CHAPTER

3

Measures of Central Tendency

INTRODUCTION

Measures of central tendency: A group of statistics that indicate where cases tend to cluster in a distribution or what is typical in a distribution. The most common measures of central tendency are the mode, median, and mean.

How much income does the typical American take home each year? Are SAT math scores higher since the implementation of the No Child Left Behind act? Is the typical level of education higher for males than for females? These kinds of questions are addressed using a group of statistics known as *measures of central tendency*. The most commonly known of these is the mean, or average. Others include the mode and the median. They tell us what is typical of a distribution of cases. In other words, they describe where most cases in a distribution are located.

In this chapter, we learn the three most commonly used **measures of central tendency**: the mode, the median, and the mean (Table 3.1). Pay particular attention to the circumstances in which each one is used, what each of them tells us, and how each one is interpreted. We also learn how to calculate each of these measures with both raw data and grouped data (data in frequency tables).

AVERAGES AND ENERGY CONSUMPTION

Average: A statistic that represents what is typical in a distribution.

Averages are one of the most commonly used statistics we encounter on a daily basis. Among a myriad of other items, averages are used to describe the weather, gas mileage in our cars, stock market performance, and our weekly expenditures. Like all statistics, averages allow us to summarize, describe, and predict events in our daily lives.

You probably learned how to calculate an average many years ago, so it may seem quite simple. It is just one of three commonly used measures of central tendency that describe what is typical about data; surprisingly, it is the complicated one of the

TABLE 3.1

Symbol	Meaning of Symbol	Formula
Measures of central tendency	A group of statistics (mode, median, and mean) that are used to describe where cases tend to fall in a distribution. They tell us what is typical in a distribution.	
Mode (*Mo*)	The most frequently occurring value or attribute	
Median (*Md*)	The value of the middle case in a rank-ordered distribution	
Mean ($\bar{X}$)	The mathematical average for sample data	$\bar{X} = \frac{\sum X}{N}$
Mean (σ)	The mathematical average for population data	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$
Position of the median	Formula used to determine which case number (in a rank-ordered distribution of cases) is the median	$\frac{N+1}{2}$
Outliers	Cases with very high or very low values	

three. In statistics, the mathematical average is referred to as the mean. The other two commonly used measures of central tendency are the mode and the median. Each of these three measures can tell us what is typical about a distribution of cases; taken together, they tell us even more. Each represents a unique description of the values around which cases tend to cluster (hence the term *central tendency*).

Averages can be found in all walks of life, but they are particularly common in the world of sports. For example, a person watching a professional football game is likely to encounter the following averages: yards gained per game, yards allowed per game, yards per play, yards per carry, yards per throw, yards per catch, yards per kickoff return, yards per punt return, and hang time of punts. Other more obscure averages are also increasingly common, for example, average number of passing yards after halftime during a snowstorm game in December while wearing Adidas shoes (just kidding, although it is possible that Adidas marketers have looked into this). Statistically, these are referred to as means, but we tend to call them averages.

Other "hidden" averages that we often do not think of as averages include the following: winning percentage, pass completion percentage, and field goal percentage. Baseball, basketball, hockey, and other sports provide viewers with a plethora of statistical averages that consumers can use to reify their sports viewing experience by using them to draft fantasy teams, bet on events, and talk sports with friends. Like all statistics, we tend to use these averages to summarize, describe, and predict the outcome of sporting events and individual performances. In fact, sports may be the social institution most responsible for single-handedly popularizing statistics among the American public. This chapter discusses these three statistics—mode, median, and mean—and demonstrates the advantages and drawbacks of each using a variety of examples.

The distribution of cases across the attributes of any variable can be described using a variety of statistics. The three most commonly used measures of central tendency are the mode, median, and mean. Each is based on a distinct logic, but all three are used for the same general purpose: They tell us where in a distribution the cases tend to cluster.

© iStockphoto.com/JMichl

The mode, median, and mean cannot be used interchangeably. Each tells us something different. The mode tells us which attribute has the greatest frequency. The median tells us the value of the case that divides the distribution into two equal-sized parts, one part containing cases with greater values and the other part containing cases with lesser values. The mean tells us the arithmetic average of the cases.

It is important to understand that certain measures of central tendency apply to only certain types of variables. When working with nominal variables, only the mode can be used. When working with ordinal variables, only the mode and median can be used. When working with interval/ratio variables, all three measures can be used. See Table 3.2 for a summary of when different measures can be used.

On October 12, 2006, the U.S. Bureau of the Census announced that the population of the United States would reach 300 million people on October 17, 2006, at about 7:46 a.m. Eastern Daylight Time. Of course, nobody really knew the exact moment when the population reached this historic milestone, but at some point during mid-October 2006, it is highly likely that somebody became the 300 millionth live American (as of 2018, the U.S. population was about 326 million). As only the fourth society in the history of the world to reach this milestone, it puts the United States in the same company as China, India, and the former Soviet Union. To put this number in perspective, it would take the combined populations of present-day Germany (83 million), France (68 million), Italy (61 million), Spain (47 million), Poland (38 million), and Republic of Ireland (5 million) to surpass the population of the United States in 2006.

Population has been a popular topic for social analysts, city planners, and environmentalists for thousands of years. As the world population grew into the 19th century, the work of the Reverend Thomas Malthus (1766–1834) took on greater significance. Malthus (1798/1993) argued that because population grows exponentially while food production grows linearly, the day will come when hunger and starvation become unavoidable aspects of life on Earth. He also argued that a series of preventive checks, such as delayed marriage and abstinence, would help to avoid or at least minimize the problems associated with food shortages. While some analysts still adhere to variations on Malthus's ideas, others claim that nutrition and starvation have economic, political, and technological dimensions that Malthus did not consider.

Nevertheless, in the current model of development, with population growth come increases in energy consumption. History shows that the more people there are, the more energy they consume. The question is, *how big of an increase in energy consumption accompanies the increases in population growth?* A comparison of total primary energy consumption (petroleum, dry natural gas, coal, nuclear, hydroelectric, and nonhydroelectric renewable) across countries helps to answer this question. According to the U.S. Energy Information Administration's international data

TABLE 3.2

	Nominal	Ordinal	Interval/Ratio
Mode	Applies	Applies	Applies
Median	NA	Applies	Applies
Mean	NA	NA	Applies

Note: NA, not applicable.

browser (https://www.eia.gov/beta/international/analysis.php), the U.S. population, which in 2003 was just under 300 million people, consumed a total of 103.3 exajoules (EJ) of energy (1 EJ = 10^{18} joules [power]). By comparison, China, with a population of 1.3 billion people, consumed 57.4 EJ, and India, with 1.1 billion people, consumed 15.2 EJ. To put all these numbers in perspective, while the population of the United States was only 23% that of China, the United States consumed 1.8 times the total amount of energy that China consumed. This means that, at 2003 consumption levels, China's population could nearly double before it consumes as much energy as the United States' population. Similarly, while the U.S. population was only 27% that of India, the United States consumed nearly seven times the total amount of energy. However, as the economies of both China and India have come to resemble the economy of the United States, their per capita energy consumption has risen dramatically. In fact, from 2003 to 2013, China's total energy consumption more than doubled to 125.1 EJ and India's grew to 24.8 EJ.

These kinds of statistics put population growth in perspective by differentiating between qualitatively different types of populations. Averages make them understandable. For example, on average, in 2003 a resident of the United States used as much energy as about 8 residents of China or 25 residents of India. As biologist Paul Ehrlich noted, data on consumption tell us a lot about the kind of society in which we live (Ehrlich & Ehrlich, 2008). With a dwindling supply of nonrenewable fuels (fossil fuels) and the effects of burning those fuels, the world might be able to accommodate many more people; however, it might not be able to accommodate too many more Americans.

DEFINITIONS OF MODE, MEDIAN, AND MEAN

Mode. The **mode** is the value or number that occurs most often in a distribution of cases. In other words, in a frequency distribution, the attribute with the greatest frequency is the mode. It is possible to have more than one mode in a distribution.

Mode: The most frequently occurring value in a distribution.

A distribution with two modes is called a bimodal distribution. Likewise, a distribution with three modes is called a trimodal distribution. The mode can be used to describe nominal, ordinal, and interval/ratio variables.

Median. In a rank-ordered distribution of cases, the **median** is the value of the middle case; it is the value of the case that divides the distribution into two equally sized groups, those with higher values and those with lower values. For example, in a distribution with the numbers 4, 5, and 6, the median is 5 because it is in the middle. In a frequency distribution with an even number of cases, such as the numbers 4, 5, 6, and 7, the median is the value between 5 and 6, or 5.5. In small datasets like these, it is easy to identify the middle.

Median: The value of the middle case in a rank-ordered distribution.

