4

COMPUTING AND UNDERSTANDING AVERAGES

Means to an End

Difficulty Scale ☺ ☺ ☺ ☺
(moderately easy)

WHAT YOU WILL LEARN IN THIS CHAPTER

- Understanding measures of central tendency
- Computing the mean for a set of scores
- Computing the median for a set of scores
- Computing the mode for a set of scores
- Understanding and applying scales or levels of measurement
- Selecting a measure of central tendency

You've been very patient, and now it's finally time to get started working with some real, live data. That's exactly what you'll do in this chapter. Once data are collected, a usual first step is to organize the information using simple indexes to describe the data. The easiest way to do this is through computing an average, of which there are several different types.

An **average** is the one value that best represents an entire group of scores. It doesn't matter whether the group of scores is the number correct on a spelling test for 30 fifth graders or the batting percentage of each of the New York Yankees (which, by the way, was not very good during the 2015 season) or the number of people who registered as Democrats or Republicans for the upcoming elections. In all of these examples, groups of data can be summarized using an average. You can also think of an average as the “middle” space or as a fulcrum on a seesaw. It’s the point where all the values in a set of values are balanced.

Averages, also called **measures of central tendency**, come in three flavors: the mean, the median, and the mode. Each provides you with a different type of information about a distribution of scores and is simple to compute and interpret.

COMPUTING THE MEAN

The **mean** is the most common type of average that is computed. It is simply the sum of all the values in a group, divided by the number of values in that group. So, if you had the spelling scores for 30 fifth graders, you would simply add up all the scores to get a total and then divide by the number of students, which is 30.

The formula for computing the mean is shown in Formula 4.1.

$$\bar{X} = \frac{\sum X}{n}, \tag{4.1}$$

where

- the letter *X* with a line above it (also sometimes called “*X* bar”) is the mean value of the group of scores or the mean;
- the $\sum$, or the Greek letter sigma, is the summation sign, which tells you to add together whatever follows it to obtain a total or sum;
- the *X* is each individual score in the group of scores; and
- the *n* is the size of the sample from which you are computing the mean.

To compute the mean, follow these steps:

1. List the entire set of values in one or more columns. These are all the *X*s.
2. Add all of the values together to obtain the total.
3. Divide the total by the number of values.

For example, if you needed to compute the average number of shoppers at three different locations, you would compute a mean for that value.

Location	Number of Annual Customers
Lanham Park Store	2,150
Williamsburg Store	1,534
Downtown Store	3,564

The mean or average number of shoppers in each store is 2,416. Formula 4.2 shows how this average was computed using the formula you saw in Formula 4.1:

$$\bar{X} = \frac{\sum X}{n} = \frac{2{,}150 + 1{,}534 + 3{,}564}{3} = \frac{7{,}248}{3} = 2{,}416. \qquad (4.2)$$

Or, if you needed to compute the average number of students in kindergarten through Grade 6, you would follow the same procedure.

Grade	Number of Students
Kindergarten	18
1	21
2	24
3	23
4	22
5	24
6	25

The mean or average number of students in each class is 22.43. Formula 4.3 shows how this average was computed using the formula you saw in Formula 4.1:

$$\bar{X} = \frac{\sum X}{n} = \frac{18 + 21 + 24 + 23 + 22 + 24 + 25}{7} = \frac{157}{7} = 22.43. \qquad (4.3)$$

See, we told you it was easy. No big deal. Let's repeat our two computations in R starting with the average number of customers.

```
> (2150 + 1534 + 3564)/3
[1] 2416
>
```

Let's break this down.

- >—the first character in a new line in the R Console.
- `(2150 + 1534 + 3564)/3`—we added the customer counts together and divided the total by the number of locations, 3.
- `[1] 2416`—the mean number of customers, also called the average.

