

# CHAPTER 3

## Measures of Central Tendency

“ *Welcome to Lake Wobegon, where all the women are strong, all the men are good-looking, and all the children are above average.*

—Garrison Keillor ”

### LEARNING OBJECTIVES

1. Describe the only measure of center appropriate for a nominal-level variable.
2. Identify the difference between the median and the mean.
3. Describe when the median may be a better measure of center compared with the mean.
4. Explain how to calculate and interpret all measures of center from both grouped and ungrouped data.

### 2 Introduction

What do you think about when you think about the average or typical prisoner released from state prisons today? Well, the most recent data available tell us that the typical state prisoner released is more likely to be a male who is 40 years of age or older and has an average of 4.9 prior convictions (Durose, Cooper, & Snyder, 2014). We have conveyed a lot of information with concepts such as “typical” and “average” in this sentence, but in this chapter you are going to learn more precise statistical concepts used to describe the most typical quality or value of a variable. In Chapter 2, we learned to describe our data using frequency distributions and graphical displays. These pieces of information are important, but they should be combined with summary statistics

that also help to describe our variable distribution. In this chapter, you will learn about summary statistics called **measures of central tendency**. Think of the two key words in this term and what they connote—central tendency—a tendency to be at the center of something, in this case the center of data. Measures of central tendency capture the “typical,” “average,” or “most likely” score or value in a distribution of scores like the 40-year-old male with 4.9 prior convictions typical state prisoner above.

We will discuss three different measures of central tendency in this chapter: the mode, the median, and the mean. Each measure captures a somewhat different notion of “central tendency,” and you should not be surprised to learn that each requires a certain level of measurement.

## 2 The Mode

The **mode** is one measure of central tendency. The mode conceptualizes “central tendency” in terms of what is the *most likely, most common, or most frequent* score in a distribution of scores. The mode can be calculated with data measured at the nominal, ordinal, or interval/ratio level. However, if you have nominal- or purely ordinal-level data (purely ordinal in the sense that the data are not continuous data that you have made ordinal by making class intervals or grouping your data), then the mode is the *only* appropriate measure of central tendency that you may legitimately use. If the data are in numerical or tabular form, the mode can be easily identified by finding the score or value in a distribution that has (a) the greatest frequency, (b) the largest proportion, or (c) the highest percentage. If the data are in graphical form, the mode can be identified by finding the score or value in the graph that has (a) the largest slice in a pie chart, (b) the longest bar in a bar chart, or (c) the highest bar in a histogram. Thus, the way the mode “interprets” central tendency is that it is the most likely or probable or the most frequent score or value in a distribution of values.

### Case Study

#### The Modal Category of Hate Crime

Let’s go through a couple of examples. In Chapter 2, we presented data that showed the distribution of different kinds of hate crimes that were reported to the police in the year 2013 (Table 2.1); these data are reproduced in Table 3.1. A hate crime is defined as one that is intended to hurt and intimidate someone because of his or her race, ethnicity, national origin, religion, sexual preference, or disability. As you can see, there were 5,922 single-bias hate crime incidents reported to the Federal Bureau of Investigation (FBI) Uniform Crime Reports program that year that fell into one of five distinct types based on the motivation or the type of hate that precipitated the crime. The variable “reported hate crime” is measured at the nominal level because the only distinction among the values of this variable are qualitative distinctions of “kind”—a hate crime driven by racial hatred is simply different from one driven by religious hatred.

Looking at the distribution of scores in this table, we can discern that the most frequent type of hate crime in 2013, or the modal hate crime, was one motivated by racial hostility. We would conclude, therefore, that the mode for this variable is “racially motivated hate crime.” There are a number of different ways we could come to this conclusion, each of which would converge on the same answer. First, we could look at the reported frequencies and note that the frequency of racial hate crimes is clearly greater than the frequency for all other kinds of reported hate crimes. Second, we could look

Get the edge on your studies.

eLearning:

- Take a quiz to find out what you’ve learned.
- Review key terms with eFlashcards.
- Explore additional data sets.

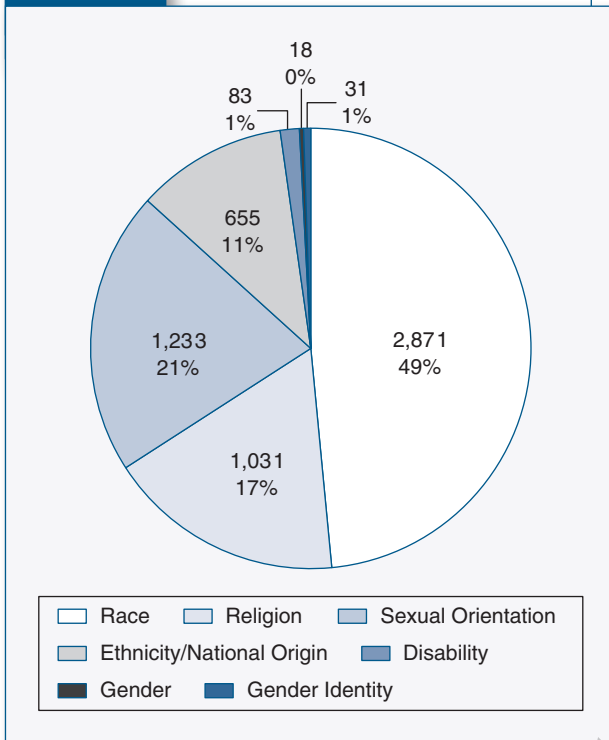
 SAGE edge™

#### Measures of central tendency:

Summary statistics that capture the “typical,” “average,” or “most likely” scores or values in variable distributions.

**Mode:** Value of a variable that occurs more often than any other value.

**Figure 3.1** Types of Hate Crime Incidents Reported to Police in 2013



**Table 3.1** Types of Hate Crime Incidents Reported to Police in 2013

Basis of Hate	f	Proportion	%
Race	2,871	.485	48.5
Religion	1,031	.174	17.4
Sexual orientation	1,233	.208	20.8
Ethnicity/National origin	655	.111	11.1
Disability	83	.014	1.4
Gender	18	.003	0.3
Gender identity	31	.005	0.5
Total	5,922	1.000	100.0

Source: Adapted from *Hate Crimes Statistics—2013* from the Federal Bureau of Investigation (2013b).

at the column of proportions, find that nearly one half (.485) of all hate crimes that were reported were racially motivated, and note that this proportion is greater than the proportion for any other kind of hate crime. Third, we could examine the row of percentages, find that 48.5% of all hate crimes in 2013 were racially motivated hate crimes, and note that this percentage is greater than the percentage for any other type of hate crime. Finally, we could use the information we have about proportions to determine the probability of each type of hate crime and then draw a conclusion about what the mode is. Since the proportion or relative frequency of a value/score can also be understood as its probability of occurring, we can see that if we were to select randomly 1 out of the 5,922 hate crimes in 2013, the probability that it would be a racially motivated hate crime would be .485, the probability that it would be motivated by religious prejudice would be .174, the probability of a sexually motivated hate crime would be .208, and so on. The greatest probability event, therefore, is a racial hate crime, a probability that exceeds those of all other possible outcomes. All of our different ways to capture the mode tell us that the modal type of hate crime was racially motivated hate crime.

Another way to determine what the mode is for a nominal- or ordinal-level variable is to examine the graph of the frequency data (or the graphed proportions or percentages). In Figure 3.1, we show the pie chart of the data in Table 3.1 with both the frequency and the percentage of each value. Note that the largest slice in the pie is for the value “race hate crime.” This is the modal hate crime for 2013.

