data. What's a typical or average value? For the ages just given, the numerical **average,** or **mean,** is 76.3 years. As another measure of "center" consider that there were 31 values in the list of ages, so the **median,** with half of the ages above and half of the ages below it, is 78. There are 15 ages below 78 and 15 ages above it.

To find the mean of a data set, simply add up all of the values and then divide by the number of values. In the age example, the sum of all of the ages is 2365 combined years! But the ages are divided up among 31 people, so the average is 2365/31 = 76.3 years. The median is the middle value after the numbers have been put in order. When the data set has an odd number of values, the median is the one in the middle of the ordered list, as in the age example. There are 31 ages, so the median of 78 has 15 ages above it and 15 ages below it. When the data set has an even number of values, the median is the average of the middle two. Sometimes,

*Hypothetical but realistic data, constructed to illustrate the concepts in this chapter.

there are multiple individuals with the same value, and some of them are tied with the median. So the formal definition is that the median is the value that has half of the ordered list of numbers at or above it and half of the ordered list at or below it.

If one or more values in a data set occur more than once, then the most common one is called the **mode**. The mode is sometimes mentioned as a measure of "center" but in fact it can occur anywhere in the data set, so it doesn't generally represent the center. For the age example, there are many values that occur twice, and the age 65 occurs three times, so the mode is 65. Most of the ages are in fact higher than 65, so the mode is not a very useful representation of "center" in this example. Occasionally, the mode makes sense as a measure of the most "typical" value. For instance, suppose the ages of the students in a kindergarten class were recorded at the start of the school year, and there were three children at each of ages 4 and 6, but 12 children at age 5. Then it would make sense to report that the mode of the ages is 5.

Outliers

You can see that for the oldest ages, the median of 78 is somewhat higher than the mean of 76.3. That's because a very low age, 36, pulled down the mean. It didn't pull down the median because, as long as that very low age was 78 or less, its effect on the median would be the same.

If one or more values are far removed from the rest of the data, they are called **outliers.** There are no hard and fast rules for determining what qualifies as an outlier, but we will learn some guidelines that are often used in identifying them. In this case, most people would agree that the age of 36 is so far removed from the other values that it definitely qualifies as an outlier. The reason for this outlier will be revealed after we look at possible reasons for outliers in general.

Here are three reasons outliers might occur, and what to do about them:

1. The outlier is a legitimate data value and represents natural variation in responses. In that case, the outlier should be retained and included in data summaries.
2. A mistake was made when recording the measurement or a question was misunderstood. In that case, if the correct value can found, replace the outlier. Otherwise, delete it from the dataset before computing numerical summaries. (But always report doing so.)
3. The individual(s) in question belong(s) to a different group than the rest of the individuals. In that case, outliers can be excluded if data summaries are desired for the majority group only. Otherwise, they should be retained.

Which of these three reasons is responsible for the outlier in the oldest ages? It turns out that the student who gave the response of 36 years misunderstood the question. He thought he was supposed to give the age of the oldest person whose *age* he actually

knew, and that was his father, who was 36 years old. He knew older people, but did not know their exact ages. Whether or not to discard that outlier depends on the question we want to answer. If the question is "What are the oldest ages of the people students know?", then the outlier should be discarded (Reason 2, question was misunderstood). If the question is "What ages would students report when asked this question?", then the value of 36 years is a real response, and should not be discarded (Reason 1). But its value and the reason for it should be noted in any narrative summary of the data.

Variability

The third kind of useful information contained in a set of data is the **variability.** How spread out are the values? Are they all close together? Are most of them together, but a few are outliers? Knowing that the mean is about 76, you might wonder if your great grandmother Margaret's age of 86 is unusually high. It would obviously have a different meaning if the reported ages ranged from 72 to 80 than if they ranged from 50 to 100.

The idea of *natural variability,* introduced in Chapter 3, is particularly important when summarizing a set of measurements. Much of our work in statistics involves comparing an observed difference to what we should expect if the difference is due solely to natural variability. For instance, to determine if global warming is occurring, we need to know how much the temperatures in a given area naturally vary from year to year. To determine if a one-year-old child is growing abnormally slowly, we need to know how much heights of one-year-old children naturally vary.

Minimum, Maximum, and Range

The simplest measure of variability is to find the minimum value and the maximum value and to compute the **range,** which is just the difference between them. In the case of the oldest ages, the reported ages went from 36 to 99, for a range of 63 years. Without the outlier, they covered a 47 year range, from 52 to 99. Margaret's age is not so surprising given that range.

Temperatures over the years on a given date in a certain location may range from a record low of 59 degrees Fahrenheit to a record high of 90 degrees, a 31-degree range. We introduce two more measures of variability, the interquartile range and the standard deviation, later in this chapter.

Shape

The fourth kind of useful information is the **shape,** which can be derived from a certain kind of picture of the data. We can answer questions such as: Are most of the values clumped in the middle with values tailing off at each end? Are there two distinct groupings, with a gap between them? Are most of the values clumped together at one end with a few very high or low values? Even knowing that ages ranged from 36 to 99, Margaret's age of 86 would have a different meaning if the ages were clumped mostly in the 60s and the 90s, for instance, than if they were spread out evenly across that range.

7.2 Picturing Data: Stemplots and Histograms

About Stemplots

A **stemplot** is a quick and easy way to put a list of numbers into order while getting a picture of their shape. The easiest way to describe a stemplot is to construct one. Let's first use the ages we've been discussing, then we will turn to some real data, where each number has an identity. Before reading any further, look at the rightmost part of Figure 7.1 so you can see what a completed stemplot looks like. Each of the digits extending to the right represents one data point. The first thing you see is 3|6. That represents the lowest reported age of 36. Each of the digits on the right represents one reported age. For instance, see if you can locate the age of the oldest person, 99. It's the last value to the right of the "stem" value of 9|.

Creating a Stemplot

Stemplots are sometimes called **stem-and-leaf plots** or **stem-and-leaf diagrams.** Only two steps are needed to create a stemplot—creating the stem and attaching the leaves.

Step 1: Create the stems.

The first step is to divide the range of the data into equal units to be used on the **stem.** The goal is to have approximately 6 to 15 stem values, representing equally spaced intervals. In the example shown in Figure 7.1, each of the seven stem values represents a range of 10 years of age. For instance, any age in the 80s, from 80 to 89, would be placed after the 8| on the stem.

Step 2: Attach the leaves.