Repeating this process to compute the mean of students across classrooms, the syntax in R is

```
> (18 + 21 + 24 + 23 + 22 + 24 + 25)/7
[1] 22.42857
>
```

Note that the average number of customers was a whole number, 2,416, but the average number of students did not divide evenly, so the answer was 22.42857. When we reported the average number of students, calculated with a calculator, we rounded to the second decimal place, giving us an average of 22.43 students per classroom. Like using a calculator, R will report numbers past two decimal places, so you will round to the number appropriate for your field. Psychologists follow reporting guidelines for the American Psychological Association (APA) in the APA style guide and round to two decimal places. Ask your instructor what is common practice in her or his field.

- The mean is sometimes represented by the letter *M* and is also called the typical or average score. If you are reading another statistics book or a research report and you see something like $M = 45.87$, it probably means that the mean is equal to 45.87.
- In the formula, a small *n* represents the sample size for which the mean is being computed. A large *N* (← like this) would represent the population size. In some books and in some journal articles, no distinction is made between the two.
- In the classroom example, we had a sample of elementary school classrooms. If we calculate a mean for all elementary school classrooms across the country, then we are computing the mean for the population of elementary classrooms in the country.
- The sample mean is the measure of central tendency that most accurately reflects the population mean.
- The mean is like the fulcrum on a seesaw. It's the centermost point where all the values on one side of the mean are equal in weight to all the values on the other side of the mean.
- Finally, for better or worse, the mean is very sensitive to extreme scores. An extreme score can pull the mean in one or the other direction and make it less representative of the set of scores and less useful as a measure of central tendency. This, of course, all depends on the values for which the mean is being computed. And, if you have extreme scores and the mean won't work as well as you want, we have a solution! More about that later.

The mean is also referred to as the **arithmetic mean**, and there are other types of means that you may read about, such as the harmonic mean. Those are used in special circumstances and need not concern you here. And if you want to be technical about it, the arithmetic mean (which is the one that we have discussed up to now) is also defined as the point about which the sum of the deviations is equal to zero (whew!). So, if you have scores like 3, 4, and 5 (whose mean is 4), the sum of the deviations about the mean (–1, 0, and +1) is 0.

Remember that the word *average* means only the one measure that best represents a set of scores and that there are many different types of averages. Which type of average you use depends on the question that you are asking and the type of data that you are trying to summarize. This is a levels-of-measurement issue, which we will cover later in this chapter when we talk about when to use which measure.

Computing a Weighted Mean

You've just seen an example of how to compute a simple mean. But there may be situations when you have the occurrence of more than one value and you want to compute a weighted mean. A weighted mean can be easily computed by multiplying the value by the frequency of its occurrence, adding the total of all the products, and then dividing by the total number of occurrences. It beats adding up every individual data point.

To compute a weighted mean, follow these steps:

1. List all the values in the sample for which the mean is being computed, such as those shown in the column labeled "Value" (the value of *X*) in the following table.
2. List the frequency with which each value occurs.
3. Multiply the value by the frequency, as shown in the third column.
4. Sum all the values in the Value × Frequency column.
5. Divide by the total frequency.

For example, here's a table that organizes the values and frequencies in a flying proficiency test for 100 airline pilots.

Value	Frequency	Value × Frequency
97	4	388
94	11	1,034
92	12	1,104

(Continued)

(Continued)

Value	Frequency	Value × Frequency
91	21	1,911
90	30	2,700
89	12	1,068
78	9	702
60 (Don't fly with this guy.)	1	60
Total	100	8,967

The weighted mean is 8,967/100, or 89.67. Computing the mean this way is much easier than entering 100 different scores into your calculator or computer program.

In basic statistics, an important distinction is made between those values associated with samples (a part of a population) and those associated with populations. To do this, statisticians use the following conventions. For a sample statistic (such as the mean of a sample), Roman letters are used. For a population parameter (such as the mean of a population), Greek letters are used. For example, the mean for the spelling score for a sample of 100 fifth graders is represented as $\bar{X}_5$, whereas the mean for the spelling score for the entire population of fifth graders is represented, using the Greek letter mu, as μ_5.

COMPUTING THE MEDIAN

The median is also a measure of central tendency but of a very different kind. The **median** is defined as the midpoint in a set of scores. It's the point at which one half, or 50%, of the scores fall above and one half, or 50%, fall below. It's got some special qualities that we will talk about later in this section, but for now, let's concentrate on how it is computed. There's no standard formula for computing the median.