In larger datasets, such as those with hundreds or thousands of cases, it is much more difficult to determine the midpoint. A simple formula can help. To determine the number of the case that falls in the middle of a rank-ordered distribution of cases (meaning they are in order from lowest to highest), use the formula $\frac{N+1}{2}$. In our example using the values 4, 5, and 6, our N is equal to 3. Using $\frac{N+1}{2}$ we find that $\frac{3+1}{2}=2$. The answer 2 tells us that we need to count over two cases in our rank-ordered distribution of cases. The second case has a value of 5 so the median is 5. It should be noted that when finding the median, you should always remember to first rank the cases in order from highest to lowest. Also remember that the median is not equal to the result of the formula $\frac{N+1}{2}$. This only tells us which case number to go to in order to find the value of the median. The median is equal to the value of the middle case.

$$Position\ of\ the\ median = \frac{N+1}{2}$$

The median is a very popular measure of central tendency when working with ecological data representative of towns, states, and other geopolitical units. For example, when describing the typical income of a city such as Buffalo, New York, it is often better to use median household income than to use average household income. Doing so helps to control for the effect of outliers (discussed later in this chapter).

Mean. The **mean** is the mathematical average. It is equal to the sum ($\sum$) of all cases divided by the number of cases (N). The formula for the mean is

Mean: The mathematical average. It is calculated by summing all the scores in a distribution and dividing by the total number of scores.

$$\bar{X} = \frac{\sum X}{N}$$

The mean is the most sophisticated of all the measures of central tendency because it incorporates the most information in its calculation, but it is also a relatively "unstable" measure.

The symbol for the mean is $\bar{X}$ and is pronounced "x bar." The mean can be greatly influenced by cases with extremely low or extremely high values. These very high or very low values are called outliers. Outliers have the effect of pulling the mean

toward them and skewing the data. **Skew** occurs when outliers pull the mean far from the other two measures of central tendency.

Skew: The degree to which a normal distribution is distorted due to the effect of outliers.

CALCULATING MEASURES OF CENTRAL TENDENCY WITH RAW DATA

Public transportation is considered a good way to reduce the cost of owning an automobile and an effective means by which to reduce the amount of carbon dioxide emissions in the atmosphere. Suppose that we are interested in learning more about the ability of improvements in public transportation to reduce the amount of driving that people in a particular neighborhood do. It is first necessary to find out how often people in this neighborhood use the bus, so we begin by sampling 15 residents on how many times they have boarded a city bus during the past two weeks. (Obviously, we would want to sample more than 15, but this is just for demonstration purposes.) Our sample yields the data shown in Table 3.3.

TABLE 3.3

Rider #	# Bus Trips
1	5
2	3
3	10
4	5
5	8
6	6
7	5
8	4
9	0
10	2
11	14
12	5
13	7
14	4
15	3

To begin our analysis, the cases (frequency values for number of bus trips) are first rank-ordered from lowest to highest values:

0 2 3 3 4 4 5 5 5 5 6 7 8 10 14

The mode is defined as the value that occurs most frequently. It is the simplest measure of central tendency to calculate because all we need to do is count the number of times each value occurs. Four respondents indicated that they boarded city buses five times during the past two weeks. Because no value has a frequency greater than 4, the mode for these data is 5. It is important to remember that the mode is not equal to the frequency of five bus rides during the last two weeks (which is 4); it is equal to five bus rides.

Mode = 5

The median is defined as the value of the case that divides a rank-ordered distribution in half. In datasets with a very small number of cases, such as our bus riding example, it is fairly simple to determine the middle case. If we select the middle case, there should be seven cases with higher values and seven cases with lower values. The middle case, the seven cases below it, and the seven cases above it add up to 15 cases. But what do we do when we have hundreds, thousands, or even tens of thousands of cases? We need a way to determine which case is in the middle. We use the formula below to solve this problem:

$$Position\ of\ the\ median = \frac{N+1}{2}$$

We take the number of cases, add 1, and divide by 2. In this case, it means that our equation looks like this:

$$\textit{Position of the median} = \frac{15+1}{2} = \frac{16}{2} = 8.$$ The middle case is case number 8.

0 2 3 3 4 4 5 **5** 5 5 6 7 8 10 14

Case 1 2 3 4 5 6 7 **8** 9 10 11 12 13 14 15

Looking back at our rank-ordered distribution of cases, we simply count (from right to left or from left to right) over to the eighth case. In these data, the value of the eighth case is equal to 5. Therefore, the median is equal to 5.

Median = 5

The mean is equal to the arithmetic average of all the cases. To calculate the mean, we add the values of all the cases to determine the sum (∑). Then we divide the sum by the number of cases (*N*). The formula for the mean looks like this:

$$\bar{X} = \frac{\sum X}{N}$$

In other words, this is the sum of each individual case divided by the number of cases. For our bus riding dataset, it looks like this:

$$\bar{X} = \frac{\sum(0, 2, 3, 3, 4, 4, 5, 5, 5, 5, 6, 7, 8, 10, 14)}{N} = \frac{81}{15} = 5.4$$

Mean = 5.4

As these results show, the mode and median have the same value of 5. The mean, on the other hand, is slightly larger with a value of 5.4. Looking at the rank-ordered distribution of cases, we can see that the cases with values higher than the median (14, 10, 8, 7, 6, 5, 5) tend to be a greater difference from the median than the cases below the median (0, 2, 3, 3, 4, 4, 4). Because the degree of difference from the median is greater on the higher value side than on the lower value side, we know that the mean is going to be pulled toward the higher values.

NOW YOU TRY IT 3.1

Now that you have seen a demonstration of how measures of central tendency are calculated using raw data, try to calculate them on your own using the data below. The data represent responses from 20 students who were asked "How many credits are you taking this semester?"

15	16	15	12	9	15	16	16	15	15	18	12	10	14	12	15	15	18	15	15

What is the value of the mode? The median? The mean?

CALCULATING MEASURES OF CENTRAL TENDENCY WITH GROUPED DATA

As discussed in Chapter 2, sociologists and other scientists tend not to work with raw data. Instead, they work with grouped data such as frequency tables. Frequency tables actually make the calculation of the three measures of central tendency easier. Table 3.4 was generated using SPSS for Windows and is based on the same data that were used in the bus riding example.

Mode. To find the mode using a frequency table, look in the Frequency column for the largest *f* value. In this example, that value is 4. Remember that this does not mean that the mode is equal to 4; instead, it means that four different respondents indicated that they took five bus rides during the past two weeks. The mode is 5. Five bus rides per week is the modal value for our distribution.

Median. To find the median using a frequency table, we have two options available. The first is to determine the middle case using the formula $\frac{N+1}{2}$. After determining the middle case, we use the Frequency column in the table to count the cases. For example, the middle case is equal to $\frac{15+1}{2}$ or 8. We then begin counting frequencies from the top of the table until we get to the eighth case. When we add up all the respondents who took zero, one, two, three, or four bus rides, we have six cases. When we include the four respondents who took five bus rides, we are up to 10 cases. Therefore, the eighth case was one of the respondents who took five bus rides. Therefore, the median is equal to 5.

A second simpler method to use to determine the median is the Cumulative Percent column. Remember that the Cumulative Percent column is a running percentage total

TABLE 3.4 Number of Bus Rides During the Past Two Weeks

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1	6.7	6.7	6.7
	2	1	6.7	6.7	13.3
	3	2	13.3	13.3	26.7
	4	2	13.3	13.3	40.0
	5	4	26.7	26.7	66.7
	6	1	6.7	6.7	73.3
	7	1	6.7	6.7	80.0
	8	1	6.7	6.7	86.7
	10	1	6.7	6.7	93.3
	14	1	6.7	6.7	100.0
	Total	15	100.0	100.0	

based on the Frequency column, so rather than use the formula to find the middle case, we can look for the "50% point" in the Cumulative Percent column.

As evidenced by Table 3.4, the 50% mark must often be imagined rather than seen. For example, if we include all the respondents who rode the bus zero times, we have included 6.7% of the sample; if we include all the respondents who rode the bus either zero times or one time, we have included 13.3% of the sample. If we continue this process until we include all the respondents who rode the bus zero, one, two, three, or four times, we have included 40.0% of the sample. Taking this one step further and including all those who rode the bus zero, one, two, three, four, or five times, we are up to 66.7% of the sample, well past the 50% mark. Therefore, the median represents those who rode more than four times but less than six times. In other words, the 50% mark must fall between 40.0% and 66.7%. This means that the median falls in the group that rode the bus five times during the past two weeks. The median is 5.

Mean. To find the mean using a frequency table, it is necessary to use both the Value column (the column at the far left of the table) and the Frequency column to determine the total number of bus rides the 15 respondents have taken. For example, if two respondents rode the bus three times each, they rode a total of six times (2 × 3). Therefore, to determine the total number of bus rides taken ($\sum X$), we need only to multiply the number of bus rides by the frequency of respondents for each row in the table.

For example, in Table 3.5, zero bus rides were taken by one respondent. Therefore, 0 × 1 = 0 bus rides. Similarly, two bus rides were taken by one respondent. Therefore, 2 × 1 = 2 bus rides. This process is repeated for each value of bus rides and the results are added to give us the sum total of bus rides ($\sum X$). After determining this, we divide the number of bus rides by the number of respondents to determine our average using the formula for the mean:

$$\bar{X} = \frac{\sum X}{N}$$

Table 3.5 shows a demonstration using the bus riding data.