The modal type of hate crime reported in 2013, then, is a racially motivated hate crime. Note that the mode is the *value* or *score* that is most frequent or most likely, not the actual numerical value of the frequency, proportion, or percentage. The mode for the variable “reported hate crime in the year 2013” is “racially motivated hate crime.” The mode is not 2,871 or .485 or even 48.5%. The mistake that students most frequently make when they are first learning statistics is that they conclude that the mode is some frequency, proportion, or percentage rather than the value of a given variable. To avoid making this mistake, just remember that the mode is the *value*, *score*, or *outcome* of a variable that is most likely or frequent, not the actual frequency of that value.

## Case Study

### The Modal Number of Prior Arrests

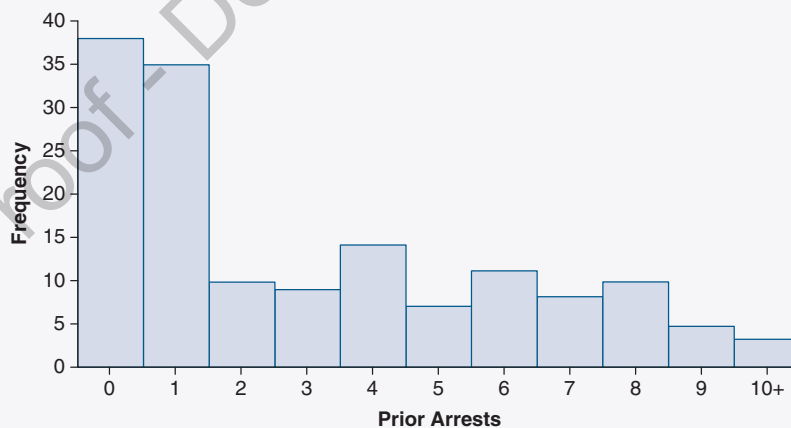
In Table 3.2, we report the frequency distribution (and percentages) of the number of prior arrests for a hypothetical sample of 150 armed robbery suspects. The data are in the form of an ungrouped frequency distribution, and the variable “number of prior arrests” is measured at the interval/ratio level. The histogram for the frequency data is shown in Figure 3.2. Note in the table that there are two values that are more frequent than all of the others: “0 prior arrests” and “1 prior arrest.” The frequencies for these values are very comparable and much greater than any of the other values. This corresponds to the height of the two largest rectangular bars in the histogram (Figure 3.3) for 0 and 1 prior arrest. Even though the frequencies for 0 and 1 prior arrest are not exactly equal, they are very comparable, and their frequencies are much greater than those of any other value. They are comparable enough that it might be misleading to say that there is one and only one distinct mode in these data. It would appear more appropriate, then, that for this variable there are two distinct modes: a mode of 0 prior arrests and a mode of 1 prior arrest. Because there are two modes in the data, this distribution has a bimodal distribution. A **bimodal distribution** is a distribution that has two distinct values with the greatest frequency or the largest probability of occurring even if their frequencies are not

**Table 3.2** Number of Prior Arrests for a Sample of Armed Robbery Suspects

Number	<i>f</i>	%
0	38	25.33
1	35	23.33
2	10	6.67
3	9	6.00
4	14	9.33
5	7	4.67
6	11	7.33
7	8	5.33
8	10	6.67
9	5	3.33
10 or more	3	2.00
Total	<i>n</i> = 150	99.99*

\*Percentages may not sum to 100% due to rounding

**Figure 3.2** Number of Prior Arrests Among 150 Suspected Armed Robbers



exactly equal. It tells us that there are two scores that are roughly the most typical or most likely scores in the distribution.

The strategy for identifying the mode when the data are in the form of a grouped frequency distribution is pretty much the same as what we have just discussed. Table 3.3

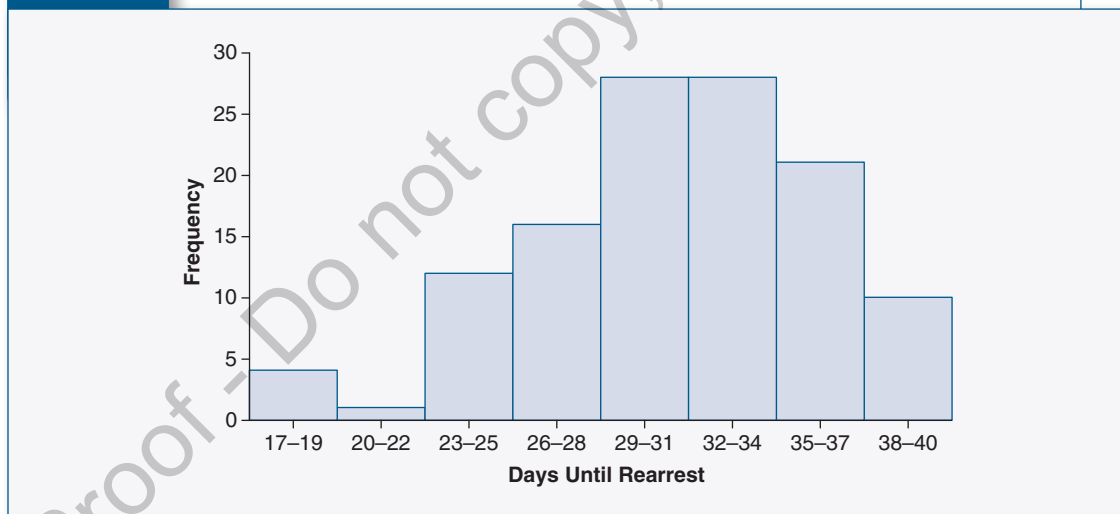
**Bimodal distribution:** Distribution that contains two distinct modes with the greatest frequency even if the frequencies are not exactly equal.

**Table 3.3** Grouped Frequency Distribution for Time-Until-Rearrest Data for 120 Released Offenders

Stated Limits (Days)	<i>f</i>	Midpoint
17–19	4	18
20–22	1	21
23–25	12	24
26–28	16	27
29–31	28	30
32–34	28	33
35–37	21	36
38–40	10	39
	<i>n</i> = 120	

provides the grouped frequency distribution for the variable “time until rearrest” for a sample of 120 male inmates who were released from prison and who were rearrested at some point after release. We first presented these data in the previous chapter. Recall that each person’s “score” reflects the number of days he was out in the community after being released before he committed a new offense. The histogram for these data is shown in Figure 3.3, where the class intervals are shown on the *x* axis. When we look at this frequency distribution, it is apparent that these are bimodal data. The two modes are represented by the class intervals 29–31 days and 32–34 days. For both of these class intervals, the frequency is higher than for any other class interval. Exactly 28 persons were arrested between 19 and 31 days after release and 28 persons were rearrested between 32 and 34 days (very close to 1 month) after their release from prison. The two modes also can be seen from the histogram as the two highest peaks in Figure 3.3.