To compute the median, follow these steps:

1. List the values in order, from either highest to lowest or lowest to highest.
2. Find the middle-most score. That's the median.

For example, here are the incomes from five different households:

$135,456

$25,500

$32,456

$54,365

$37,668

Here is the list ordered from highest to lowest:

$135,456

$54,365

$37,668

$32,456

$25,500

There are five values. The middle-most value is $37,668, and that's the median.

Now, what if the number of values is even? Let's add a value ($34,500) to the list so there are six income levels. Here they are sorted with the largest value first:

$135,456

$54,365

$37,668

$34,500

$32,456

$25,500

When there are an even number of values, the median is simply the mean of the two middle values. In this case, the middle two cases are $34,500 and $37,668. The mean of those two values is $36,084. That's the median for that set of six values.

What if the two middle-most values are the same, such as in the following set of data?

$45,678

$25,567

$25,567

$13,234

Then the median is same as both of those middle-most values. In this example, it's $25,567.

If we had a series of values that was the number of days spent in rehabilitation for a sports-related injury for seven different patients, the numbers might look like this:

43

34

32

12

51

6

27

As we did before, we can order the values (51, 43, 34, 32, 27, 12, 6) and then select the middle value as the median, which in this case is 32. So, the median number of days spent in rehab is 32.

Placing these numbers in order is straightforward when you have fewer than 10 numbers. Once you start dealing with data sets with more observations, like some used later in this chapter and the book, R can easily reorder your data. Let R sort your numbers from smallest to largest like this:

```
> sort(c(43, 34, 32, 12, 51, 6, 27))
[1] 6 12 27 32 34 43 51
>
```

The `sort()` function expects a vector of values, so we used the `c()` function to put all of our numbers together in a vector. What did we get back? A reordered vector of seven numbers, in which we can easily see the median is 32 because it is the middle value.

If you know about medians, you should also know about **percentile points**. Percentile points are used to define the percentage of cases equal to and below a certain point in a distribution or set of scores. For example, if a score is "at the 75th percentile," it means that the score is at or above 75% of the other scores in the distribution. The median is also known as the 50th percentile, because it's the point below which 50% of the cases in the distribution fall. Other percentiles are useful as well, such as the 25th percentile, often called Q_1, and the 75th percentile, referred to as Q_3. So what's Q_2? The median, of course.

Here comes the answer to the question you've probably had in the back of your mind since we started talking about the median. Why use the median instead of the mean?

For one very good reason. The median is unaffected by extreme scores, whereas the mean is affected by extreme values.

When you have a set of scores in which one or more scores are extreme, the median better represents the centermost value of that set of scores than any other measure of central tendency. Yes, even better than the mean.

What do we mean by *extreme?* It's probably easiest to think of an extreme score as one that is very different from the group to which it belongs. For example, consider the list of five incomes that we worked with earlier (shown again here):

$135,456

$54,365

$37,668

$32,456

$25,500

The value $135,456 is more different from the other four than is any other value in the set. We would consider that an extreme score.

The best way to illustrate how useful the median is as a measure of central tendency is to compute both the mean and the median for a set of data that contains one or more extreme scores and then compare them to see which one best represents the group. Here goes.