The second step is to attach a **leaf** to represent each data point. The next digit in the number is used as the leaf, and if there are any remaining digits they are simply dropped. Let's use the unordered list of ages first displayed:

75, 90, 60, 95, 85, 84, 76, 74, 92, 62, 83, 80, 90, 65, 72, 79, 36, 78, 65, 98, 70, 88, 99, 60, 82, 65, 79, 76, 80, 52, 75

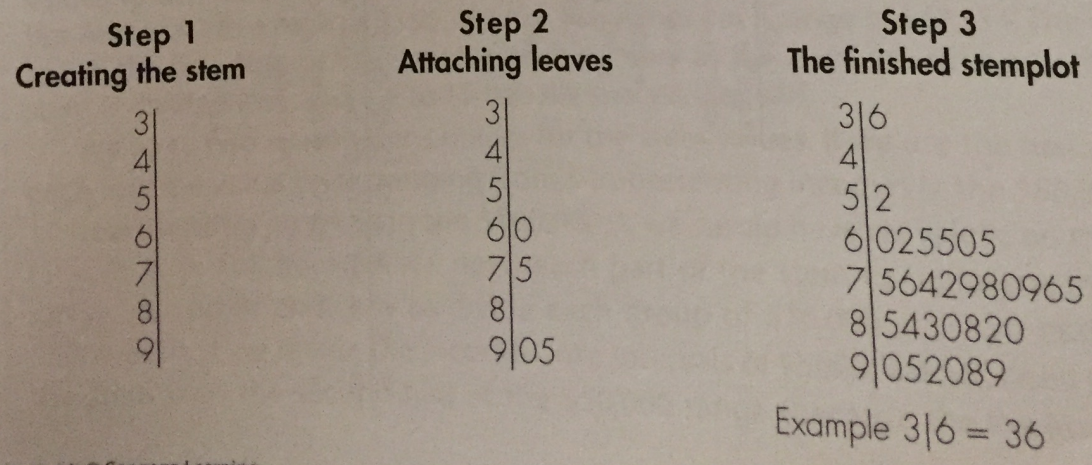


Figure 7.1
Building a stemplot of oldest ages

The middle part of Figure 7.1 shows the picture after leaves have been attached for the first four ages, 75, 90, 60, and 95. The finished picture, on the right, has the leaves attached for all 31 ages. Sometimes an additional step is taken and the leaves are ordered numerically on each branch.

Further Details for Creating Stemplots

Suppose you wanted to create a picture of what your own pulse rate is when you are relaxed. You collect 25 values over a few days and find that they range from 54 to 78. If you tried to create a stemplot using the first digit as the stem, you would have only three stem values (5, 6 and 7). If you tried to use both digits for the stem, you could have as many as 25 separate values, and the picture would be meaningless.

The solution to this problem is to reuse each of the digits 5, 6, and 7 in the stem. Because you need to have equally spaced intervals, you could use each of the digits two or five times. If you use them each twice, the first listed would receive leaves from 0 to 4, and the second would receive leaves from 5 to 9. Thus, each stem value would encompass a range of five beats per minute of pulse. If you use each digit five times, each stem value would receive leaves of two possible values. The first stem for each digit would receive leaves of 0 and 1, the second would receive leaves of 2 and 3, and so on. Notice that if you tried to use the initial pulse digits three or four times each, you could not evenly divide the leaves among them because there are always 10 possible values for leaves. Figure 7.2 shows two possible stemplots for the same hypothetical pulse data. Stemplot *A* shows the digits 5, 6, and 7 used twice; stemplot *B* shows them used five times. (The first two 5's are not needed and not shown.)

EXAMPLE 7.1 **Stemplot of Median Income for Families of Four**

Table 7.1 lists the estimated median income for a four-person family in 2014 for each of the 50 states and the District of Columbia, information released by the U.S. government in May 2013 for use in setting aid levels in the Low Income Home

Figure 7.2
Two stemplots for the same pulse rate data

Stemplot *A*

```
5|4
5|7 8 9
6|0 2 3 3 4 4
6|5 5 5 6 7 7 8 9
7|0 0 1 2 4
7|5 8
```

Stemplot *B*

```
5|4
5|7
5|8 9
6|0
6|2 3 3
6|4 4 5 5 5
6|6 7 7
6|8 9
7|0 0 1
7|2
7|4 5
7|
7|8
```

TABLE 7.1 Estimated 2014 Median Income for a Family of Four

Alabama	$64,899	Montana	$68,905
Alaska	$87,726	Nebraska	$74,484
Arizona	$64,434	Nevada	$69,475
Arkansas	$56,994	New Hampshire	$94,838
California	$77,679	New Jersey	$103,852
Colorado	$84,431	New Mexico	$57,353
Connecticut	$103,173	New York	$83,648
Delaware	$83,557	North Carolina	$66,985
District of Columbia	$87,902	North Dakota	$82,605
Florida	$65,406	Ohio	$73,924
Georgia	$67,401	Oklahoma	$63,580
Hawaii	$85,350	Oregon	$69,573
Idaho	$61,724	Pennsylvania	$80,937
Illinois	$81,770	Rhode Island	$87,793
Indiana	$70,504	South Carolina	$62,965
Iowa	$76,905	South Dakota	$71,207
Kansas	$74,073	Tennessee	$64,042
Kentucky	$65,968	Texas	$66,880
Louisiana	$68,964	Utah	$68,017
Maine	$74,481	Vermont	$81,408
Maryland	$105,348	Virginia	$90,109
Massachusetts	$102,773	Washington	$83,238
Michigan	$73,354	West Virginia	$63,863
Minnesota	$87,283	Wisconsin	$79,141
Mississippi	$57,662	Wyoming	$76,868
Missouri	$70,896		

Source: Federal Registry, May 15, 2013, http://www.gpo.gov/fdsys/pkg/FR-2013-05-15/html/2013-11575.htm.

Energy Assistance Program. Scanning the list gives us some information, but it would be easier to get the big picture if it were in some sort of numerical order. We could simply list the states by value instead of alphabetically, but that would not give us a picture of the location, variability, shape, and possible outliers.

Let's create a stemplot for the 51 income levels. The first step is to decide what values to use for the stem. The median family incomes range from a low of $56,994 (for Arkansas) to a high of $105,348 (for Maryland), for a range of $48,354. The goal is to use the first digit or few digits in each number as the stem, in such a way that the stem is divided into about 6 to 15 equally spaced intervals.

We have two reasonable choices for the stem values. If we use the first digit in each income value once, ranging from 5 (representing incomes in the $50,000s) to 10 (representing incomes in the $100,000s), we would have six values on the stem (5, 6, 7, 8, 9, 10). Because we need each part of the stem to represent the same range, our other choice is to divide each group of $10,000 into two intervals of $5000 each. If we divide the incomes into intervals of $5000, we will need to begin the stem with the second half of the $50,000 range (because the lowest value is

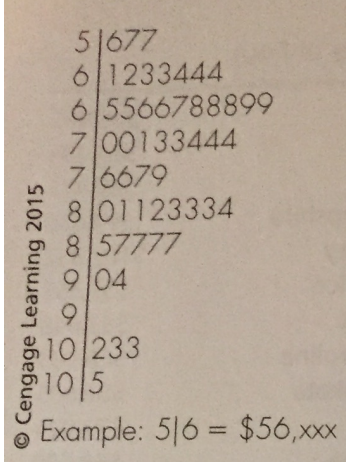


Figure 7.3
Stemplot of median incomes for families of four

$56,994) and end it with the second half of the $100,000 range, resulting in stem values of 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10 for a total of 11 stem values. For a dataset as large as this one, it is better to spread the values out somewhat, so we will divide incomes into intervals of $5000.