**Figure 3.3** Histogram of Grouped Frequency Data for Time Until Rearrest



### Advantages and Disadvantages of the Mode

As a measure of central tendency, the mode has both advantages and disadvantages relative to other measures. One advantage of the mode is that it is very simple to determine and is appealing conceptually. It is the score or value that is the most frequent and that has the greatest probability of occurring in a distribution of scores. This simple elegance means that the mode is very easy for readers to comprehend and understand, which is a key quality that statistical measures should have. The mode is also very simple to “calculate.” In fact, there is no real arithmetic calculation to determine the mode—nothing to add or subtract. We just identify the score with the greatest frequency (or proportion or percentage) in either a tabular presentation of data or a graph (find the tallest or longest bar or the largest slice of pie). Finally, the mode is a very general measure of central tendency since it can be determined for variables measured at any level. The mode is an appropriate measure of central tendency for data measured at the nominal, ordinal, and interval/ratio levels.

Copyright ©2017 by SAGE Publications, Inc.

The simplicity of understanding and determining the mode is offset by its disadvantage. Since the mode is based only on the most frequent score or scores, it does not take into account all or even most of the information available in a distribution. One thing statisticians do not like to do is ignore data or information, but that is exactly what the mode does; it ignores all information in the data except the values/scores with the highest frequency. By ignoring or throwing out information, the mode may at times give us a very misleading notion of the central tendency of our data. For example, Table 3.4 shows the number of subsequent charges of domestic violence accumulated over a 1-year period by a sample of 60 men who had been arrested for intimate partner assault. According to the table, the mode would be “0 new charges since arrest” since this value has the greatest frequency. Note, however, that the ease of identifying “0 new charges” as the mode comes at the price of ignoring the fact that many of the men in this sample *did* have numerous new charges of being abusive toward their partner after they were first arrested. Although it is technically correct that the modal number of new charges is zero, this is somewhat misleading and does not represent the entire distribution of values. Because of this deficiency with the mode, when we have interval/ratio-level data, we most often use an alternative measure of central tendency such as the median or the mean.

Table 3.4

**Number of New Charges for Domestic Violence for 60 Men Arrested for Domestic Abuse**

Number of New Charges	<i>f</i>
0	14
1	7
2	5
3	8
4	6
5	4
6	3
7	3
8	5
9	3
10 or more	2
	<i>n</i> = 60

## 2 The Median

The **median** is an appropriate measure of central tendency for quantitative data measured at the interval/ratio level or for data that may have originally been measured at the interval/ratio level but now consist of grouped data (class intervals for grouped frequency distributions that have real limits). The easiest way to think of the median is that it is the value that is at the 50th percentile in a rank-ordered distribution of scores. The 50th percentile is also known as the second quartile when the data are divided into four quartiles (one quartile is equal to 25 percentiles). The median score, in other words, is the score in the exact middle of a rank-ordered distribution of quantitative scores such that the median is the point above which one half of the scores are and below which the other one half of the scores fall.

When the data are continuous (not grouped), the median value is very easy to find by following these two steps:

**Median:** Score at the 50th percentile in a rank-ordered distribution of scores. Thus, one half of a variable's values are less than the median and one half are greater than the median.

### *Steps to Find the Median*

**Step 1:** Rank-order all scores from lowest score to highest score.

**Step 2:** Find the position of the score (*x*) that is the median score by the following formula: Median position =  $(n + 1) / 2$ . This formula says that the *position* of the median score is found by adding 1 to the total number of scores and then dividing by 2. *This formula will not give you the value of the median but will give you the position of the median score.* To find the value of the median, find the score in the position indicated by the formula in the rank-ordered array of scores.

When there are an odd number of scores, this formula is very easy to use.

Copyright ©2017 by SAGE Publications, Inc.

This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.



## Case Study

### The Median Police Response Time and Vandalism Offending

Rank	Score
1	1 minute
2	2 minutes
3	3 minutes
4	6 minutes
5	9 minutes
6	12 minutes
7	15 minutes

Rank	Score
1	1 minute
2	2 minutes
3	3 minutes
4	6 minutes
5	9 minutes
6	12 minutes
7	15 minutes
8	18 minutes

Let's say that we have seven scores that represent the number of minutes it takes the police to respond to a "911" call for service: 9, 1, 3, 6, 12, 2, and 15. To find the median number of minutes it takes the police to respond, step 1 instructs us first to rank-order the scores from low to high:

Then we find the position of the median with our positional locator,  $(n + 1) / 2$ , which in this case is  $(7 + 1) / 2 = 8 / 2 = 4$ th position. The median score is in the fourth position in our rank-ordered scores. Again, we emphasize that the median is not 4, but it falls in the fourth position of our rank-ordered scores. The scores, therefore, must all be put in rank order before you can find the median. To find the value of the median, find the score in the fourth position. We can do this either by starting at the top of the rank-ordered scores (the lowest score) and counting down until we find the fourth score or by starting at the bottom (the highest score) and counting up until we find the fourth score. The result is the same; the score in the fourth position of our rank-ordered scores is 6 minutes. The median amount of time it took to respond to a 911 call for service, therefore, is 6 minutes. Note that exactly one half of the scores in this distribution are lower (1, 2, and 3 minutes) and exactly one half are higher (9, 12, and 15 minutes). The median score, then, sits in the exact middle of the distribution of rank-ordered scores. This is the way the median "interprets" central tendency—it is a positional measure. We can say that 50% of the scores in the distribution of police response time fall below 6 minutes and 50% fall above this value.

Now let's add one more call for service to these data. In this case, it took the police 18 minutes to respond—the longest time so far. We now have a total of eight scores, and the rank order of these eight scores is as follows:

When we use our positional locator formula for the median, we find that the position of the median is  $(8 + 1) / 2 = 9 / 2 = 4.5$ th score (notice that we now have eight data points, so the numerator in our formula is 8). What does a position of 4.5 mean? It means that the median score is the score that is at the midpoint between the fourth and fifth scores in our rank-ordered distribution of scores. Since our fourth score (starting from the lowest score) is 6 minutes and our fifth score (again from the lowest score) is 9 minutes, the median score is the midpoint between these scores. Had we found the fourth and fifth scores by starting at the bottom or highest score and counted up, the two scores would still have been 9 and 6. To find the midpoint between our two scores, we have to add the two scores and then divide by 2. The midpoint between 6 minutes and 9 minutes is  $(6 + 9) / 2 = 15 / 2 = 7.5$  minutes. The median number of minutes it took the police to respond in this second set of scores, then, is 7.5 minutes. Note that one half of the scores are greater than this time and one half are less. The median measures central tendency as the score in the middle in a set of rank-ordered scores, or 50th percentile. That is what the median means.

Another way to identify the median in a set of continuous scores is to find the 50th percentile in a cumulative percentage distribution (you already learned how to make a cumulative percentage distribution, so this should be good practice). Table 3.5 reports the distribution of scores for a variable called "number of times committing vandalism" for a sample of 77 boys; the values range from 0 times to 10 or more times. The first column in Table 3.5 reports the value or score (the number of times a boy reported committing an act of vandalism), the second column shows the frequency for each value, the third column reports the cumulative frequency, the fourth column is the percentage for each value, and the fifth and final column compiles the cumulative percentages.

There are two ways to find the median number of acts of vandalism in this distribution. One is to use the formula for the position of the median we just learned. Since we have  $n = 77$  total scores, the median is in the  $(77 + 1) / 2 = 78$

$/2 = 39$ th position. To find the score at the 39th position, all we have to do is use the column of cumulative frequencies. We can see from this column that 30 scores are at the value of 2 or lower and that 41 scores are at the value of 3 or lower. If the 30th score is a 2, and that is the last 2 in the distribution, we have to go to the next value to find the next scores. This means that the 31st score is a 3, the 32nd score is a 3, ..., the 39th score is a 3, and the 40th and 41st scores as well (the 42nd score is a 4). Since the median score is the 39th score, the median must be “3 acts of vandalism.” We could also have discovered this by looking at the column of cumulative percentages. Since the median is the 50th percentile, all we have to do is find the score at the 50th percentile. We can see that values of “2 acts of vandalism” are at the 39th percentile and that values of “3 acts of vandalism” go from the 40th to the 53rd percentile—the 50th percentile is contained here. The 50th percentile, then, is at the score of “3 acts of vandalism.” In words, then, 50% of the boys in the sample committed 3 or fewer acts of vandalism and 50% of the boys committed 3 or more acts of vandalism.