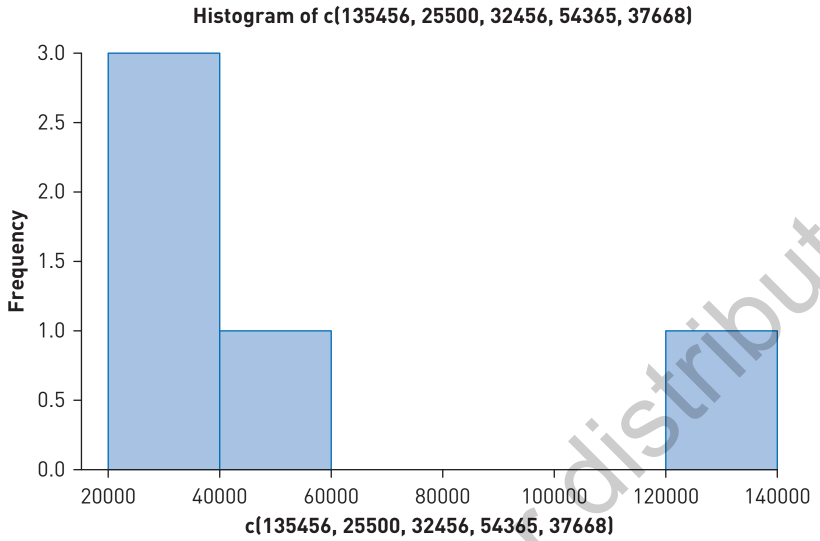
The mean of the set of five scores you see above is the sum of the set of five divided by 5, which turns out to be $57,089. On the other hand, the median for this set of five scores is $37,668. Which is more representative of the group? The value $37,668, because it clearly lies more in the middle of the group, and we like to think about "the average" (in this case, we are using the median as a measure of central tendency) as being representative or assuming a central position. In fact, the mean value of $57,089 falls above the fourth highest value ($54,365) and is not very central or representative of the distribution.

It's for this reason that certain social and economic indicators (often involving income) are reported using a median as a measure of central tendency—"The median income of the average American family is . . ."—rather than using the mean to summarize the values. There are just too many extreme scores that would **skew**, or significantly distort, what is actually a central point in the set or distribution of scores.

Another way to identify when the median might be more useful than the mean is to create a picture. We will spend more time in Chapter 6 talking about visualizing data and how to personalize those graphics, but let's look at one option now. The R function we are going to use is `hist()`. Let's input the numbers from the first income example to create a histogram like Figure 4.1.

```
> hist(c(135456, 25500, 32456, 54365, 37668))
>
```

FIGURE 4.1 ● Histogram of the vector of income numbers.



R created a title displayed across the top of the figure, "`Histogram of c(135456, 25500, 32456, 54365, 37668).`" It is not necessarily the best title in the world, but R lets us know what we are looking at. The *x*-axis ranges from 20,000 to 140,000 in 20,000 increments. R also described the vector of numbers used as input for the function. The *y*-axis shows the frequency (aka count) of observations in each block. What is most interesting about this histogram? The gap. Nothing is happening between 60,000 and 120,000, and that tells you the value on the far right is very different from the other items. Therefore, the median is probably a better measure of central tendency to use with these data.

With only six numbers, we can easily tell that 135,456 is extreme without the histogram. When someone gives you some data with 50, 100, or more values, the histogram can be quite helpful!

You learned earlier that sometimes the mean is represented by the capital letter *M* instead of $\bar{X}$. Well, other symbols are used for the median as well. We like the letter *M*, but some people confuse it with the mean, so they use *Med* or *Mdn* for median. Don't let that throw you—just remember what the median is and what it represents, and you'll have no trouble adapting to different symbols.

Here are some interesting and important things to remember about the median.

- Although the mean is the middle point of a set of values, the median is the middle point of a set of cases.
- Because the median deals with how many cases there are and not the values of those cases, extreme scores (sometimes called outliers) don't overly influence the median.

COMPUTING THE MODE

The third and last measure of central tendency that we'll cover, the mode, is the most general and least precise measure of central tendency, but it plays a very important part in understanding the characteristics of a set of scores. The **mode** is the value that occurs most frequently. There is no formula for computing the mode.

To compute the mode, follow these steps:

1. List all the values in a distribution but list each value only once.
2. Tally the number of times that each value occurs.
3. The value that occurs most often is the mode.

For example, an examination of the political party affiliation of 300 people might result in the following distribution of scores.

Party Affiliation	Number or Frequency
Democrats	90
Republicans	70
Independents	140

The mode is the value that occurs most frequently, which in the preceding example is Independents. That's the mode for this distribution.

If we were looking at the modal response on a 100-item multiple-choice test, we might find that A was chosen more frequently than any other choice. The data might look like this.

Response Choices Selected	A	B	C	D
Number of times	57	20	12	11

On this 100-item multiple-choice test where each item has four choices (A, B, C, and D), A was the answer selected 57 times. It's the modal response.