Figure 7.3 shows the completed stemplot. Notice that the leaves have been put in order. Notice also that the income values have been *truncated* instead of *rounded*. To truncate a number, simply drop off the unused digits. Thus, the lowest income of $56,994 (for Arkansas) is truncated to $56,000 instead of rounded to $57,000. Rounding could be used instead. ■

Obtaining Information from the Stemplot

Stemplots help us determine the "shape" of a data set, identify outliers, and locate the center. For instance, the pulse rates in Figure 7.2 have a "bell shape" in which they are centered in the mid-60s and tail off in both directions from there. There are no outliers. The stemplot of ages in Figure 7.1 clearly illustrates the outlier of 36. Aside from that and the age of 52, they are somewhat *uniformly* distributed in the 60s, 70s, 80s, and 90s.

From the stemplot of median income data in Figure 7.3, we can make several observations. First, there is a wide range of values, with the median income in Maryland, the highest, being close to twice that of Arkansas, the lowest. Second, there appear to be four states with unusually high median family incomes, all over $100,000. From Table 7.1, we can see that these are Massachusetts, Connecticut, New Jersey, and Maryland. Then there is a gap before reaching New Hampshire, at $94,838. Other than the four values over $100,000, the incomes tend to be almost "bell-shaped" with a center around the mid $70,000s. There are no obvious outliers.

If we were interested in what factors determine income levels, we could use this information from the stemplot to help us. We would pursue questions like "What is different about the four high-income states?" We might notice that much of their population works in high-income cities. Many New York City employees live in Connecticut and New Jersey, and Washington, D.C. employees live in Maryland. Much of the population of Massachusetts lives and works in the Boston area.

Creating a Histogram

Histograms are pictures related to stemplots. For very large data sets, a histogram is more feasible than a stemplot because it doesn't list every data value.

To create a histogram, divide the range of the data into intervals in much the same way as we did when creating a stemplot. But instead of listing each individual value, simply count how many values fall into each part of the range. Draw a bar with height equal to the count for each part of the range. Or, equivalently, make the height equal to the *proportion* of the total count that falls in that interval.

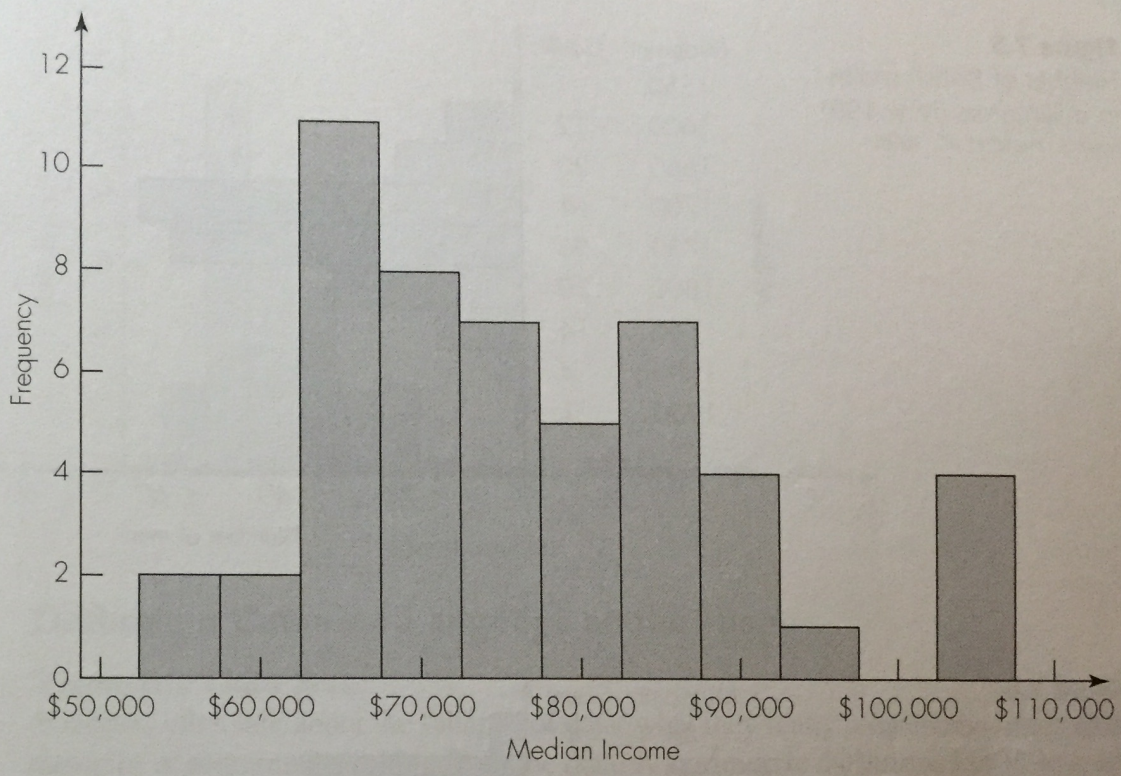


Figure 7.4
Histogram of estimated 2014 median income for families of four, for states in the United States

Figure 7.4 shows a histogram for the income data from Table 7.1. Each bar represents an interval of $5000, with the first one centered at $55,000 (ranging from $52,500 to $57,500). Notice that the heights of the bars are represented as frequencies. For example, there are four values in the highest median income range, centered on $105,000. If this histogram had used the proportion in each category on the vertical axis instead, then the height of the bar centered on $105,000 would be 4/51, or about 0.08. In that case, the heights of all of the bars must sum to 1, or 100%. If you wanted to know what proportion of the data fell into a certain interval or range, you would simply sum the heights of the bars for that range. Also, notice that if you were to turn the histogram on its side, it would look very much like a stemplot except that the labels would differ slightly.

EXAMPLE 7.2 **Heights of British Males**

Figure 7.5 (next page) displays a histogram of the heights, in millimeters, of 199 randomly selected British men. (Marsh, 1988, p. 315; data reproduced in Hand et al., 1994, pp. 179–183). The histogram is rotated sideways from the one in Figure 7.4. Some computer programs display histograms with this orientation. Notice that the heights create a "bell shape" with a center in the mid-1700s (millimeters). There are no outliers. ■

EXAMPLE 7.3 **The Old Faithful Geyser**

Figure 7.6 (next page) shows a histogram of the times between eruptions of the "Old Faithful" geyser. Notice that the picture appears to have two clusters of values, with one centered around 50 minutes and another, larger cluster centered around 80 minutes. A picture like this may help scientists figure out what causes the geyser to erupt when it does. ■