TABLE 3.5

	Rides	Frequency	Frequency of Bus Rides
Valid	0	1	0 × 1 = 0
	2	1	2 × 1 = 2
	3	2	3 × 2 = 6
	4	2	4 × 2 = 8
	5	4	5 × 4 = 20
	6	1	6 × 1 = 6
	7	1	7 × 1 = 7
	8	1	8 × 1 = 8
	10	1	10 × 1 = 10
	14	1	14 × 1 = 14
	Total	15	Sum of bus rides ($\sum X$) = 81

Once we have determined the sum of bus rides, we only need to divide by N to calculate the mean, just as we did with the example using raw data. Therefore, the mean, $\bar{X} = \frac{\Sigma X}{N}$, is equal to $\frac{81}{15}$ or 5.4. The process of calculating the mean using a frequency table may seem cumbersome at first, but it can save a great deal of time, particularly when working with frequency tables that have hundreds or thousands of cases.

The mean is 5.4. Notice how a small number of respondents who rode the bus a lot have the effect of increasing the mean relative to the mode and median.

STATISTICAL USES AND MISUSES

In his now classic book *How to Lie With Statistics* (1954), Darrell Huff discussed the idea of the "well chosen average" (p. 27). As Huff noted, the term *average* can be used to refer to any one of the three different measures of central tendency discussed here; however, as has been shown in this chapter, each of these measures tells us something quite different. Using Huff's real estate example, it is often said that buyers are better off with a low-cost house in a high-income neighborhood than with a high-cost house in a low-income neighborhood. If a real estate agent tells you that your $150,000 house is the average cost of a home in your neighborhood, are they referring to the mean (the mathematical average), the median, or the mode? Unless you know what questions to ask, you may never find out until after you have bought the most expensive house on the block (the one house that is pulling the average home price up to $150,000, i.e., the outlier).

Similarly, in his book *Full House* (1997), Stephen J. Gould calls for further analysis of the median life expectancy for those diagnosed with often fatal diseases. For example, when a person is diagnosed with a deadly disease, the doctor may tell him or her that the median life expectancy of someone with this disease is two years. In other words, the person is being told that he or she has

© iStockphoto.com/fstop123

(Continued)

(Continued)

two years to live. According to Gould, this leaves out at least one important consideration: How advanced was the disease when it was detected? The longer a disease goes undetected, the shorter the life expectancy of the person with the disease because of the lack of treatment; however, if the disease is detected in its early stages, then life expectancy is much longer.

As these two examples show, it is necessary to have a working knowledge of both the statistics (measures of central tendency) and the context (real estate, medical) in which the statistics are being used. Without this knowledge, people risk being misinformed, misled, and maybe even "hustled" by someone with something to gain from others' lack of statistical know-how!

Education and Central Tendency

Roughly a quarter of the U.S. population receives a four-year bachelor's degree or higher. In fact, according to the 2016 General Social Survey (GSS) (National Opinion Research Center, 2016), 18.7% reported receiving a bachelor's degree and 11.1% reported receiving a graduate degree. Unbeknownst to many, on receiving a bachelor's degree we become a member of the top 1% of the world's most educated people (formally educated). This says quite a lot about the degree of educational inequality nationally and globally.

Table 3.6 shows the distribution of education degrees in the United States in 2016. As the table and the corresponding chart below show, the most common degree that people receive in this country is a high school diploma ($f = 1{,}461$), and the second most common degree is a bachelor's degree ($f = 536$).

Figure 3.1 provides a visual representation of this pattern. The highest frequency in a frequency table, which is also the tallest bar in the bar chart, represents an attribute around which cases tend to cluster. The attribute with the greatest frequency of cases is referred to as the mode.

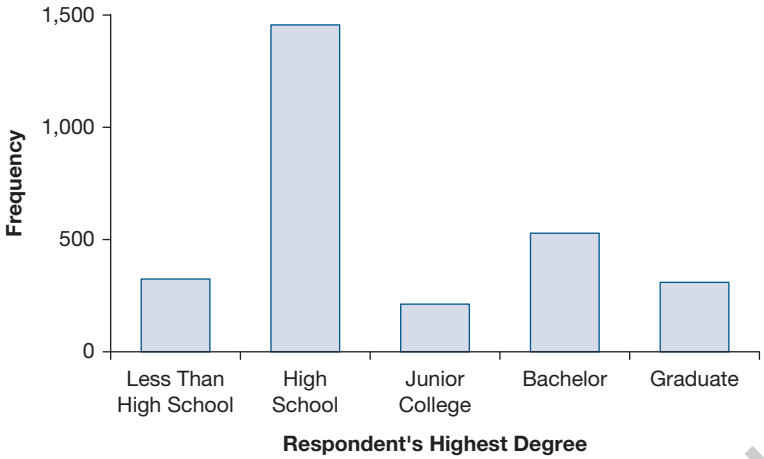
Mode = High School

TABLE 3.6 Respondent's Highest Degree

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Less than high school	328	11.4	11.5	11.5
	High school	1,461	51.0	51.1	62.6
	Junior college	216	7.5	7.6	70.1
	Bachelor	536	18.7	18.7	88.9
	Graduate	318	11.1	11.1	100.0
	Total	2,859	99.7	100.0	
Missing	No answer	8	.3		
Total		2,867	100.0		

Source: Data from the National Opinion Research Center, General Social Survey.

FIGURE 3.1 Respondent's Highest Degree



Source: Data from the National Opinion Research Center, General Social Survey.

We can also see that the median is equal to high school diploma in Table 3.6 because the 50% mark in the Cumulative Percent column comes somewhere after the less than high school respondents and before the junior college respondents.

Another way to calculate the median is to find the middle case and use the Frequency column of Table 3.6 to determine in which attribute the middle case falls. For example, we know that $N = 2{,}859$. Therefore, using the formula for the position of the median, $\frac{N+1}{2}$, we know which case is the middle case: $\frac{2{,}859+1}{2} = 1{,}430$. Now we must find the 1,430th case.

If we add up all the respondents with less than high school, we are up to 328 cases.

If we add up all the respondents with less than high school and high school, we are up to 1,789 cases.

At this point, we have surpassed the middle case number of 1,430. Therefore, the middle case is somewhere in the high school group. The median is high school:

Median = High School

We can see that education is not distributed evenly across the population. Wealthier people are more likely to have higher degrees, as are people whose parents went to college, who are white, and who are male. In a study published in the *American Sociological Review,* sociologists Claudia Buchman and Thomas DiPrete noted that in 1960 65% of all bachelor's degrees were awarded to men. This held true until 1982, and since then more bachelor's degrees have been awarded to women than to men (Buchman & DiPrete, 2006, p. 515). According to more recent data from the GSS, women receive about 55% of all bachelor's degrees and 59% of all graduate degrees (National Opinion Research Center, 2016).

This interesting trend is exemplified in another 2016 GSS variable, Years of Education, which is operationalized at the interval/ratio level. We use this variable to investigate the relationships between education and health.

NOW YOU TRY IT 3.2

Now that you have seen a demonstration of how measures of central tendency are calculated using frequency tables, try to calculate them on your own using the data below. The table shows data from 50 respondents who were asked how many brothers and sisters they have.

Number of brothers and sisters

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	4	8.0	8.0	8.0
	1	15	30.0	30.0	38.0
	2	11	22.0	22.0	60.0
	3	8	16.0	16.0	76.0
	4	6	12.0	12.0	88.0
	5	6	12.0	12.0	100.0
	Total	50	100.0	100.0	

What is the value of the mode? The median? The mean?

Central Tendency, Education, and Your Health

Most people hope to live long healthy lives. Epidemiological research shows that, among other factors, social class is a major determinant of health and longevity (California Newsreel, 2008). In other words, people of upper-class standing are more likely to stay healthy and live longer than people of lower-class standing. Greater access to health care, healthier lifestyles, less dangerous working conditions, and healthier diets are just a few of the benefits that come with greater levels of wealth and income. Not surprisingly, education is one of the keys to achieving higher class standing. So just how much education does the typical resident of the United States have?

Each year the National Opinion Research Center at the University of Chicago conducts its GSS, in which thousands of Americans are interviewed on topics ranging from their age to their education to their opinions about government spending. One of the questionnaire items included each year is the number of years of education that each respondent has. More precisely, interviewers ask respondents something to the effect of "What is the highest year of school you have completed?" The responses for the year 2016 are shown in Table 3.7.