**Table 3.5** Reported Number of Times Committing Vandalism for 77 Boys

# of Times	<i>f</i>	<i>cf</i>	%	<i>c%</i>
0	15	15	19	19
1	10	25	13	32
2	5	30	7	39
3	11	41	14	53
4	7	48	9	62
5	8	56	10	72
6	5	61	7	79
7	4	65	5	84
8	5	70	7	91
9	4	74	5	96
10 or more	3	77	4	100
Total	$n = 77$		100	

## The Median for Grouped Data

What do we do when we have grouped data and want to find the median score? Things get a little more complicated, but with a formula and a little work, we can determine the median with grouped data as well. Table 3.6 reports the grouped frequency distribution for the example of 120 prison inmates who had served their sentences and were released into the community for whom we have the number of days they were “free” before they were rearrested. Note that this table reports the real limits of the class intervals. As you will see, you need to know the real limits when calculating the median for grouped data. The presence of real limits tells us that these ordinal-level data were once measured at the interval or ratio level, permitting us to calculate a median. The procedure for determining the value of the median for these grouped data is comparable to that for ungrouped data.

First, we have to rank-order the values of the variable. In Table 3.6, the values of the variable “time until rearrest” consist of class intervals, and they are already rank-ordered from low to high. Now that the class intervals are rank-ordered, we need to find the interval that contains the median. We can use our positional locator  $(n + 1) / 2$  and the column of cumulative frequencies to find the interval that contains the median. Since  $n = 120$ , the median is in the  $(120 + 1) / 2 = 121 / 2 = 60.5$ th position, or the score that is at the midpoint between the 60th and 61st scores. We now have to locate the class interval that has the median. The class interval 26–28 days contains the 18th score to the 33rd score, according to the column of cumulative frequencies. The 34th score to the 61st score, then, can be found in the class interval

**Table 3.6** Grouped Frequency Distribution for Time-Until-Rearrest Data for 120 Inmates

Stated Limits	Real Limits	<i>f</i>	<i>cf</i>
17–19 days	16.5–19.5 days	4	4
20–22 days	19.5–22.5 days	1	5
23–25 days	22.5–25.5 days	12	17
26–28 days	25.5–28.5 days	16	33
29–31 days	28.5–31.5 days	28	61
32–34 days	31.5–34.5 days	28	89
35–37 days	34.5–37.5 days	21	110
38–40 days	37.5–40.5 days	10	120
		$n = 120$	



29–31 days (see the column of cumulative frequencies). This is the interval that contains the median since the median score is the midpoint between the 60th and 61st scores and the 60th and 61st scores are at the end of that interval. With this information, we are now ready to calculate the actual value of the median. We can determine the value of the median (and not just the location in this case) with the following formula:

$$X_{\text{median}} = L + \left( \frac{\left( \frac{n+1}{2} \right) - cf}{f} \right) w_i \quad (3-1)$$

where

$X_{\text{median}}$  = the value of the median

$L$  = the lower real limit of the class interval that contains the median

$cf$  = the cumulative frequency of the class interval just before the class interval that contains the median

$f$  = the frequency of the interval that contains the median

$w_i$  = the width of the class interval

$n$  = the total number of observations in the sample

Now let's calculate what the median number of days until rearrest is in these data:

$$\begin{aligned} X_{\text{median}} &= L + \left( \frac{\left( \frac{n+1}{2} \right) - cf}{f} \right) w_i \\ X_{\text{median}} &= 28.5 + \left( \frac{\left( \frac{120+1}{2} \right) - 33}{28} \right) 3 \\ X_{\text{median}} &= 28.5 + \left( \frac{60.5 - 33}{28} \right) 3 \\ X_{\text{median}} &= 28.5 + (.98)3 \\ X_{\text{median}} &= 28.5 + 2.94 \\ X_{\text{median}} &= 31.44 \text{ days until rearrest} \end{aligned}$$

The median number of days until rearrest, then, is 31.44 days. In this distribution of grouped data, we can say that one half of the inmates were rearrested within 31.44 days, or about 1 month, and one half were arrested at 31.44 days or later. Remember that all of the people in this sample were eventually arrested.

### Advantages and Disadvantages of the Median

The median has a number of advantages as a measure of central tendency. First, unlike the mode that can have more than one value, there will always be only one median. Second, as the score in the exact middle of a rank-ordered distribution of scores, the median value has intuitive appeal—it is easy to understand. For example, if when you took the ACT or SAT before college you found out that you scored at the 50th percentile on the test, you knew that you were in the middle of the rank-ordered scores, that one half of the students taking the test at the same time you did scored

higher than you and one half scored lower than you. Third, the median is a useful measure of central tendency that is used in some graphical displays of data. Finally, because the median does not use all of the scores in our data, it is not influenced by extremely high or extremely low scores. As we learned in the last chapter, extremely high or low scores in a distribution are referred to as outliers. Since the median locates the score in the middle of the distribution, or at the 50th percentile, it does not matter whether there are outliers in the data. Let's explain.

Table 3.7 records three columns of data. Each column represents the rate of forcible rape per 100,000 people for a sample of U.S. cities in 2013. There are seven cities in the first column, and the rape rates have already been rank-ordered. The position of the median in these data is the 4th score, starting at either the lowest or highest score:

$$\left( \frac{7+1}{2} = 4 \right)$$

The median rape rate for these seven cities is 28.1 rapes per 100,000. In the next column of cities, we simply add one more city, Anchorage, Alaska, with a rape rate of 133.2 per 100,000. This is an extremely high rape rate, and adding it to this list of seven cities makes it a high outlier. What happens to the value of the median? Well, with eight cases now in the distribution, the median score is the midpoint between the 4th and 5th scores:

$$\left( \frac{8+1}{2} = 4.5 \right)$$

As such, we have to take the average of the scores in the 4th and 5th positions, which gives us:

$$\left( \frac{28.1+28.4}{2} = 28.25 \right) \text{ rapes per 100,000}$$

As you can clearly see, adding this very high outlier did not change the median rape rate much at all, only from 28.1 rapes per 100,000 to 28.25 rapes per 100,000. Despite the outlier, the median still gives a very accurate assessment of the central tendency of rape rates in these data. The median is sturdy or robust in the presence of a high outlier.