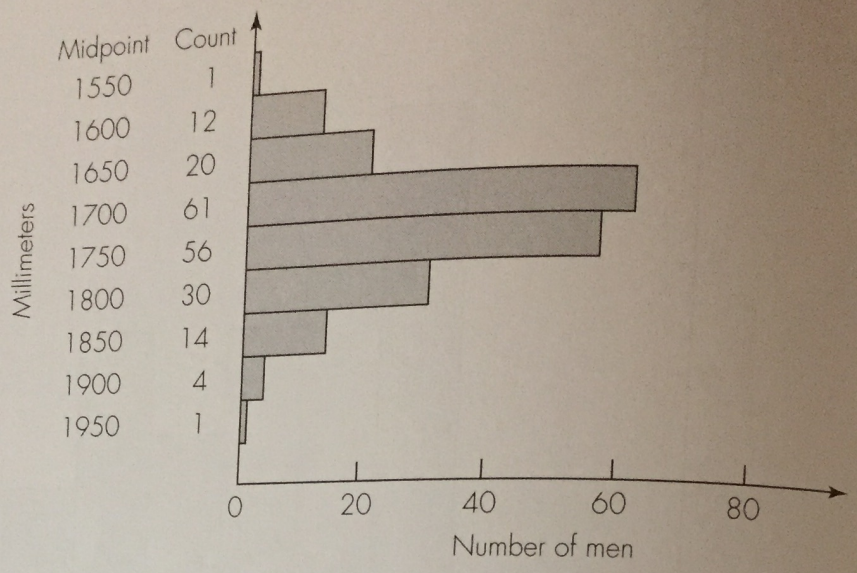


Figure 7.5
Heights of British males in millimeters ($N = 199$)
Source: Hand et al., 1994.

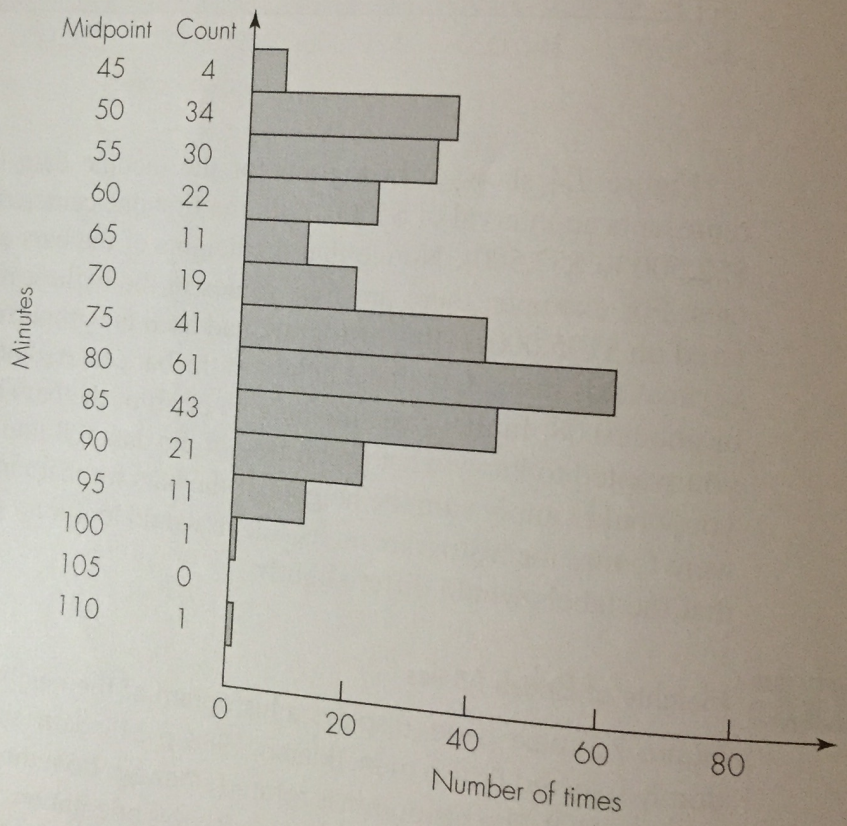


Figure 7.6
Times between eruptions of "Old Faithful" geyser ($N = 299$)
Source: Hand et al., 1994.

EXAMPLE 7.4 **How Much Do Students Exercise?**

Students in an introductory statistics class were asked a number of questions on the first day of class. Figure 7.7 shows a histogram of 172 responses to the question "How many hours do you exercise per week (to the nearest half hour)?" Notice that the bulk of the responses are in the range from 0 to 10 hours, with a mode of 2 hours. But there are responses trailing out to a maximum of 30 hours a week, with five responses at or above 20 hours a week.

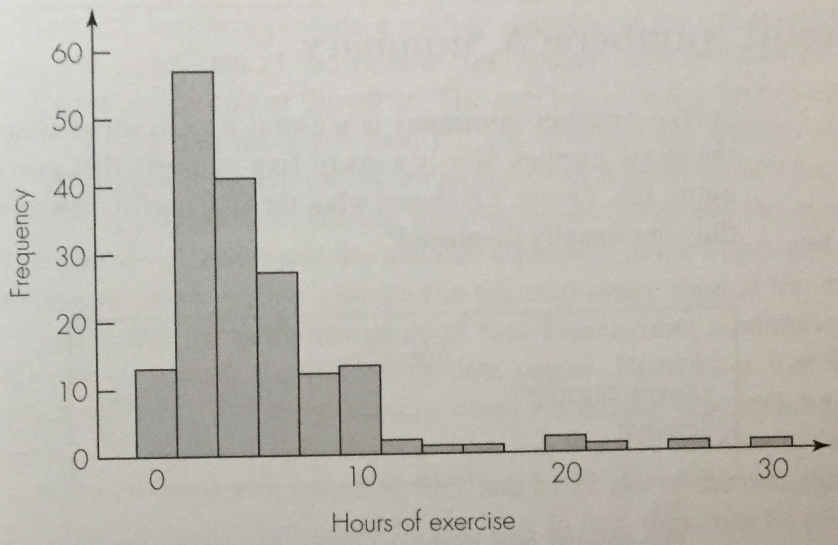


Figure 7.7
Self-reported hours of exercise for 172 college students
Source: The author's students.

Defining a Common Language about Shape

Symmetric Data Sets

Scientists often talk about the "shape" of data; what they really mean is the shape of the stemplot or histogram resulting from the data. A **symmetric** data set is one in which, if you were to draw a line through the center, the picture on one side would be a mirror image of the picture on the other side. A special case, which will be discussed in detail in Chapter 8, is a **bell-shaped** data set, in which the picture is not only symmetric but also shaped like a bell. The stemplots in Figure 7.2, displaying pulse rates, and the histogram in Figure 7.5, displaying male heights, are approximately symmetric and bell-shaped.

Unimodal or Bimodal

Recall that the mode is the most common value in a set of data. If there is a single prominent peak in a histogram or stemplot, as in Figures 7.2 and 7.5, the shape is called **unimodal,** meaning "one mode." If there are two prominent peaks, the shape is called **bimodal,** meaning "two modes." Figure 7.6, displaying the times between eruptions of the Old Faithful geyser, is bimodal. There is one peak around 50 minutes and a higher peak around 80 minutes.