Table 3.7 is based on N = 2,858. Measures of central tendency give us an idea of the number of years of education that are typical. Figure 3.2 represents this table in graphic form. It is useful because it gives us a visual representation of the

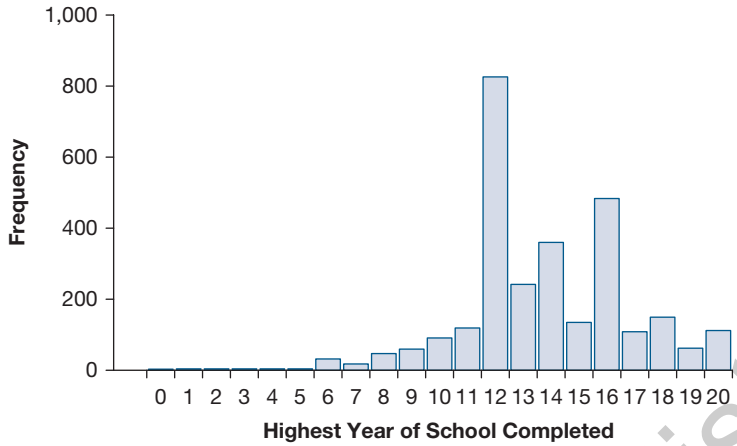
© iStockphoto.com/filadendron

TABLE 3.7 Highest Year of School Completed

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	2	.1	.1	.1
	1	3	.1	.1	.2
	2	3	.1	.1	.3
	3	3	.1	.1	.4
	4	2	.1	.1	.5
	5	4	.1	.1	.6
	6	31	1.1	1.1	1.7
	7	18	.6	.6	2.3
	8	48	1.7	1.7	4.0
	9	59	2.1	2.1	6.1
	10	90	3.1	3.1	9.2
	11	118	4.1	4.1	13.3
	12	824	28.8	28.8	42.2
	13	242	8.5	8.5	50.6
	14	359	12.6	12.6	63.2
	15	137	4.8	4.8	68.0
	16	485	17.0	17.0	85.0
	17	108	3.8	3.8	88.7
	18	149	5.2	5.2	93.9
	19	63	2.2	2.2	96.2
	20	110	3.8	3.8	100.0
	Total	2,858	100.0	100.0	

Source: Data from the National Opinion Research Center, General Social Survey.

FIGURE 3.2 Highest Year of School Completed



Source: Data from the National Opinion Research Center, General Social Survey.

data from the table and allows us to see trends that are not readily apparent in the table. Relatively few respondents have less than 8 or 9 years of education, whereas a significant number of them cluster between 12 and 16 years of education. Those respondents with less than 6 years of education qualify as outliers and have the effect of (1) lowering the mean and (2) negatively skewing the distribution.

As we can see in Figure 3.2, 824 respondents indicated that they have 12 years of education, most likely the equivalent of a high school diploma. We also see that many respondents have 14 years of education (a two-year college degree) and 16 years of education (a four-year college degree). Because no other attribute has a greater frequency than 12, the mode is equal to 12.

Mode = 12

Using Figure 3.2 to determine the median is virtually impossible. Instead, we use Table 3.7. Remember that the median is the 50% mark, or the point at which half the cases fall above and half fall below. As shown earlier, using the equation $\frac{N+1}{2}$ we find that $\frac{2{,}859+1}{2}=1{,}430$. In other words, case number 1,430 is the middle case in the distribution, so the number of years of education associated with case number 1,430 is the median. By adding all the cases with 0 to 12 years of education, we reach a total of 1,205, still short of the middle case. When we add the 242 respondents with 13 years of education, we reach a total of 1,447, beyond the middle case of 1,430. Therefore, the middle case fell among the respondents with 13 years of education. This is a laborious way to find the median. Rather than adding all the cases, we can use the Cumulative Percent column to locate the 50% mark in the distribution.

After including all respondents up to and including 12 years of education, we are up to only 42.2% of the sample; however, after including all respondents up to and including 13 years of education, we are up to 50.6% of the sample. Therefore, the middle case needed to be a person with 13 years of education. Consequently, the median is equal to 13.

Median = 13

Calculating the mean is also done most easily with the table. First, we must determine the total number of years of education that all 2,858 respondents have achieved. Then we divide by the number of respondents. Table 3.8 shows these calculations.

Using the formula $\bar{X} = \frac{\sum X}{N}$, we find that $\frac{39{,}261}{2{,}858} = 13.7$. The mean number of years of education for respondents in this sample is 13.7. Therefore, if the sampling was done in a methodologically sound manner, we can predict that the average number of years of education for Americans is 13.7.

Mean = 13.7

To test the hypothesis that income is associated with health, we would need to gather additional data on our respondents' health status so that we could statistically determine whether those respondents with the highest levels of education are in fact those who tend to have higher levels of overall health. A variety of statistical tools

TABLE 3.8 Highest Year of School Completed

		Frequency	Years of Education × Frequency
Valid	0	2	0 × 2 = 0
	1	3	1 × 3 = 3
	2	3	2 × 3 = 6
	3	3	3 × 3 = 9
	4	2	4 × 2 = 8
	5	4	5 × 4 = 20
	6	31	6 × 31 = 186
	7	18	7 × 18 = 126
	8	48	8 × 48 = 384
	9	59	9 × 59 = 531
	10	90	10 × 90 = 900
	11	118	11 × 118 = 1,298
	12	824	12 × 824 = 9,888
	13	242	13 × 242 = 3,146
	14	359	14 × 359 = 5,026
	15	137	15 × 137 = 2,055
	16	485	16 × 485 = 7,760
	17	108	17 × 108 = 1,836
	18	149	18 × 149 = 2,682
	19	63	19 × 63 = 1,197
	20	110	20 × 110 = 2,200
	Total	2,858	Total years of education = 39,261

called *measures of association* can be used to test our prediction; however, they are presented in a later chapter.

Table 3.9 shows that there is some connection between education and perception of personal health. For example, of those with less than high school, we see that only 10.6% of respondents consider their health to be excellent. This is quite low compared with all other levels of education, including high school (16.4%), junior college (28.8%), bachelor (33.1%), and graduate (36.2%). These results are shown in Figure 3.3.

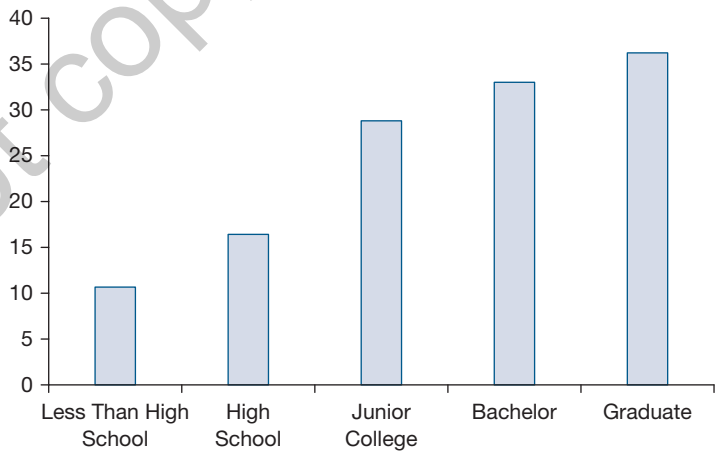
We might also want to view these data when education is operationalized as an interval/ratio variable. In this example, education is operationalized as the number of years of education that a respondent indicates he or she has completed. In this example, the axes have been switched and respondents' health is now the categorical variable on the *X*-axis of Figure 3.3. The height of each bar indicates the mean number of years of education for each category of health.

TABLE 3.9 Condition of Health

		Respondent's Highest Degree					
		Less Than High School	High School	Junior College	Bachelor	Graduate	Total
Condition of health	Excellent	10.6%	16.4%	28.8%	33.1%	36.2%	22.2%
	Good	38.7%	51.0%	46.8%	48.6%	51.1%	48.8%
	Fair	36.4%	25.4%	20.1%	16.1%	10.4%	22.8%
	Poor	14.3%	7.2%	4.3%	2.2%	2.3%	6.3%
Total		100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Source: Data from the National Opinion Research Center, General Social Survey.

FIGURE 3.3 Percentage With Excellent Health by Highest Degree



Source: Data from the National Opinion Research Center, General Social Survey.

As Figure 3.4 shows, those who feel they are in excellent health have an average of 15 years of education, those in good health average 14 years, those in fair health average 13 years, and those in poor health average 12 years. These relatively simple averages tell us that sociology and the study of inequality has a great deal to offer in the way of public health improvements. The social bases of health may be far more important than anyone previously thought.

Income and Central Tendency: The Effect of Outliers

Outliers: Cases with values either higher or lower relative to the typical pattern in a distribution.

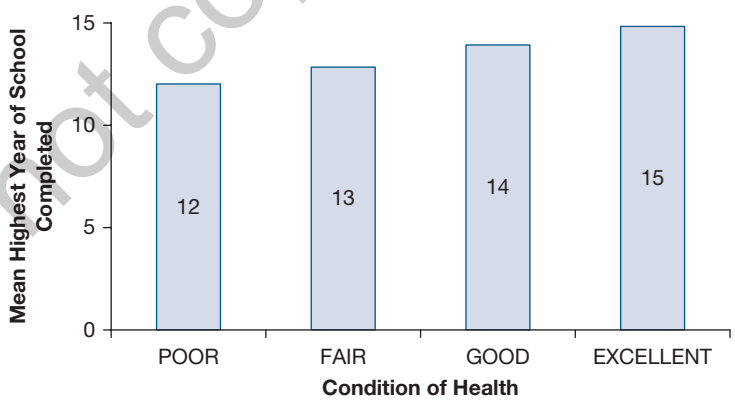
Outliers are cases with values that far exceed the normal range of values in a distribution. They have the effect of skewing data (significantly pulling the distribution in one direction) and, in some cases, should even be removed from the data; however, their removal needs to be justified.