In the next set of cities, we remove Anchorage and substitute Goldsboro, North Carolina, which in 2013 had a rate of forcible rape of only 4.0 per 100,000, one of the lowest rape rates of all major U.S. cities (since these are rape rates, or the number of rapes per population, it does not matter that Goldsboro is a small town; Goldsboro has a population that is greater than that of Redmond, Oregon). The rape rate for Goldsboro is an example of a low outlier. We once again find the median for these eight scores as the midpoint between the 4th and 5th scores:

$$\frac{28.0+28.1}{2} = 28.05 \text{ rapes per 100,000}$$

By comparing the three medians we calculated with seven cities, the median rape rate was 28.1 per 100,000. When we added Anchorage with a high outlier, the median rate was 28.25. And when we added Goldsboro with a low outlier, the median rape rate was 28.05. In each case, the measure of central tendency tells us that the median rape rate is around 28 per 100,000 population. Even when there are outlying scores, then, the median is a very stable measure of central tendency because it is defined as the 50th percentile and does not take the value of each and every score into account. This is an important advantage of the median as a measure of central tendency. One disadvantage of the median is that, like the mode, it uses only one or two pieces of information. The next measure of central tendency we will discuss, the mean, uses all the information in the data to determine its central tendency.

**Table 3.7** Rape Rates (per 100,000 People) for Selected U.S. Cities in 2013

Rank	City	Rate	Rank	City	Rate	Rank	City	Rate
1	Binghamton, NY	22.2	1	Binghamton, NY	22.2	1	Goldsboro, NC	4.0
2	Albany, GA	23.5	2	Albany, GA	23.5	2	Binghamton, NY	22.2
3	Redmond, OR	28.0	3	Redmond, OR	28.0	3	Albany, GA	23.5
4	Cedar Rapids, IA	28.1	4	Cedar Rapids, IA	28.1	4	Redmond, OR	28.0
5	Charleston, SC	28.4	5	Charleston, SC	28.4	5	Cedar Rapids, IA	28.1
6	Boston, MA	33.8	6	Boston, MA	33.8	6	Charleston, SC	28.4
7	Akron, OH	38.4	7	Akron, OH	38.4	7	Boston, MA	33.8
			8	Anchorage, AK	133.2	8	Akron, OH	38.4

Source: Adapted from *Crime In the United States* from the Federal Bureau of Investigation (2013a).

## 2 The Mean

The third and final measure of central tendency that we will examine is the mean. Like the median, the mean requires that the data be measured at the interval/ratio level. However, it too can be calculated if you have ordinal data in the form of a grouped frequency distribution where you have taken continuous data and created class intervals.

### Case Study

#### Calculating the Mean Time Until Rearrest

The **mean** is defined as the arithmetic average of a group of scores and is calculated by summing all of the scores and then dividing by the total number of scores. You are already very familiar with the mean. Your college grade point average (GPA) is a mean. For example, suppose you took five classes last semester and earned two A's, a B, a C, and one D (in math, of course). Let's assume that your college assigns 4.0 for an A grade, 3.0 for a B grade, 2.0 for a C grade, and 1.0 for a D grade. Your GPA for last semester, then, would be  $(4 + 4 + 3 + 2 + 1) / 5 = 14 / 5 = 2.8$ . This is the mean grade you received in all five of your classes. Your average would be almost a B, reflecting the fact that you did get two A's but also received a C and a D.

Before we get some practice in calculating the mean, we need to distinguish between the mean of a population and the mean of a sample. Recall from our discussion in Chapter 1 that a population consists of the universe of cases we are interested in. For example, if we are interested in the relationship between IQ scores and delinquency among male youths between the ages of 12 and 18 years in the United States, then our population would consist of all male youths between the ages of 12 and 18 who reside in the United States. The population we are interested in, then, is generally very large and both difficult and costly to study directly. Our population of male adolescents, for example, would number in the millions. A sample, you will remember, is a subset of the population. We select a sample from the population and study the sample with the intention of making an inference to the population based on what we know about the sample. The sample is much smaller than the population.

It would be possible (although it would involve a great deal of work and money) to calculate the mean of a population. The mean of a population, therefore, is unknown but knowable. In statistics, the mean of a population is denoted by the symbol  $\mu$  (the Greek letter mu) and is defined as the sum of all scores in the population divided by the total number of observations in the population:

Copyright ©2017 by SAGE Publications, Inc.

**Mean:** Arithmetic average of a group of scores calculated as the sum of the scores divided by the total number of scores. The mean is an appropriate measure of central tendency for interval/ratio-level data.

$$\mu = \frac{\sum_{i=1}^N X_i}{N} \quad (3-2)$$

where

$X_i$  = each  $X$  score in the population  
 $N$  = the total number of observations in the population

To calculate the population mean, therefore, sum all scores in the population, starting with the first and ending with the last or  $N$ th, and then divide by the total number of observations ( $N$ ). Note that the mean takes all scores into account since we have to sum all scores before calculating the mean.

When we have a sample and wish to calculate the mean of the sample, the formula we use is slightly different:

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} \quad (3-3)$$

where

$\bar{X}$  = the symbol used for the sample mean and is pronounced “ $x$  bar”  
 $x_i$  = the  $i$ th raw score in a distribution of sample scores  
 $\sum_{i=1}^n x_i$  = the instruction to sum all  $x_i$  scores, starting with the first score ( $i = 1$ ) and continuing until the last score ( $i = n$ )  
 $n$  = the total number of scores

This formula is telling you that to calculate the sample mean, you begin by summing all of the scores in your sample, starting with the first score and ending with the last or  $n$ th score, and then divide this sum by the total number of scores in your sample. For example, if you had a distribution of 10 scores, you would calculate the mean of those scores by taking the sum of all 10 scores and then dividing by 10:  $\bar{X} = (x_1 + x_2 + x_3 + \dots + x_{10}) / 10$ . Unlike the median, the mean is not a positional measure of central tendency. Since the mean takes into account all of your scores, you do not need to rank-order them beforehand; you can simply start summing numbers from the very first score.

As an example of calculating the mean, let's begin by calculating the mean rape rate for the seven cities that make up the first column of Table 3.7. The mean rate of forcible rape would be as follows:

$$\begin{aligned} \bar{X} &= \frac{\sum_{i=1}^n X_i}{n} \\ \bar{X} &= \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7}{7} \\ \bar{X} &= \frac{22.2 + 23.5 + 28 + 28.1 + 28.4 + 33.8 + 38.4}{7} \\ \bar{X} &= \frac{202.4}{7} \\ \bar{X} &= 28.91 \text{ rapes per } 100,000 \end{aligned}$$

As practice, let's also calculate the mean rape rate per 100,000 for the second column of cities in Table 3.7:

$$\begin{aligned} \bar{X} &= \frac{22.2 + 23.5 + 28 + 28.1 + 28.4 + 33.8 + 38.4 + 133.2}{8} \\ \bar{X} &= \frac{335.6}{8} \\ \bar{X} &= 41.95 \text{ rapes per } 100,000 \end{aligned}$$

Like the median, the mean also is a sort of balancing score. The mean exactly balances the distance of each score from the mean. If we were to subtract the mean of a sample from each score in the sample, the negative differences from the mean would exactly equal the positive differences. Let's take a simple example. We have a set of five scores (2, 4, 6, 8, and 10). We calculate the mean and find the following:

$$\bar{X} = \frac{(2 + 4 + 6 + 8 + 10)}{5} = \frac{30}{5} = 6$$

We then subtract this mean from each score:

$$2 - 6 = -4$$

$$4 - 6 = -2$$

$$6 - 6 = 0$$

$$8 - 6 = 2$$

$$10 - 6 = 4$$

Subtracting the mean from each score yields what is called the *mean deviation*. Note that the sum of the negative differences is  $-6$  and the sum of the positive differences is  $+6$ , so the sum of all differences from the mean is equal to zero. This will always be true. It is in this sense that the mean is a balancing score of the differences. The mean is the only measure of central tendency that has this characteristic. The sum of the differences of each score from the mean, then, will always be zero. In mathematical terms, this means that  $\Sigma(x_i - \bar{X}) = 0$ .