Skewed Data Sets

In common language, something that is skewed is off-center in some way. In statistics, a **skewed** data set is one that is basically unimodal but is substantially off from being bell-shaped. If it is **skewed to the right,** the higher values are more spread out than the lower values. Figure 7.7, displaying hours of exercise per week for college students, is an example of data skewed to the right. If a data set is **skewed to the left,** then the lower values are more spread out and the higher ones tend to be clumped. This terminology results from the fact that before computers were used, shape pictures were always hand drawn using the horizontal orientation in Figure 7.4. Notice that a picture that is skewed to the right, like Figure 7.7, extends further to the right of the highest peak (the tallest bar) than to the left. Most students think the terminology should be the other way around, so be careful to learn this definition! The direction of the "skew" is the direction with the unusual values, and *not* the direction with the bulk of the data.

7.3 Five Useful Numbers: A Summary

A **five-number summary** is a useful way to summarize a long list of numbers. As the name implies, this is a set of five numbers that provide a good summary of the entire list. Figure 7.8 shows what the five useful numbers are and the order in which they are usually displayed.

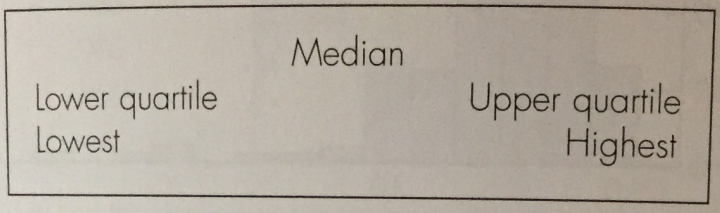


Figure 7.8
The five-number summary display

The lowest and highest values are self-explanatory. The median, which we discussed earlier, is the number such that half of the values are at or above it and half are at or below it. If there is an odd number of values in the data set, the median is simply the middle value in the ordered list. If there is an even number of values, the median is the average of the middle two values. For example, the median of the list 70, 75, 85, 86, 87 is 85 because it is the middle value. If the list had an additional value of 90 in it, the median would be 85.5, the average of the middle two numbers, 85 and 86. Make sure you find the middle of the *ordered* list of values.

The median can be found quickly from a stemplot, especially if the leaves have been ordered. Using Figure 7.3, convince yourself that the median of the family income data is the 26th value ($51 = 25 + 1 + 25$) from either end, which is the lowest of the $74,000 values. Consulting Table 7.1, we can see that the actual value is $74,073, the value for Kansas.

The **quartiles** are simply the medians of the two halves of the ordered list of numbers. The **lower quartile**—because it's halfway into the first half—is one quarter of the way from the bottom of the ordered list. Similarly, the **upper quartile** is one quarter of the way down from the top of the ordered list. Complicated algorithms exist for finding exact quartiles. We can get close enough by simply finding the median of the data first, then finding the medians of all the numbers below it and all the numbers above it. For the family income data, the lower quartile is the median of the 25 values below the median of $74,073. Notice that this would be the 13th value from the bottom because $25 = 12 + 1 + 12$. Counting from the low end of the stemplot in Figure 7.3, the 13th value is the first occurrence of $66,000. Consulting Table 7.1, the value is $66,880 (Texas). The upper quartile is the median of the upper 25 values, which is the highest of the values in the $83,000s. Consulting Table 7.1, we see that it is $83,648 (New York). This tells us that three-fourths of the states have median family incomes at or below that for New York, which is $83,648.

The five-number summary for the family income data is thus:

$74,073
$66,880 $83,648
$56,994 $105,348

These five numbers provide a useful summary of the entire set of 51 numbers. We can get some idea of the middle, the spread, and whether or not the values are clumped at one end or the other. The gap between the first quartile and the median ($7193) is somewhat lower than the gap between the median and the third quartile ($9575), indicating that the values in the lower half are somewhat closer together than those in the upper half. The gap between the extremes and the quartiles are larger than between the quartiles and the median, especially at the upper end, indicating that the values are more tightly clumped in the mid-range than at the ends.

Note that, in using stemplots to find five-number summaries, we won't always be able to consult the full set of data values. Remember that we dropped the last three digits on the family incomes when we created the stemplot. If we had used the stemplot only, the family income values in the five-number summary (in thousands) would have been $56, $66, $74, $83, and $105. All of the conclusions we made in the previous paragraph would still be obvious. In fact, they may be more obvious, because the arithmetic to find the gaps would be much simpler. Truncated values from the stemplot are generally close enough to give us the picture we need.

7.4 Boxplots

A visually appealing and useful way to present a five-number summary is through a **boxplot,** sometimes called a **box and whisker plot.** This simple picture also allows easy comparison of the center and spread of data collected for two or more groups.

EXAMPLE 7.5 How Much Do Statistics Students Sleep?

During the spring semester, 190 students in a statistics class at a large university were asked to answer a series of questions in class one day, including how many hours they had slept the night before (a Tuesday night). A five-number summary for the reported number of hours of sleep is

	7	
6		8
3		16

Two individuals reported that they slept 16 hours; the maximum for the remaining 188 students was 12 hours. A boxplot for the sleep data is shown in Figure 7.9. ■

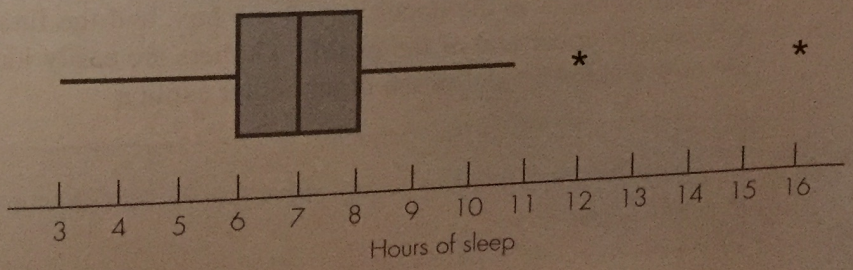


Figure 7.9
Boxplot for hours of sleep

Creating a Boxplot

The boxplot for the hours of sleep is presented in Figure 7.9 and illustrates how a boxplot is constructed. Here are the steps:

1. Draw a horizontal or vertical line, and label it with values from the lowest to the highest values in the data. For the example in Figure 7.9, a horizontal line is used and the labeled values range from 3 to 16 hours.
2. Draw a rectangle, or box, with the ends of the box at the lower and upper quartiles. In Figure 7.9, the ends of the box are at 6 and 8 hours.
3. Draw a line in the box at the value of the median. In Figure 7.9, the median is at 7 hours.
4. Compute the width of the box. This distance is called the **interquartile range** because it's the distance between the lower and upper quartiles. It's abbreviated as "IQR." For the sleep data, the IQR is 2 hours.
5. Compute 1.5 times the IQR. For the sleep data, this is $1.5 \times 2 = 3$ hours. Define an *outlier* to be any value that is more than this distance from the closest end of the box. For the sleep data, the ends of the box are 6 and 8, so any value below $(6 - 3) = 3$, or above $(8 + 3) = 11$, is an outlier.
6. Draw a line or "whisker" at each end of the box that extends from the ends of the box to the farthest data value that isn't an outlier. If there are no outliers, these will be the minimum and maximum values. In Figure 7.9, the whisker on the left extends to the minimum value of 3 hours but the whisker on the right stops at 11 hours.
7. Draw asterisks to indicate data values that are beyond the whiskers and are thus considered to be outliers. In Figure 7.9, we see that there are two outliers, at 12 hours and 16 hours.