For example, suppose you live in a town with a population of 40 ($N = 40$), with 10 residents earning $25,000 per year, 20 residents earning $50,000 per year, and 10 residents earning $75,000 per year. This is an interval/ratio variable, and all three measures of central tendency (mean, median, and mode) are equal to $50,000 per year. A frequency table and **histogram** with a normal curve for these data are shown in Table 3.10 and Figure 3.5.

Histogram: Used with interval/ratio variables. The height of each bar represents the frequency of each attribute.

Now suppose that two wealthy computer software developers move to your town, each with an income of $200,000 per year. This brings the town population to 42 and has the effect of raising the mean income to $57,143. The mean income for the town rises dramatically but is not representative of most people in the town. After all, 40 of the 42 residents still make an average of $50,000 per year. The wealthy computer software moguls are considered outliers, cases that have the effect of pulling the average toward the value of the outlier. Just like outliers with very high values can raise the average, so too can outliers with very low values lower the average. If a very poor person with an income of $1 per year moved into town in place of the person with very high income, it would have the effect of pulling down the mean income for the town. The effect that the two very high-salaried residents have on the town's income distribution is shown in the frequency table and histogram with normal curve in Table 3.11 and Figure 3.6.

FIGURE 3.4 Mean of Highest Year of School Completed by Condition of Health



Source: Data from the National Opinion Research Center, General Social Survey.

TABLE 3.10 Income

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	$25,000	10	25.0	25.0	25.0
	$50,000	20	50.0	50.0	75.0
	$75,000	10	25.0	25.0	100.0
	Total	40	100.0	100.0	

FIGURE 3.5 Chart Showing a Perfectly Even Distribution of Cases Around a Mean of $50,000

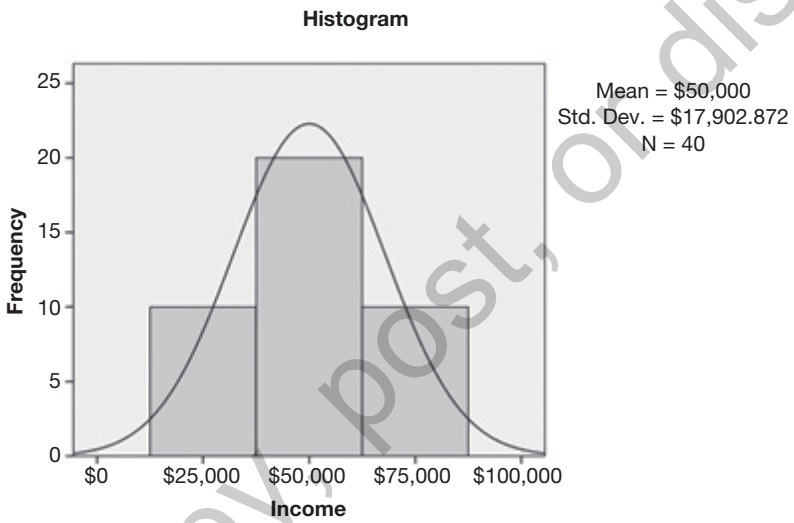
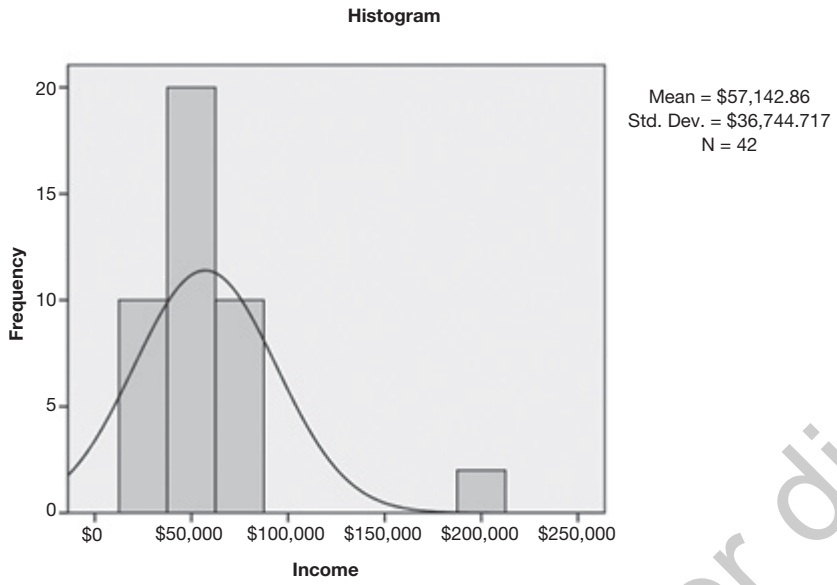


TABLE 3.11 Income

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	$25,000	10	23.8	23.8	23.8
	$50,000	20	47.6	47.6	71.4
	$75,000	10	23.8	23.8	95.2
	$200,000	2	4.8	4.8	100.0
	Total	42	100.0	100.0	

FIGURE 3.6 Histogram of Income With Outliers in the Data



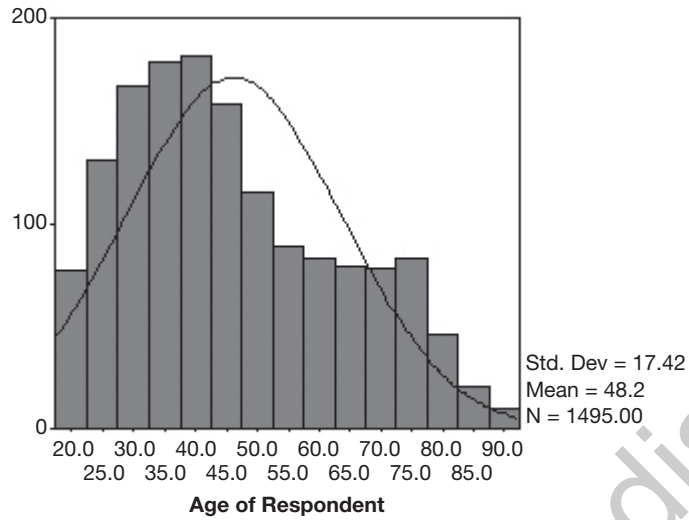
Notice that in the first histogram (Figure 3.5) the normal curve line goes to the bottom of the chart at about $100,000, and in the second histogram (N = 42) (Figure 3.6) the normal curve line goes to the bottom of the chart at about $200,000. This has the effect of making the normal curve have a "tail" on the right-hand side and skewing it positively (toward larger values).

Unlike the mean, the median is unaffected by the outlier because the person with very high income represents only a single case or a very small number of cases. As the point at which half the cases fall below and half fall above, the addition of either a computer software mogul or a very poor person does not affect the median income of the town at all. The median remains $50,000 per year. Often, particularly in the case of income data that are likely to contain outliers, it is more common to use the median than to use the mean. In these cases, the median gives us a better representation of the typical income of the community.

Like the median, the mode is not affected by the presence of outliers. More people make $50,000 than any other amount; therefore, the mode remains at $50,000 per year.

In general, it is useful to use all three measures of central tendency when describing where cases fall in a distribution. Figure 3.7 shows a histogram with a normal curve of age. Some summary statistics are offered to the right of the chart, including N and the mean. We see that the tallest bar in the histogram represents 40.0 years. Therefore, the mode, or the modal age, is 40. We also see that the peak of the normal curve is at an age higher than that of the mode. The peak of the normal curve is the mean. In this case, it is 46.2 years.

FIGURE 3.7 Histogram of Age



Source: Data from the National Opinion Research Center, General Social Survey.

Because the mean is greater than the mode, we know that more cases with higher values (possibly outliers) are present in the data. This also means that the mean, being affected more by outliers than the mode or median, will fall to the right of the mode.

Because most distributions approximate a normal curve but do not take the exact shape of a normal curve, we say that distributions are skewed. A distribution can be skewed positively (which is usually, but not always, to the right) or negatively (which is usually, but not always, to the left). Remember, *skew* is the result of *outliers*. Skew refers to the tail of a distribution. Outliers are cases in the distribution with either very high or very low values. They have the effect of pulling the mean toward them. Therefore, a distribution of age with a few extremely old people would have a higher average than without the extremely old people.

The mean is the measure of central tendency that is most affected by outliers. The median is generally not affected. Therefore, a distribution with outliers tends to be more skewed than a distribution without outliers. Often outliers are excluded from statistical analyses because they distort the normal trends in the data so significantly that the resulting statistics no longer represent what is typical, or average, about a population.

The examples below further demonstrate the effect of outliers. Table 3.12 represents a distribution of 40 respondents between the ages of 18 and 20 years. Table 3.13 represents those same respondents except that a 20-year-old has been replaced with a respondent who is 120 years old, an outlier.

As you can see, the addition of the outlier has no effect on the median; it remains unchanged.

The mean, however, does change.