Formula 3-3 for calculating the sample mean is very simple and easy to use when there are only a few scores. When there are a large number of scores, however, this formula is a bit cumbersome. To calculate the sample mean when there are many scores in a frequency distribution, the following formula is easier to use:

$$\bar{X} = \frac{\sum X_i f_i}{n} \quad (3-4)$$

where

$\bar{X}$  = the sample mean

$x_i$  = the  $i$ th score

$f_i$  = the frequency for the  $i$ th score

$X_i f_i$  = the  $x$ th score multiplied by its frequency

$n$  = the total number of scores

Formula 3-4 may seem a bit complicated, so let's illustrate its use step by step with an example.

## Case Study

### Calculating the Mean Police Response Time

Table 3.8 shows an ungrouped frequency distribution of response times to 911 calls to the police for assistance. Each response time was rounded to the nearest minute. Just so there is no confusion here, note that there were five occasions when the police responded to a 911 call within 1 minute, six times when they responded within 2 minutes, nine times when they responded within 3 minutes, and so on, concluding with one time when they

responded to a call for assistance within 11 minutes. You may recall that we used these data in the last chapter. The first step in calculating the mean is to create a new column of scores where each entry in the column is the product of each.

The  $x$  score is multiplied by its frequency  $f$  (this column is labeled  $x_i f_i$ ). For example, the first entry in the  $x_i f_i$  column is 5, which represents the fact that the police responded to a 911 call within 1 minute five times. Normally, to calculate the mean, we would add these five scores of 1 by doing  $1 + 1 + 1 + 1 + 1 = 5$ . By taking the product of the score and its frequency ( $x_i f_i$ ) instead, we are simply taking advantage of the fact that  $1 + 1 + 1 + 1 + 1 = 1 \times 5 = 5$ . For the second entry of the third column we are taking advantage of the fact that  $2 + 2 + 2 + 2 + 2 + 2 = 2 \times 6 = 12$ . We take each  $x_i$  score and multiply it by its frequency to form the column of  $x_i f_i$ . The second step in calculating the mean is to sum all of these products. The sum of the column of  $x_i f_i$  in Table 3.8 is 224. This is what we would have obtained had we taken the first approach and summed all  $x_i$  scores ( $1 + 1 + 1 + 1 + 1 + \dots + 7 + 7 + 7 + \dots + 10 + 11 = 224$ ). The third step in calculating the mean is to divide the sum of the product  $x_i f_i$  by the total number of sample scores. In this case, since there were 50 911 calls, we can calculate the mean or average response time to a 911 call as  $\bar{X} = 224 / 50 = 4.48$  minutes. Since .48 minute is equal to 28.8 seconds ( $.48 \times 60$  seconds), the average response time was 4 minutes and 28.8 seconds, or about 4.5 minutes.

Remember that the total number of sample scores is  $n$  and is the sum of the number of frequencies. Very often, students will use the number of different scores in the frequency distribution, rather than the total number of scores, as the denominator for the mean. For example, rather than using 50 as the denominator in the earlier problem since there were 50 response times recorded, many students are tempted to use 11 because there are 11 different values. There may be only 11 values for the variable “police response time,” but there were a total of 50 calls for police services, and this is the total number of observations.

Table 3.8

## Response Times to 911 Calls for Police Assistance

Minutes	$f_i$	$x_i f_i$
1	5	5
2	6	12
3	9	27
4	8	32
5	6	30
6	7	42
7	3	21
8	2	16
9	2	18
10	1	10
11	1	11
	$n = 50$	$\Sigma = 224$

### Steps to Calculate the Mean From an Ungrouped Frequency Distribution

**Step 1:** Multiply each  $x_i$  score by its frequency ( $f_i$ ). This will give you a column of products ( $x_i f_i$ ).

**Step 2:** Sum the obtained products from step 1:

$$\Sigma(x_i f_i)$$

**Step 3:** Divide this by the total number of scores ( $n$ ):

$$\bar{X} = \frac{\Sigma(x_i f_i)}{n}$$

## The Mean for Grouped Data

The procedures for calculating the mean when the data are in a grouped frequency distribution are very similar to those used when the data are in an ungrouped frequency distribution. The first thing you have to determine is that the underlying measurement of the data is continuous even though the data are grouped. If you are satisfied that the data are continuous and have been put into a grouped frequency distribution for convenience and clarification, then you may proceed. Keep in mind that since the data are in the form of a grouped frequency distribution, there are no individual  $x$  scores. Rather, the data are now in the form of class intervals, and although we know which class interval a score falls into, we do not know the exact score. To calculate a mean with grouped data, then, we are going to have to make a simplifying assumption. We must make the assumption that *each score within a class interval is located*



*exactly at its midpoint.* Once we make this assumption, we do not exactly have a distribution of  $x_i$  scores, but we have a distribution of  $m_i$  scores, where the  $m_i$  refers to the midpoint of the  $i$ th class interval.

Earlier in this chapter, Table 3.3 provided the grouped frequency distribution data for the time until arrest for a sample of 120 offenders released into the community. Recall that these data are a count of the number of days a released offender was in the community until he was rearrested. These are grouped data that were originally continuous, so we can legitimately calculate a mean. Our simplifying assumption is that each score within a class interval lies at its midpoint. So, for purposes of calculating the mean, we are going to assume that all four cases in the first class interval are at the midpoint of 18 days, the one score in the second class interval is at the midpoint of 21, the 12 scores in the third class interval are at the midpoint of 24, and so on. Recall that we need to make this assumption because to calculate a mean, we need to have a specific score (e.g., 18 days) rather than an interval within which a score lies (17–19 days). Since we are making this assumption, we are getting only an estimate of the mean for these data. This estimate is probably not going to be exactly what the value of the mean would be if we calculated it from the original continuous data. In a moment, we will check and see how accurate we are in making this assumption.

Once we have made this assumption, we are ready to calculate our mean. Recall that when we have data in the form of a frequency distribution, we can use formula 3-4 to calculate the mean by taking the product of each  $x$  score and its frequency ( $x_i f_i$ ), summing these products over all  $x_i$  scores, and then dividing by the total number of scores. We are going to modify this formula only slightly and use it to calculate a mean from grouped data. With grouped data, we do not have an individual  $x_i$  score, but we do have  $m_i$  scores since we are assuming that each score within its class interval lies at its midpoint ( $m_i$ ). To calculate the mean from grouped data, then, we just substitute  $m_i$  for  $x_i$  in formula 3-4 and take the product of each midpoint and the number of scores that are assumed to lie at that midpoint:

$$\bar{X} = \frac{\sum m_i f_i}{n} \quad (3-5)$$

where

$\bar{X}$  = the mean

$m_i$  = the midpoint for the  $i$ th class interval

$f_i$  = the frequency for the  $i$ th class interval

$m_i f_i$  = the  $m_i$  midpoint multiplied by its frequency

$n$  = the total number of scores

In other words, to calculate the mean, we multiply the midpoint of each class interval by the frequency of that class interval. Once we have done this for each class interval, we sum these products over all intervals and then divide by the total number of scores. We will illustrate the use of the mean formula for grouped data with the time-until-rearrest data. Table 3.9 provides the information we need.