If all you have is the information contained in a five-number summary, you can draw a **skeletal boxplot** instead. The only change is that the whiskers don't stop until they reach the minimum and maximum, and thus outliers are not specifically identified. You can still determine if there are any outliers at each end by noting whether the whiskers extend more than $1.5 \times$ IQR. If so, you know that the minimum or maximum value is an outlier, but you don't know if there are any other, less extreme outliers.

Interpreting Boxplots

Boxplots essentially divide the data into fourths. The lowest fourth of the data values is contained in the range of values below the start of the box, the next fourth is contained in the first part of the box (between the lower quartile and the median), the next fourth is in the upper part of the box, and the final fourth is between the box and the upper end of the picture. Outliers are easily identified. Notice that we are now making the definition of an outlier explicit.

> An **outlier** is defined to be any value that is more than $1.5 \times$ IQR beyond the closest quartile.

In the boxplot in Figure 7.9, we can see that one-fourth of the students slept between 3 and 6 hours the previous night, one-fourth slept between 6 and 7 hours, one-fourth slept between 7 and 8 hours, and the final fourth slept between 8 and 16 hours. We can thus immediately see that the data are skewed to the right because the final fourth covers an 8-hour period, whereas the lowest fourth covers only a 3-hour period.

As the next example illustrates, boxplots are particularly useful for comparing two or more groups on the same measurement. Although almost the same information is contained in five-number summaries, the visual display makes similarities and differences much more obvious.

EXAMPLE 7.6 **Who Are Those Crazy Drivers?**

The survey taken in the statistics class in Example 7.5 also included the question "What's the fastest you have ever driven a car? _____ mph." The boxplots in Figure 7.10 illustrate the comparison of the responses for males and females. Here are the corresponding five-number summaries. (There are only 189 students because one didn't answer this question.)

Males (87 Students)		Females (102 Students)	
110		89	
95	120	80	95
55	150	30	130

Some features are more immediately obvious in the boxplots than in the five-number summaries. For instance, the lower quartile for the men is equal to the upper quartile for the women. In other words, 75% of the men have driven 95 mph or faster, but only 25% of the women have done so. Except for a few outliers (120 and 130), all of the women's maximum driving speeds are close to or below the median for the men. Notice how useful the boxplots are for comparing the maximum driving speeds for the sexes. ■

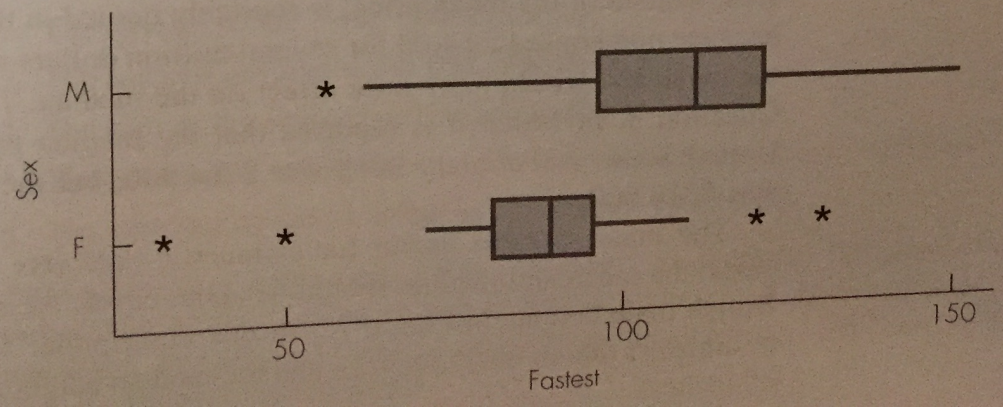


Figure 7.10
Boxplots for fastest ever driven a car

7.5 Traditional Measures: Mean, Variance, and Standard Deviation

The five-number summary has come into use relatively recently. Traditionally, only two numbers have been used to describe a set of numbers: the **mean,** representing the center, and the **standard deviation,** representing the spread or variability in the values. Sometimes the **variance** is given instead of the standard deviation. The standard deviation is simply the square root of the variance, so once you have one you can easily compute the other.

The mean and standard deviation are most useful for symmetric sets of data with no outliers. However, they are very commonly quoted, so it is important to understand what they represent, including their uses and their limitations.

The Mean and When to Use It

As we discussed earlier, the **mean** is the numerical average of a set of numbers. In other words, we add up the values and divide by the number of values. The mean can be distorted by one or more outliers and is thus most useful when there are no extreme values in the data. For example, suppose you are a student taking four classes, and the number of students in each is, respectively, 20, 25, 35, and 200. What is your typical class size? The median is 30 students. The mean, however, is 280/4 or 70 students. The mean is severely affected by the one large class size of 200 students.

As another example, refer to Figure 7.7 from Example 7.4, which displays hours per week students reportedly exercise. The majority of students exercised 10 hours or less, and the median is only 3 hours. But because there were a few very high values, the mean amount is 4.5 hours a week. It would be misleading to say that students exercise an average of 4.5 hours a week. In this case, the median is a better measure of the center of the data.

Data involving incomes or prices of things like houses and cars often are skewed to the right with some large outliers. They are unlikely to have extreme outliers at the lower end because monetary values can't go below 0. Because the mean can be distorted by the high outliers, data involving incomes or prices are usually summarized using the median. For example, the median price of a house in a given area, instead of the mean price, is routinely quoted in the economic news. That's because one house that sold for several million dollars would substantially distort the mean but would have little effect on the median. This is evident in Thought Question 3, in which it is reported that the median price of new homes in the United States in February 2013 was $264,900, but the average price, the mean, was $310,000.

The mean is most useful for symmetric data sets with no outliers. In such cases, the mean and median should be about equal. As an example, notice that the British male heights in Figure 7.5 fit that description. The mean height is 1732.5 millimeters (about 68.25 inches), and the median height is 1725 millimeters (about 68 inches).

The Standard Deviation and Variance

It is not easy to compute the **standard deviation** of a set of numbers, but most calculators and computer programs such as Excel now handle that task for you. For example, in Excel if the data are listed in rows 1 to 20 of Column A, type "=STDEV.S(A1:A20)" into any cell and the standard deviation will be shown. It is more important to know how to interpret the standard deviation, which is a useful measure of how spread out the numbers are. Consider the following two sets of numbers, both with a mean of 100:

Numbers	Mean	Standard Deviation
100, 100, 100, 100, 100	100	0
90, 90, 100, 110, 110	100	10

The first set of numbers has no spread or variability to it at all. It has a standard deviation of 0. The second set has some spread to it; on average, the numbers are about 10 points away from the mean, except for the number that is exactly at the mean. That set has a standard deviation of 10.

Computing the Standard Deviation

Here are the steps necessary to compute the standard deviation:

1. Find the mean.
2. Find the deviation of each value from the mean: value − mean.
3. Square the deviations.
4. Sum the squared deviations.
5. Divide the sum by (the number of values) − 1, resulting in the variance.
6. Take the square root of the variance. The result is the standard deviation.