TABLE 3.12 Age of Respondent

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18	5	12.5	12.5	12.5
	19	17	42.5	42.5	55.0
	20	18	45.0	45.0	100.0
	Total	40	100.0	100.0	

Statistics

Age of Respondent

N	Valid	40
	Missing	0
Mean		19.33
Median		19.00
Mode		20

TABLE 3.13 Age of Respondent

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18	5	12.5	12.5	12.5
	19	17	42.5	42.5	55.0
	20	17	42.5	42.5	97.5
	120	1	2.5	2.5	100.0
	Total	40	100.0	100.0	

In Table 3.13, which contains the 120-year-old respondent, we see a significant increase in the mean. In fact, the mean increases from 19.33 to 21.83 years. We know that if we plotted a normal curve for each of these distributions, the distribution containing the 120-year-old respondent would be skewed positively to the right. This demonstrates the usefulness of having more than one measure of central tendency to describe how cases are distributed. Some examples of how to calculate measures of central tendency are now offered.

EYE ON THE APPLIED

THE TECHNOLOGICAL FIX TO CLIMATE CHANGE

Many people feel that by adopting new technologies, we can save both money and the environment by burning less fossil fuel; however, some trend data from the Bureau of Transportation Statistics reveal that putting our faith in this "technological fix" may do more harm than good.

Between 1990 and 2016, average new vehicle fuel efficiency for light-duty vehicles increased from 18.8 to 22.0 miles per gallon (mpg), a 17% gain. The gain is even greater when analyzing the switch from regular cars to the hybrid versions of new cars, but for now we consider only the overall average.

Can the gains in fuel economy help to curb the overall level of carbon dioxide originating among drivers? The answer is a flat *no*. U.S. Bureau of Transportation Statistics data show a steady increase in the total number of vehicle miles traveled. Americans increased their aggregate number of vehicle miles traveled by 108% (meaning it more than doubled) between 1980 and 2016. Between 1990 and 2016, the increase was 48%. Despite well-intentioned attempts on behalf of individuals to reduce their transportation fuel consumption by purchasing more fuel-efficient vehicles, rising population and a growing automobile culture result in societal driving totals that far outstrip any gains made in fuel efficiency. In other words, despite all the effort to curb auto emissions, as a society we drive far more than ever before with far greater carbon dioxide emissions from driving. The data suggest that an immediate and total conversion to hybrid vehicles could initially reduce the total amount of transportation fuels consumed to around those levels found during the mid to late 1990s (why the 1990s should be considered a benchmark era for sustainability is another question); however, such an effort seems highly unlikely.

The reality of these aggregate statistics is that we consume more than ever despite gains made in efficiency, just as we drive more than ever despite gains in fuel economy. It may be time to remove the phrase *miles per gallon* from the discourse on climate change; this phrase may do more harm than good. This idea reflects sociologist Emile Durkheim's claim that the whole is greater than the sum of its parts. In other words, we can't rely on an analysis of individual drivers to solve climate change. Instead, we need to understand driving as a social phenomenon, a characteristic of society, not as an individual behavior.

Example 1: Calculating Measures of Central Tendency With Raw Data

Suppose the following nine cases constitute our data:

4, 5, 8, 3, 2, 5, 4, 5, 5

Rank-ordering the cases from low to high results in: 2 3 4 4 5 5 5 5 8.

Mode. There are more 5s than any other value. So the mode = 5.

Median. Because there are nine cases, the middle case is $\frac{N+1}{2}$ or $\frac{9+1}{2}=5$. The median is equal to the value of case 5. If we count over four cases, we see that the value of the fifth case is equal to 5. So the median = 5.

Mean. The mean is obtained by summing the value of each case and dividing by the number of cases using the formula $\bar{X}=\frac{\Sigma X}{N}$. Therefore, $\bar{X}=\frac{2+3+4+4+5+5+5+5+8}{9}=\frac{41}{9}=4.6$. So the mean = 4.6.

Example 2: Calculating Measures of Central Tendency Using Raw Data

As global warming becomes an increasing threat to our way of life on this planet, I was curious to learn more about the most fuel-efficient cars on the market. Visiting the U.S. Department of Energy website (http://www.fueleconomy.gov), I learned that the most fuel-efficient midsize cars, large cars, and midsize station wagons on the market for 2009 were the Toyota Prius, Nissan Versa, Hyundai Sonata, Honda Accord, Volkswagen Jetta SportWagen, Kia Rondo, and Saab 9-5 SportCombi. They are listed along with the city and highway gas mileage ratings in Table 3.14.

Table 3.14 tells us which cars to consider if we are looking for a fuel-efficient family sedan; however, it does not tell us how fuel efficiency has changed over time. In other words, we can use Table 3.14 to shop for the most fuel-efficient vehicle, but we need more data to determine whether our choices will make any difference over time.

To investigate this a little further, I searched for all the family sedans that get more than 20 mpg for 1994 and 2009. The results of this search are presented in Table 3.15.

Mode. For the model year 1994, the most frequently occurring value is 21 mpg ($f = 6$). Therefore, the mode for 1994 is equal to 21 mpg.

For the model year 2009, the most frequently occurring values are 21 and 22. For each, $f = 6$. Because there are two modes, we call this a bimodal distribution. Comparing the mode from 1994 with that from 2009, we see very little difference in fuel efficiency (21 mpg vs. 21 or 22 mpg).

TABLE 3.14 City and Highway Mileage Ratings in 2009

Make and Model	City (mpg)	Highway (mpg)
Toyota Prius	48	45
Nissan Versa	26	31
Hyundai Sonata	22	32
Honda Accord	22	31
Volkswagen Jetta SportWagen	30	41
Kia Rondo	20	27
Saab 9-5 SportCombi	18	27

Source: Data from the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy.

TABLE 3.15 Fuel Efficiency for Family Sedans (>20 mpg), 1994 and 2009

1994			2009		
Make and Model	City (mpg)	Highway (mpg)	Make and Model	City (mpg)	Highway (mpg)
Mazda 626	23	31	Toyota Prius (hybrid)	48	45
Honda Accord	22	29	Nissan Altima (hybrid)	35	33
Chevrolet Corsica	22	28	Toyota Camry (hybrid)	33	34
Buick Century	22	28	Chevrolet Malibu (hybrid)	26	34
Oldsmobile Cutlass Ciera	22	28	Saturn Aura (hybrid)	26	34
Oldsmobile Achieva	21	32	Hyundai Elantra	25	33
Pontiac Grand Am	21	32	Kia Spectra	24	32
Infiniti G20	21	29	Nissan Altima	23	32
Mitsubishi Galant	21	28	Saturn Aura	22	33
Dodge Spirit	21	27	Kia Optima	22	32
Plymouth Acclaim	21	27	Hyundai Sonata	22	32
Subaru Legacy	20	28	Honda Accord	22	31
Toyota Camry	20	27	Chevrolet Malibu	22	30
Hyundai Sonata	19	26	Toyota Camry	21	31
Chrysler LeBaron	19	25	Volkswagen Passat	21	31
Ford Taurus	18	27	Mazda 6	21	30
Mercury Sable	18	27	Chrysler Sebring	21	30
Eagle Vision	18	26	Dodge Avenger	21	30
			Ford Fusion	20	29
			Mercury Milan	20	29
			Mitsubishi Galant	20	27
			Subaru Legacy	20	27
			Nissan Maxima	19	26
			Nissan Altima	19	26
			Mercury Sable	18	28
			Hyundai Azera	18	26
			Buick LaCrosse/Allure	17	28

Source: Data from the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy.

To calculate the three measures of central tendency, let's assume that we want to compare city driving conditions in 1994 with those in 2009.

Median. As you can see, the city mpg is already ranked from high to low in Table 3.14. Therefore, we need only to determine the middle case using the following formula:

$$Position\ of\ the\ median = \frac{N+1}{2}$$

Once we have determined the position of the median, we determine the value of the median by reading the mileage for that particular automobile. For the 1994 model year,

$$Position\ of\ the\ median = \frac{18+1}{2} = \frac{19}{2} = 9.5$$

When we count down 9.5 cases, we see that the median is right between the mileage for the Mitsubishi Galant (21 mpg) and that for the Dodge Spirit (also 21 mpg). Therefore, the median mpg for these 1994 family sedans is 21 mpg.

For the 2009 model year,

$$Position\ of\ the\ median = \frac{27+1}{2} = \frac{28}{2} = 14$$

Therefore, when we count down to the 14th case, we find that the median is equal to the mpg of the Toyota Camry. In this case, the city mpg of the 2009 Toyota Camry is 21. Therefore, the median mpg for these 2009 family sedans is 21 mpg.

So far, our comparison of modes and means offers little hope of finding overall increases in fuel efficiency between 1994 and 2009. A comparison of the means, however, may tell a different story.

Mean. To determine the average mpg using our formula for the mean $\bar{X} = \frac{\Sigma X}{N}$, we must determine (1) the number of cases (N) and (2) the sum of all the different cars'

© iStockphoto.com/Tramino

mpg values ($\sum X$). For the model year 1994, $N = 18$ and $\sum X = 369$. Therefore, our formula for the mean is equal to

$$\bar{X} = \frac{23+22+22+22+22+21+21+21+21+21+21+20+20+19+19+18+18+18}{18}$$

$$= \frac{369}{18} = 20.5$$

This means that, on average, 1994 model year family sedans got 20.5 mpg.