The sum of each midpoint multiplied by its frequency is 3,723. Now, to calculate the mean, all we have to do is divide this sum by the total number of scores or observations. The mean number of days free until rearrest, therefore, is

$$\bar{X} = \frac{3,723}{120} = 31.02 \text{ days}$$

On average, then, these offenders were free for 31.02 days before being rearrested and returned to prison. When you calculate the mean from grouped data using this method, make sure that you use the correct sample size for the denominator. The  $n$  in the formula is the total number of sample observations or scores you have. In this example, our data consist of 120 observations.

Recall that these time-until-rearrest data were originally measured at the interval/ratio level from which we created class intervals. In the previous example, we estimated the mean number of days an offender was free in the community

based on the class interval scores. The question to answer now is whether we were accurate in our estimation of the mean using this grouped data. To determine our precision, let's calculate the mean number of days until rearrest from the original variable measured at the interval/ratio level and compare it with our estimate with formula 3-5. Table 3.10 gives the frequency distribution for the time-until-rearrest data in their original form, and we provide the necessary  $(f_i x_i)$  column. Using the ungrouped data, then, we can calculate the mean as  $3,729 / 120 = 31.075$  days. Our estimate of the mean with the grouped data was 31.02 days, so we were pretty close to the value of the mean had the data remained in its original interval/ratio form. In general, you will find that you will not lose much accuracy in estimating the mean when you use grouped data rather than ungrouped data if the grouping was carefully done.

### Advantages and Disadvantages of the Mean

The mean has a number of advantages as a measure of central tendency. First, it is intuitively appealing. Everyone is familiar with an average. The mean also uses all of the information in a data set, and this is an advantage as long as there are no outliers in the data. The mean is also an efficient measure of central tendency. In other words, if we had a population of scores (with a mean and a median), and from this population we took many samples and calculated both the mean and the median for each sample, the medians of these samples would differ more from each other and the population median than the means would differ from each other and the population mean.

Because we usually draw only one sample from a population, we want to have the measure of central tendency that is the most precise. This is the mean. The one disadvantage of the mean is a by-product of one of its strengths; because it takes every score into account, the mean may be distorted by high or low outliers. When we sum every score to calculate the mean, we may at times be adding uncharacteristically high or uncharacteristically low scores. When this happens, the value of the mean will give us a distorted sense of the central tendency of the data. To illustrate this point, let's return to Table 3.7, which provided three columns of rape rates per 100,000 people for selected U.S. cities. The first column consists of seven cities. Calculate the mean rape rate for these cities. You should have obtained a value of  $\bar{X} = 202.4 / 7 = 28.91$  rapes per 100,000. What happens to the mean when we include the high outlier of Anchorage from the second column? The mean now is  $\bar{X} = 335.6 / 8 = 41.95$  rapes per 100,000. Note what happened to the value of the mean when we included this high outlier. The magnitude of the mean increased dramatically, and it is now more than 64 rapes per 100,000. The inclusion of a high outlier, then, had the effect of inflating our mean.

Now look at the third column of cities in Table 3.7. What happens to the mean when we drop the high outlier of Anchorage and add the low outlier of Goldsboro, with a rape rate of 4.0 per 100,000? The value of the mean is now  $\bar{X} = 202.4 / 8 = 25.3$  rapes per 100,000. The mean has now declined slightly, although it does not distort the central tendency as much as our high outlier did. As you can see, however, the effect of a low outlier is to lessen the magnitude of the mean.

**Table 3.9** Calculating a Mean Using Grouped Data: Time Until Rearrest for 120 Inmates

Stated Limits (Days)	F	Midpoint	$m f_i$
17–19	4	18	72
20–22	1	21	21
23–25	12	24	288
26–28	16	27	432
29–31	28	30	840
32–34	28	33	924
35–37	21	36	756
38–40	10	39	390
	$n = 120$		$\Sigma = 3,723$

#### Steps to Calculate the Mean From an Ungrouped Frequency Distribution

**Step 1:** Multiply each  $x_i$  score by its frequency ( $f_i$ ). This will give you a column of products ( $x_i f_i$ ).

**Step 2:** Sum the obtained products from step 1:

$$\Sigma(x_i f_i)$$

**Step 3:** Divide this by the total number of scores ( $n$ ):

$$\bar{X} = \frac{\Sigma m_i f_i}{n}$$

**Table 3.10**  
**Calculating a Mean Using**  
**Ungrouped Data: Time Until**  
**Rearrest for 120 Inmates**

$x_i$	$f_i$	$x_i f_i$
17	1	17
18	1	18
19	2	38
20	1	20
21	0	0
22	0	0
23	3	69
24	4	96
25	5	125
26	3	78
27	7	189
28	6	168
29	11	319
30	7	210
31	10	310
32	7	224
33	12	396
34	8	272
35	8	280
36	8	288
37	6	222
38	4	152
39	2	78
40	4	160
	$n = 120$	$\Sigma = 3,729$

The purpose of this exercise is to show that sometimes the mean can provide a distorted sense of the central tendency in our data. Since the mean uses every score in our distribution, high outliers can inflate the mean, and low outliers can deflate the mean, relative to what the value of the mean would be without the outliers. For this reason, it is generally a good idea to report *both* the mean and the median when you are discussing the central tendency in your data. With respect to Table 3.7, we saw how the mean increases or decreases with the inclusion of outliers in the data.

Reporting both the mean and the median can also tell us something important about the shape of our data. In Chapter 2, we illustrated the difference between symmetrical and skewed distributions. In Chapter 5, we will be discussing the normal or “bell-shaped” distribution, which is a very important theoretical probability distribution in statistics. A normal distribution has one mode (it has a single peak), and it is symmetrical. If a line were drawn down the center of the distribution, the left half would be a mirror image of the right half. In a symmetrical or normal distribution, the mean, median, and mode are all the same, located right at the center of the distribution. If a distribution is not normal, recall that it is said to be a skewed distribution. In a negatively skewed distribution, the mean is less than the median. This is because there are low outlying scores on the left of the distribution pulling the value of the mean down. Stated differently, in a negatively skewed distribution, the mean is lower in magnitude than the median because low outliers are deflating the mean. This is what we saw in the third column of Table 3.7. Thus, knowing that in a distribution of scores the mean is much lower than the median, we might suspect that the distribution has a negative skew. The greater the difference there is between the mean and the median, the greater the negative skew. Conversely, in a positively skewed distribution, the mean is greater than the median because high outliers are inflating the magnitude of the mean relative to the median, as we saw in the second column of Table 3.7.

## 2 Summary

In this chapter, we focused on measures of central tendency. These measures of central tendency are used as summary indicators of the typical, usual, most frequent, or average score in a distribution of scores. There are three measures of central tendency: the mode, the median, and the mean.

The mode is the score or value with the highest frequency. Therefore, it is the score or value that has the greatest probability or likelihood of occurring. There may be more than one mode in a given distribution of scores. As a measure of central tendency, the mode is probably the easiest to obtain since it requires no real calculations and is an appropriate measure of central tendency for nominal, ordinal, or interval/ratio-level data.

The median is the score at the 50th percentile. Thus, it is the score or value that divides a rank-ordered distribution of scores into two equal halves. A characteristic of the median, then, is that one half of the scores will be greater than the median and one half will be less than it. The median requires continuous-level data (interval/ratio) or continuous-level data that have been made ordinal through the creation of a grouped frequency distribution. Since the median locates the score at the 50th percentile, it is not affected by outlying scores in a distribution. For this reason, it is a very good measure of central tendency when the data are skewed.