Let's try this for the set of values 90, 90, 100, 110, 110.

1. The mean is 100.
2. The deviations are −10, −10, 0, 10, 10.
3. The squared deviations are 100, 100, 0, 100, 100.
4. The sum of the squared deviations is 400.
5. The (number of values) − 1 = 5 − 1 = 4, so the variance is 400/4 = 100.
6. The standard deviation is the square root of 100, or 10.

Although it may seem more logical in step 5 to divide by the number of values, rather than by the number of values minus 1, there is a technical reason for subtracting 1. The reason is beyond the level of this discussion but concerns statistical bias, as discussed in Chapter 3.

The easiest interpretation is to recognize that *the standard deviation is roughly the average distance of the observed values from their mean.* Where the data have a bell shape, the standard deviation is quite useful indeed. For example, the Stanford-Binet IQ test (5th edition) is designed to have a mean of 100 and a standard deviation of 15. If we were to produce a histogram of IQs for a large group representative of the whole population, we would find it to be approximately bell-shaped. Its center would be at 100. If we were to determine how far each person's IQ fell from 100, we would find an average distance, on one side or the other, of about 15 points. (In the next chapter, we will see how to use the standard deviation of 15 in a more useful way.) For shapes other than bell shapes, the standard deviation is useful as an intermediate tool for more advanced statistical procedures; it is not very useful on its own, however.

EXAMPLE 7.7 **Putting it All Together with Women's Heights**

Let's look at how to combine the information learned in this chapter into one coherent story. Women in a college statistics class were asked to report various measurements, including height, for which 94 women responded. Most of them reported height to the nearest inch, but a few reported it to the nearest half inch.

We probably can consider these women to be representative of all college women for this measurement. What can we learn about college women's heights from these 94 individuals? Figure 7.11 illustrates a boxplot of these measurements, and the five-number summary used to construct it. From these, we learn the following:

- The heights ranged from 59 inches (4 feet, 11 inches) to 70 inches (5 feet, 10 inches).
- The median of 64 means that half of the women reported heights of 64 inches or more, and half reported 64 inches or less.
- The lowest one-fourth of the women ranged from 59 to 63 inches, the next one-fourth from 63 to 64 inches, the next one-fourth from 64 to 66 inches, and the final one-fourth from 66 to 70 inches. Therefore, heights are more spread out in the extremes than in the middle and are slightly more spread out in the upper half than in the lower half.
- There are no outliers.

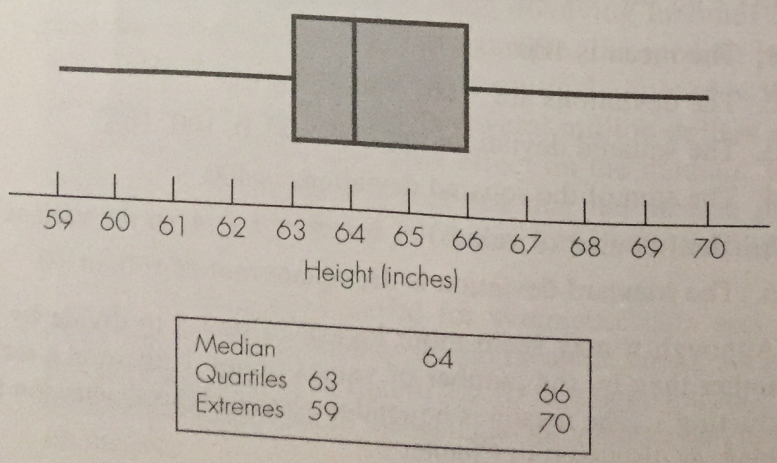


Figure 7.11 Boxplot of heights of 94 college women, and five-number summary used to create it

Figure 7.12
Stemplot and histogram for heights of 94 college women for Example 7.7

Stem–and–leaf of height N = 94
Leaf Unit = 1.0

```
5 | 9
6 | 000001111111
6 | 22222333333333333333
6 | 44444444444444445555555555555
6 | 666666666666777777777
6 | 88889999
7 | 000
```

Example: 5 | 9 = 59 inches

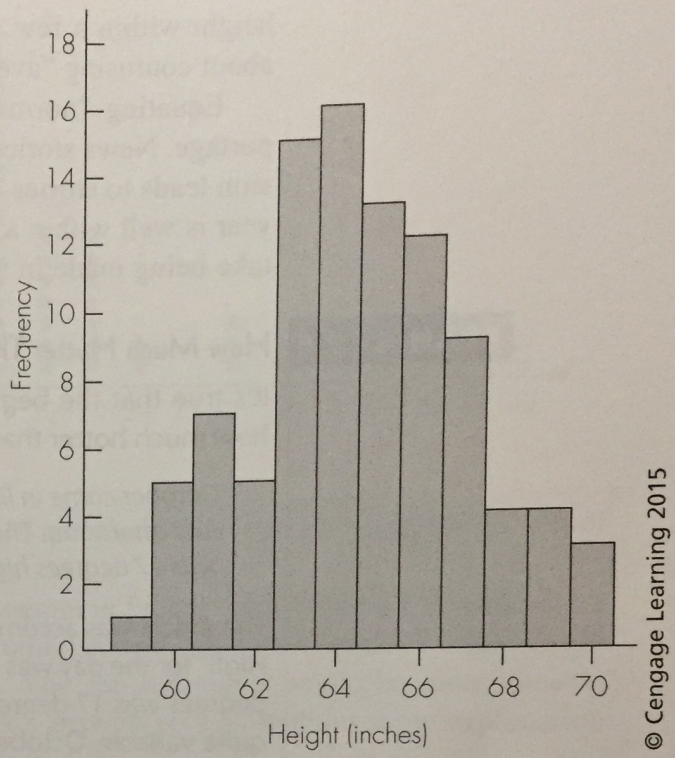


Figure 7.11 gives information about location, spread, and outliers but does not tell us much about the shape. For that, we need to use a stemplot or a histogram. Both of these (created by the Minitab Statistical Software program) are shown in Figure 7.12. Note that the stemplot uses seven intervals, allocating 2 inches to each interval, while the histogram uses 12 intervals, one for each inch of height. Both pictures show that the pattern is bell-shaped, although the histogram obscures the smooth shape somewhat because of the level of detail.

The only measures that aren't directly obvious from either Figure 7.11 or Figure 7.12 are the mean and standard deviation. Those can be computed from the original 94 observations. The mean is 64.455 inches, and the standard deviation is 2.51 inches. We will round these off to 64.5 and 2.5 inches. In Chapter 8, we will revisit this example and learn that these two measures are very useful for bell-shaped distributions. ■

7.6 Caution: Being Average Isn't Normal

By now, you should realize that it takes more than just an average value to describe a set of measurements. Yet, it is a common mistake to confuse "average" with "normal." For instance, if a young boy is tall for his age, people might say something like "He's taller than normal for a three-year-old." In fact, what they mean is that he's taller than the *average* height of three-year-old boys. There is quite a range of possible heights for three-year-old boys, and as we will learn in Chapter 8, any

height within a few standard deviations of the mean is quite "normal." Be careful about confusing "average" and "normal" in your everyday speech.