Doing the same for the 2009 model year, we find that $N = 27$ and $\sum X = 626$. Therefore, our formula for the mean is equal to

$$\bar{X} = \frac{\begin{array}{l}48+35+33+26+26+25+24+23+22+22+22+22+22+\\21+21+21+21+21+20+20+20+20+19+19+18+18+17\end{array}}{18}$$

$$= \frac{626}{27} = 23.2$$

Therefore, between 1994 and 2009, the average mpg of family sedans increased from 20.5 to 23.2, a 2.7-mpg (13%) increase. A significant portion of this increase is the result of hybrid technology. In fact, when the hybrid cars are removed from the 2009 model year data, the average mpg is 20.8, a .3-mpg (1%) increase over 1994 levels. This modest increase may indicate that internal combustion engines might not have been any more fuel efficient than they were 15 years earlier and that gains in fuel efficiency were due mainly to a whole new type of engine—the hybrid engine.

The statistics we calculated here represent only city driving fuel efficiency. Using the data provided, we will now generate frequency tables and calculate these statistics for highway driving.

Example 3: Calculating Measures of Central Tendency Using Grouped Data

Tables 3.16 and 3.17 represent the data for 1994 model year highway driving and 2009 model year highway driving, respectively.

TABLE 3.16 Highway mpg for 1994 Model Year

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	25	1	5.6	5.6	5.6
	26	2	11.1	11.1	16.7
	27	5	27.8	27.8	44.4
	28	5	27.8	27.8	72.2
	29	2	11.1	11.1	83.3
	31	1	5.6	5.6	88.9
	32	2	11.1	11.1	100.0
	Total	18	100.0	100.0	

Source: Data from the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy.

TABLE 3.17 Highway mpg for 2009 Model Year

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	26	3	11.1	11.1	11.1
	27	2	7.4	7.4	18.5
	28	2	7.4	7.4	25.9
	29	2	7.4	7.4	33.3
	30	4	14.8	14.8	48.1
	31	3	11.1	11.1	59.3
	32	4	14.8	14.8	74.1
	33	3	11.1	11.1	85.2
	34	3	11.1	11.1	96.3
	45	1	3.7	3.7	100.0
	Total	27	100.0	100.0	

Mode. For cars built in 1994, you can see from the Frequency column in Table 3.16 that the two largest frequencies are equal to 5 ($f = 5$), cars that get 27 and 28 mpg on the highway. Therefore, the distribution is bimodal and the modes are equal to 27 and 28.

For cars built in 2009, you can see from the Frequency column in Table 3.17 that the two largest frequencies are equal to 4 ($f = 4$), cars that get 30 and 32 mpg on the highway. Again the distribution is bimodal. This means that the modes are equal to 30 and 32 mpg.

Median. Using the Cumulative Percent column in Table 3.16, we see that the 1994 model year cars that get 27 mpg or less make up only 44.4% of the sample. Cars that get 28 mpg make up an additional 27.8% of the sample. Because the 50% mark in the frequency table comes somewhere between cars that get 27 mpg and cars that get 29 mpg, the median is equal to 28 mpg.

Using the Cumulative Percent column in Table 3.17, we see that 2009 model year cars that get 30 mpg or less make up 48.1% of the sample. Cars that get 31 mpg make up an additional 11.2%. Because the 50% mark in the frequency table comes somewhere between cars that get 30 mpg and cars that get 32 mpg, the median is equal to 31 mpg.

TABLE 3.18

Frequency × mpg	ΣX
1 × 25 =	25
2 × 26 =	52
5 × 27 =	135
5 × 28 =	140
2 × 29 =	58
1 × 31 =	31
2 × 32 =	64
$N = 18$	505

TABLE 3.19

Frequency × mpg	ΣX
3 × 26 =	78
2 × 27 =	54
2 × 28 =	56
2 × 29 =	58
4 × 30 =	120
3 × 31 =	93
4 × 32 =	128
3 × 33 =	99
3 × 34 =	102
1 × 45 =	45
N = 27	833

Mean. To calculate the mean mpg for 1994 model year cars, we multiply the frequency of each mpg by the number of cars that get that many miles per gallon. For example, for 1994 only 1 car gets 25 mpg, so 1 × 25 = 25. Two cars get 26 mpg, so 2 × 26 = 52. The equations are in Table 3.18.

Using the formula for the mean,

$$\bar{X} = \frac{\Sigma X}{N} = \frac{505}{18} = 28.1 \text{ mpg}$$

This means that, on average, the 18 cars analyzed for 1994 get 28.1 mpg on the highway.

For 2009, the equations are in Table 3.19.

Using the formula for the mean,

$$\bar{X} = \frac{\Sigma X}{N} = \frac{833}{27} = 30.9 \text{ mpg}$$

This means that, on average, the 27 cars analyzed for 2009 get 30.9 mpg on the highway.

Table 3.20 offers another example of how to calculate the mode, median, and mean with grouped data.

Example 4: Calculating Measures of Central Tendency Using Grouped Data

Mode. In this example, there are 414 respondents with zero children. There are more zeros than any other value, so the mode = 0.

Median. To find the median, we can use two different methods. First, we can find the middle case by taking $\frac{1{,}395+1}{2}$, which gives a value of 698. If we begin adding cases as we move down the Frequency column, we find that after summing all the 0s, we

TABLE 3.20 Number of Children

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	414	29.7	29.7	29.7
	1	242	17.3	17.3	47.0
	2	398	28.5	28.5	75.6
	3	226	16.2	16.2	91.8
	4	115	8.2	8.2	100.0
	Total	1,395	100.0	100.0	

have 414 cases. After summing all the 0s and 1s, we have 656 cases, still not up to the middle case number of 698. If we sum all the 0s, 1s, and 2s, we have 1,054 cases. Therefore, case number 698 came after the 1s but before the end of the 2s. Therefore, the median is equal to 2.

The second method by which to calculate the median in this example is to locate the 50% mark in the Cumulative Percent column. The 50% mark represents case 698 (the middle case). Starting at the top of the Cumulative Percent column, imagine that the percentages from 0 to 100% are all listed. Move down the column until you come to the 50% mark. After including all the 0s, we are up to 29.7%. After including all the 0s and 1s, we are up to 47.0%. After including all the 0s, 1s, and 2s, we are up to 75.6%. Therefore, the 50% mark came after the 1s but before the end of the 2s. Therefore, the median must be equal to 2.

Mean. To calculate the mean, we must begin by determining the total number of children that the 1,395 respondents have. To do this, we use the following technique.

We have 414 cases with a value of 0, so

$$414 \times 0 = 0$$

We have 242 cases with a value of 1, so

$$242 \times 1 = 242$$

We do this for all the values as follows:

$$414 \times 0 = 0$$

$$242 \times 1 = 242$$

$$398 \times 2 = 796$$

$$226 \times 3 = 678$$

$$115 \times 4 = 460$$

$$\text{Total} = 2{,}176$$

We now know that our 1,395 respondents (N = 1,395) have a total of 2,176 children. We can now put these numbers into our formula for the mean:

$$\bar{X} = \frac{\sum X}{N}$$

$$\bar{X} = \frac{\sum X}{N} = \frac{2{,}176}{1{,}395} = 1.6$$

Therefore, on average, each of our respondents has 1.6 children. Of course, it is not possible to have 1.6 children, but means are not always based on whole numbers.

INTEGRATING TECHNOLOGY

Statistical software programs not only help researchers to speed up the process of data analysis but also help to improve the quality of data quickly and easily. Often we want to analyze only a portion of our sample. SPSS for Windows and most other programs include a function that allows users to select only certain cases for analysis.

For example, suppose we are conducting a multicampus nationwide study of injuries associated with alcohol consumption. Our data may indicate that physical injury is two times as likely to occur among college students when alcohol was consumed. Being able to select out particular cases allows us to conduct the same type of analysis across different subgroups of our sample, thereby allowing us to compare males with females, student athletes with nonathletes, and campus-based students with off-campus students. Or we could compare male student athletes who live on campus with female student athletes who live on campus. The ability to apply a common statistical analysis across different groups enables researchers to (1) statistically test hypotheses and (2) build better theoretical models of human behavior by specifying the social contexts in which alcohol-related injuries are most likely to occur.

An article in the *Journal of American College Health* (Yusko, Buckman, White, & Pandina, 2008) indicated that although student athletes consume alcohol less often than nonathletes, they are more likely to binge-drink when they do consume alcohol, perhaps as a result of attempting to "maximize" their relatively limited amount of "party time." These trends are particularly true for male student athletes. This is just one example of a comparative study in which the data were analyzed using a statistical software package.

Statistical software programs provide the possibility of conducting such comparisons in as little as a few minutes or even seconds.

CHAPTER SUMMARY

This chapter focused on measures of central tendency, statistics that are used to describe where cases tend to cluster in a distribution. The three most commonly used measures are the mode, the median, and the mean. While the mode can be used with nominal, ordinal, and interval/ratio variables, the median can be used only with ordinal and interval/ratio variables. The mean can be used only when working with interval/ratio variables. Therefore, each measure makes a unique contribution to our understanding of data. It is important to be able to calculate, understand, and apply all three measures of central tendency using both raw data and grouped data (frequency tables). It is also important to be able to understand how these measures can be used in graphic representations of data.