The mean is the arithmetic average of all scores. It is calculated by summing all scores and dividing by the total number of scores. Calculation of the mean requires the same level of measurement as does the median. Because the

mean uses all of the scores, it can be substantially affected by the presence of outliers in the data. In a normal distribution, the mode, median, and mean are the same. In a negatively skewed distribution, the mean is less than the median, and in a positively skewed distribution, the mean is greater than the median. Because the presence of outliers may distort the mean as a measure of central tendency, it is generally a good policy to report both the median and the mean.

## Key Terms

▶ Review key terms with eFlashcards. 

bimodal distribution 63  
mean 70

measures of central tendency 61  
median 65  
mode 61

## Key Formulas

Sample median for grouped data (equation 3-1):

$$X_{\text{median}} = L + \left( \frac{\left( \frac{n+1}{2} \right) - cf}{f} \right) w_i$$

where

- $X_{\text{median}}$  = the value of the median
- $L$  = the lower real limit of the class interval that contains the median
- $cf$  = the cumulative frequency of the class interval just before the class interval that contains the median
- $f$  = the frequency of the interval that contains the median
- $w_i$  = the width of the class interval
- $n$  = the total number of observations in the sample

Sample mean of a population (equation 3-2):

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

where

- $X_i$  = each  $X$  score in the population
- $N$  = the total number of observations in the population

Sample mean (equation 3-3):

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

where

- $\bar{X}$  = the symbol used for the sample mean (pronounced “x bar”)
- $x_i$  = the  $i$ th raw score in a distribution of scores
- $\sum_{i=1}^n x_i$  = the instruction to sum all  $x_i$  scores, starting with the first score ( $i = 1$ ) and continuing until the last score ( $i = n$ )
- $n$  = the total number of scores

Sample mean for data in a frequency distribution (equation 3-4):

$$\bar{X} = \frac{\sum m_i f_i}{n}$$

where

- $\bar{X}$  = the mean
- $x_i$  = the  $i$ th score
- $f_i$  = the frequency for the  $i$ th score
- $x_i f_i$  = the  $x$ th score multiplied by its frequency
- $n$  = the total number of scores


Sample mean for grouped data (equation 3-5):

$$\bar{X} = \frac{\sum m_i f_i}{n}$$

where

- $\bar{X}$  = the mean
- $m_i$  = the midpoint for the  $i$ th class interval
- $f$  = the frequency for the  $i$ th class interval
- $m_i f_i$  = the  $m_i$  midpoint multiplied by its frequency
- $n$  = the total number of scores

## Practice Problems

► Test your understanding of chapter content. Take the practice quiz. 

- As a measure of central tendency, the mode is the most common score. Consider the following information on a variable called “the number of delinquent friends that someone has.” What is the mode for these data, and what does it tell you? Why can’t you calculate the “mean number of delinquent friends”?

### Number of Delinquent Friends

<i>X</i>	<i>f</i>
None	20
Some	85
Most	30
All	10

- Say you asked a random sample of seven correctional officers what their annual salary was, and their responses were as follows:

\$25,900

\$32,100

\$28,400

\$31,000

\$29,500

\$27,800

\$26,100

What is the median salary, and what is the mean salary, for this sample?

- The following data show the homicide rate per 100,000 people for 10 American cities. Given these data, which measure of central tendency would you use and why?

<i>City</i>	<i>Homicide Rate</i>
Boston, MA	6.8
Cincinnati, OH	4.5
Denver, CO	6.0
Las Vegas, NV	8.8
New Orleans, LA	43.3
New York, NY	8.7
Pittsburgh, PA	10.5
Salt Lake City, UT	5.6
San Diego, CA	4.3
San Francisco, CA	7.7

- Rachel Sutherland and her colleagues (2015) have investigated the relationship between injection drug use and criminal activity. The hypothetical data that follow represent the number of crimes committed during a 2-year period by 20 heroin addicts. Using ungrouped data, calculate the mean and the median for these 20 persons. Which measure of central tendency do you think best summarizes the central tendency of these data and why?

<i>Person Number</i>	<i>Number of Crimes Committed</i>	<i>Person Number</i>	<i>Number of Crimes Committed</i>
1	4	11	4
2	16	12	11
3	10	13	10
4	7	14	88
5	3	15	9
6	112	16	12
7	5	17	8
8	10	18	5
9	6	19	7
10	2	20	10

- In a study of police interventions and mental illness in a large Canadian city, Yannick Charette, Anne Crocker, and Isabelle Billette (2014, p. 513) reported the following distribution of the reasons for police intervention when the subject was without mental illness:

<i>Request</i>	<i>Frequency</i>
Offense against person	213
Offense against property	496
Other criminal offense	238
Potential offense	3,784
Individual in distress	139
Noncriminal incident	986

What is the measure of central tendency most appropriate for these data? Why? What does this measure of central tendency tell you about the “most typical” reason for a police intervention when the subject was without mental illness?

- The following hypothetical data show the distribution of the percentage of total police officers who do narcotics investigation in 100 American cities. Determine the mode, median, and mean.

**Percentage of Force Doing Investigation**

<i>Narcotics Investigation (%)</i>	<i>Frequency</i>
0–9	5
10–19	13
20–29	26
30–39	38
40–49	14
50–59	2
60–69	2

7. The following data represent the number of persons executed in the United States from 2007 to 2014.

<i>Year</i>	<i># of Executions</i>
2007	42
2008	37
2009	52
2010	46
2011	43
2012	43
2013	39
2014	35

What were the mean number and median number of executions over this time period? What happens to the median and mean when we add the year 2006, in which there were 53 executions? Which measure of central tendency would you use to describe the 2007–2014 distribution?

8. One seemingly inconsistent finding in criminological research is that women have a greater subjective fear of crime than men even though their objective risk of being the victim of a crime is lower. In one study, Jodi Lane and Kathleen Fox (2013) tried to explain this fact in part through the shadow of sexual assault thesis by suggesting that women are more afraid of crime because of their fear of sexual assault and the intense physical and emotional consequences they would face if raped. They suggest that

women transfer this fear of sexual assault to a fear of crime generally. The hypothetical data that follow represent the responses of a sample of 200 women who were asked to report to an interviewer the number of times that they had been sexually assaulted during the previous 5 years. Using these data, calculate the mean, median, and mode.

<i>Number of Times Assaulted</i>	<i>Frequency</i>
0–1	85
2–3	70
4–5	30
6–7	15

9. Research reported by Adrian Raine, Annis Lai Chu Fung, Jill Portnoy, Olivia Choy, and Victoria Spring (2014) suggests that there is a link between low resting heart rates and aggression and psychopathic traits. They define those with resting heart rates below 67 beats per minute as having low resting heart rates. In a random sample of 20 violent offenders currently incarcerated in a state penitentiary, the prison doctor finds the following resting heart rates. Calculate the mean and median for these data. Are the mean and median the same or different? Why do you think this is so?

<i>Person</i>	<i>Resting Heart Rate</i>	<i>Person</i>	<i>Resting Heart Rate</i>
1	59	11	60
2	62	12	55
3	69	13	52
4	62	14	70
5	64	15	52
6	70	16	57
7	54	17	53
8	66	18	61
9	51	19	64
10	56	20	63

**STUDENT STUDY SITE****WANT A BETTER GRADE?**

Get the tools you need to sharpen your study skills. Access practice quizzes, eFlashcards, data sets, and exercises at [edge.sagepub.com/bachmansccj4e](http://edge.sagepub.com/bachmansccj4e).