Equating "normal" with average is particularly common in weather data reportage. News stories often confuse these. When reporting rainfall data, this confusion leads to stories about drought and flood years when in fact the rainfall for the year is well within a "normal" range. If you pay attention, you will notice this mistake being made in almost all news reports about the weather.

EXAMPLE 7.8 **How Much Hotter Than Normal Is Normal?**

It's true that the beginning of October, 2001 was hot in Sacramento, California. But how much hotter than "normal" was it? According to the *Sacramento Bee:*

> *October came in like a dragon Monday, hitting 101 degrees in Sacramento by late afternoon. That temperature tied the record high for Oct. 1 set in 1980—and was 17 degrees higher than normal for the date.* (Korber, 2001)

The article was accompanied by a drawing of a thermometer showing that the "Normal High" for the day was 84 degrees. This is the basis for the statement that the high of 101 degrees was 17 degrees higher than normal. But the high temperature for October 1 is quite variable. October is the time of year when the weather is changing from summer to fall, and it's quite natural for the high temperature to be in the 70s, 80s, or 90s. While 101 was a record high, it was not "17 degrees higher than normal" if "normal" includes the range of possibilities likely to occur on that date. ■

CASE STUDY 7.1 **Detecting Exam Cheating with a Histogram**

SOURCE: Boland and Proschan (Summer 1990), pp. 10–14.

A class of 88 students at a university in Florida was taking a 40-question multiple-choice exam when the proctor happened to notice that one student, whom we will call C, was probably copying answers from a student nearby, whom we will call A. Student C was accused of cheating, and the case was brought before the university's supreme court.

At the trial, evidence was introduced showing that of the 16 questions *missed* by both A and C, both had made the same wrong guess on 13 of them. The prosecution argued that a match that close by chance alone was very unlikely, and student C was found guilty of academic dishonesty. The case was challenged, however, partly because in calculating the odds of such a strong match, the prosecution had used an unreasonable assumption. They assumed that any of the four wrong answers on a missed question would be equally likely to be chosen. Common sense, as well as data from the rest of the class, made it clear that certain wrong answers were more attractive choices than others.

A second trial was held, and this time the prosecution used a more reasonable statistical approach. The prosecution created a measurement for each student in the class except A (the one from whom C allegedly copied), resulting in 87 data values. For each student, the prosecution simply counted how many of his or her 40 answers matched the answers on A's paper. For Student C, there were

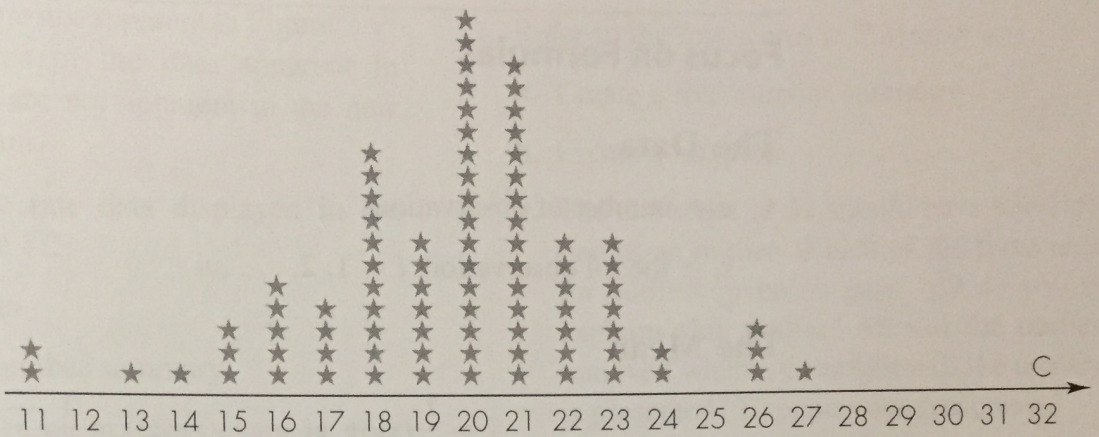


Figure 7.13
Histogram of the number of matches to A's answers for each student
Source: Data from Boland and Proschan, Summer 1990, p. 14.

32 matches to Student A, including 19 questions they both got right and 13 out of the 16 questions they both got wrong.

The results are shown in the histogram in Figure 7.13. Student C is coded as a C, and each asterisk represents one other student. Student C is an obvious outlier in an otherwise bell-shaped picture. You can see that it would be quite unusual for that particular student to match A's answers so well without some explanation other than chance.

Unfortunately, the jury managed to forget that the proctor observed Student C looking at Student A's paper. The defense used this oversight to convince them that, based only on the histogram, A could have been copying from C. The guilty verdict was overturned, despite the compelling statistical picture and evidence. ■

Thinking About Key Concepts

- Knowing the *mean* (average) of a list of numbers is not very information without additional information because *variability* is inherent in almost all measurements. It is useful to know how spread out the numbers are, what *range* they cover, what basic *shape* they have, and whether there are any *outliers*.
- *Outliers* are values that are far removed from the bulk of the data. They distort the mean and standard deviation, but have little effect on the median or the interquartile range. There are three basic reasons outliers occur, and how they should be treated depends on which of these reasons holds.
- The *mean* and *median* for a set of measurements can be very different from each other if there are *outliers* or extreme *skewness* in the numbers. The median is generally a more appropriate representation of a "typical" value than is the mean in that case.
- Useful pictures of data include *stemplots*, *histograms*, and *boxplots*. Shape can be determined from stemplots and histograms, but boxplots are the most useful type of display for comparing two or more groups.
- "Normal" should not be equated with "average." Any number in a range of values that routinely occur should be considered to be normal.

Focus on Formulas

The Data

n = number of observations

x_i = the *ith* observation, $i = 1, 2, \ldots, n$

The Mean

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The Variance

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

The Computational Formula for the Variance (easier to compute directly with a calculator)

$$s^2 = \frac{1}{(n-1)}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)$$

The Standard Deviation

Use either formula to find s^2, then take the square root to get the standard deviation s.

Exercises

Exercises with numbers divisible by 3 (3, 6, 9, etc.) are included in the Solutions at the back of the book. They are marked with an asterisk (*).

1. At the beginning of this chapter, the following "oldest ages" were listed, and a stemplot was shown for them in Figure 7.1: 75, 90, 60, 95, 85, 84, 76, 74, 92, 62, 83, 80, 90, 65, 72, 79, 36, 78, 65, 98, 70, 88, 99, 60, 82, 65, 79, 76, 80, 52, 75

a. Create a five-number summary for these ages.

b. Create a boxplot using the five-number summary from part (a).

2. Refer to the previous exercise, and the stemplot for the "oldest ages" in Figure 7.1.

a. Create a stemplot for the oldest ages using each 10s value twice instead of once on the stem.

b. Compare the stemplot created in part (a) with the one in Figure 7.1. Are any features of the data apparent in the new