CHAPTER EXERCISES

Use the dataset below to calculate the mode, median, and mean:

2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 5, 6, 6, 6, 7, 7, 8, 9

1. Mode ____
2. Median ____
3. Mean ____

Use Table 3.21 to calculate the mode, median, and mean.

TABLE 3.21 Number of Days Exercised During Past Two Weeks

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2.00	5	17.2	17.2	17.2
	3.00	8	27.6	27.6	44.8
	4.00	4	13.8	13.8	58.6
	5.00	3	10.3	10.3	69.0
	6.00	4	13.8	13.8	82.8
	7.00	3	10.3	10.3	93.1
	8.00	1	3.4	3.4	96.6
	9.00	1	3.4	3.4	100.0
	Total	29	100.0	100.0	

4. Mode ____
5. Median ____
6. Mean ____

Using Table 3.22, indicate whether or not each measure of central tendency applies. If it does, indicate its value. If not, indicate so by answering "NA" (not applicable).

TABLE 3.22 Blues or R & B Music

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Like very much	204	13.6	14.2	14.2
	Like it	619	41.3	43.2	57.4
	Mixed feelings	348	23.2	24.3	81.7
	Dislike it	206	13.7	14.4	96.0
	Dislike it very much	57	3.8	4.0	100.0
	Total	1,434	95.6	100.0	
Missing	Don't know much about it	54	3.6		
	Not applicable	12	.8		
	Total	66	4.4		
Total		1,500	100.0		

Source: Data from the National Opinion Research Center, General Social Survey.

7. Mode applies. Like it.
8. Median applies. Like it.
9. Mean does not apply. NA.

Using Table 3.23, indicate whether or not each measure of central tendency applies. If it does, indicate its value. If not, indicate so by answering "NA" (not applicable).

TABLE 3.23 How Often Respondent Watches TV Dramas or Sitcoms

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Daily	313	20.9	21.0	21.0
	Several days in week	550	36.7	36.9	57.9
	Several days in month	258	17.2	17.3	75.2
	Rarely	275	18.3	18.5	93.7
	Never	94	6.3	6.3	100.0
	Total	1,490	99.3	100.0	
Missing	Don't know	3	.2		
	Not applicable	7	.5		
	Total	10	.7		
Total		1,500	100.0		

Source: Data from the National Opinion Research Center, General Social Survey.

10. Mode applies. Several days in week.
11. Median applies. Several days in week.
12. Mean does not apply. NA.

IN-CLASS EXERCISES

Use the data below to calculate the mode, median, and mean:

Dataset: 10, 15, 17, 18, 24, 23, 11, 15, 13, 12, 15, 14

1. Mode ___
2. Median ___
3. Mean ___

Use Table 3.24 to calculate the mode, median, and mean.

TABLE 3.24 Number of Courses This Semester

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	5	5.5	5.5	5.5
	2	7	7.7	7.7	13.2
	3	10	11.0	11.0	24.2
	4	16	17.6	17.6	41.8
	5	31	34.1	34.1	75.8
	6	16	17.6	17.6	93.4
	7	6	6.6	6.6	100.0
	Total	91	100.0	100.0	

4. Mode ____
5. Median ____
6. Mean ____

Use Table 3.25 to answer Questions 7 and 8.

TABLE 3.25 Years of Education

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Less than 12 years	781	17.3	17.3	17.3
	12 years	1,204	26.7	26.7	44.0
	More than 12 years	2,525	56.0	56.0	100.0
	Total	4,510	100.0	100.0	

Source: Data from the National Opinion Research Center, General Social Survey.

7. Which measures of central tendency apply to Table 3.25?
8. What is the median number of years of education?

Use Table 3.26 to answer Questions 9 to 12.

TABLE 3.26 Education: Race of Respondent

Race of Respondent			Frequency	Percent	Valid Percent	Cumulative Percent
White	Valid	Less than 12 years	428	13.0	13.0	13.0
		12 years	917	27.9	27.9	41.0
		More than 12 years	1,939	59.0	59.0	100.0
		Total	3,284	100.0	100.0	
Black	Valid	Less than 12 years	137	21.6	21.6	21.6
		12 years	165	26.0	26.0	47.6
		More than 12 years	332	52.4	52.4	100.0
		Total	634	100.0	100.0	
Other	Valid	Less than 12 years	216	36.5	36.5	36.5
		12 years	122	20.6	20.6	57.1
		More than 12 years	254	42.9	42.9	100.0
		Total	592	100.0	100.0	

Source: Data from the National Opinion Research Center, General Social Survey.

9. What is the median level of education for whites?
10. What is the median level of education for others?
11. Which group is most likely to have less than 12 years of education?
12. Which group is most likely to have more than 12 years of education?

Use Table 3.27 to answer Questions 13 to 16.

TABLE 3.27 How Many Sex Partners Respondent Had During Last Year

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No partners	129	25.2	25.2	25.2
	1 partner	321	62.7	62.7	87.9
	2 partners	29	5.7	5.7	93.6
	3 partners	15	2.9	2.9	96.5
	4 partners	13	2.5	2.5	99.0
	5 partners	5	1.0	1.0	100.0
	Total	512	100.0	100.0	

Source: Data from the National Opinion Research Center, General Social Survey.

13. How many attributes does this variable have?
14. What is the modal number of sex partners?
15. What is the median number of sex partners?
16. What is the average number of sex partners?

HOMEWORK ASSIGNMENT

Use Tables 3.28 to 3.31 to answer the questions that follow.

TABLE 3.28 Age of Respondent

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18	19	2.2	2.2	2.2
	19	47	5.5	5.5	7.8
	20	50	5.9	5.9	13.6
	21	52	6.1	6.1	19.8
	22	72	8.5	8.5	28.2
	23	63	7.4	7.4	35.6
	24	61	7.2	7.2	42.8
	25	74	8.7	8.7	51.5
	26	83	9.8	9.8	61.3
	27	86	10.1	10.1	71.4
	28	80	9.4	9.4	80.8
	29	79	9.3	9.3	90.1
	30	84	9.9	9.9	100.0
	Total	850	100.0	100.0	

Source: Data from the National Opinion Research Center, General Social Survey.

For the variable *age:*

1. Is it nominal, ordinal, or interval/ratio?
2. What is the value of the mode?
3. If you can use the median, indicate its value (otherwise use NA [not applicable] for your answer).
4. If you can use the mean, indicate its value (otherwise use NA for your answer).

TABLE 3.29 Subjective Class Identification

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Lower class	33	3.9	5.7	5.7
	Working class	305	35.9	52.6	58.3
	Middle class	228	26.8	39.3	97.6
	Upper class	14	1.6	2.4	100.0
	Total	580	68.2	100.0	
Missing	Not applicable	263	30.9		
	Don't know	7	.8		
	Total	270	31.8		
Total		850	100.0		

Source: Data from the National Opinion Research Center, General Social Survey.

For the variable *class:*

5. Is it nominal, ordinal, or interval/ratio?
6. What is the value of the mode?
7. If you can use the median, indicate its value (otherwise use NA [not applicable] for your answer).
8. If you can use the mean, indicate its value (otherwise use NA for your answer).

TABLE 3.30 Marital Status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Married	247	29.1	29.1	29.1
	Widowed	3	.4	.4	29.4
	Divorced	36	4.2	4.2	33.6
	Separated	14	1.6	1.6	35.3
	Never married	550	64.7	64.7	100.0
	Total	850	100.0	100.0	

Source: Data from the National Opinion Research Center, General Social Survey.

For the variable *marital:*

9. Is it nominal, ordinal, or interval/ratio?
10. What is the value of the mode?
11. If you can use the median, indicate its value (otherwise use NA [not applicable] for your answer).
12. If you can use the mean, indicate its value (otherwise use NA for your answer).

TABLE 3.31 Is Life Exciting or Dull?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Exciting	201	23.6	51.5	51.5
	Routine	175	20.6	44.9	96.4
	Dull	14	1.6	3.6	100.0
	Total	390	45.9	100.0	
Missing	Not applicable	455	53.5		
	Don't know	5	.6		
	Total	460	54.1		
Total		850	100.0		

Source: Data from the National Opinion Research Center, General Social Survey.

For the variable *life:*

13. Is it nominal, ordinal, or interval/ratio?
14. What is the value of the mode?
15. If you can use the median, indicate its value (otherwise use NA [not applicable] for your answer).
16. If you can use the mean, indicate its value (otherwise use NA for your answer).

Use the dataset below to calculate the mode, median, and mean.

Dataset: 4, 7, 3, 5, 8, 6, 5, 4, 6, 7, 6, 3, 6, 4

17. Mode ___
18. Median ____
19. Mean ____
20. Explain what happens to the mode if we add an outlier with a very high value.
21. Explain what happens to the median if we add an outlier with a very high value.
22. Explain what happens to the mean if we add an outlier with a very high value.

KEY TERMS

Average, 82
Histogram, 99
Mean, 86
Measures of central tendency, 82
Median, 86
Mode, 85
Outliers, 99
Skew, 87

NOW YOU TRY IT ANSWERS

#3.1: Mode = 15, Median = 15, Mean = 14.4

#3.2: **M**ode = 1, Median = 2, Mean = 2