Want to know the easiest and most commonly made mistake when computing the mode? It's selecting the number of *times* a category occurs, rather than the *label* of the category itself. Instead of the mode being Independents (in our first example above), it's easy for someone to mistakenly conclude the mode is 140. Why? Because he or she is looking at the number of times the value occurred, not the value that occurred most often! This is a simple error to make, so be on your toes when you are asked about these things.

Apple Pie à la Bimodal

If every value in a distribution contains the same number of occurrences, then there really isn't a single mode. But if more than one value appears with equal frequency, the distribution is multimodal. The set of scores can be bimodal (with two modes), as the following set of data using pie flavors illustrates.

Slices of Pie Flavors	Number or Frequency
Apple	28
Cherry	17
Pecan	45
Pumpkin	28

In the above example, the distribution is bimodal because the frequency of the values of apple and pumpkin occurs equally. You can even have a bimodal distribution when the modes are relatively close together but not exactly the same, such as 45 slices of apple pie and 44 slices of pumpkin pie. The question becomes, "How much does one class of occurrences stand apart from another?"

Can you have a trimodal distribution? Sure—where three values have the same frequency. It's unlikely, especially when you are dealing with a large set of **data points**, or observations, but certainly possible. The real answer to the above stand-apart question is that categories have to be mutually exclusive—in a restaurant, you cannot order a slice of pie and get both apple and cherry pie, unless the chef created a new pie flavor that combines the two. In that case, you have a new flavor category to count.

WHEN TO USE WHAT MEASURE OF CENTRAL TENDENCY (AND ALL YOU NEED TO KNOW ABOUT SCALES OF MEASUREMENT FOR NOW)

Which measure of central tendency you use depends on certain characteristics of the data you are working with—specifically the **scale of measurement** at which those data occur. And that scale or level dictates the specific measure of central tendency you will use.

But let's step back for just a minute and make sure that we have some vocabulary straight, beginning with the idea of what measurement is.

Measurement is the assignment of values to outcomes following a set of rules—simple. The results are the different scales we'll define in a moment, and an outcome is anything we are interested in measuring, such as hair color, gender, test score, or height.

These scales of measurement, or rules, are the particular levels at which outcomes are observed. Each level has a particular set of characteristics, and scales of measurement come in four flavors (there are four types): nominal, ordinal, interval, and ratio.

Let's move on to a brief discussion and examples of the four scales of measurement and then discuss how these levels of scales fit with the different measures of central tendency discussed earlier.

A Rose by Any Other Name: The Nominal Level of Measurement

The **nominal level of measurement** is defined by the characteristics of an outcome that fit into one and only one class or category. For example, sex can be a nominal variable (female, male, and other), as can ethnicity (Caucasian or African American), as can political affiliation (Republican, Democrat, or Independent). Nominal-level variables are "names" (*nominal* in Latin), and the nominal level can be the least precise level of measurement. Nominal levels of measurement have categories that are mutually exclusive; for example, political affiliation cannot be both Republican and Democrat. Another commonly used word to describe a nominal variable is categorical.

Any Order Is Fine With Me: The Ordinal Level of Measurement

The *ord* in **ordinal level of measurement** stands for *order,* and the characteristic of things being measured here is that they are ordered. The perfect example is a rank of candidates for a job. If we know that Amy is ranked #1, Joe is ranked #2, and Hannah is ranked #3, then this is an ordinal arrangement. We have no idea how much higher on this scale Amy is relative to Joe than Joe is relative to Hannah. We just know that it's "better" to be #1 than #2 or #3, but not by how much.

1 + 1 = 2: The Interval Level of Measurement

Now we're getting somewhere. When we talk about the **interval level of measurement**, a test or an assessment tool is based on some underlying continuum such that we can talk about how much more a higher performance is than a lesser one. For example, if you get 10 words correct on a vocabulary test, that is twice as many as getting 5 words correct. A distinguishing characteristic of interval-level scales is that the intervals or spaces or points along the scale are equal to one another. Ten words correct is 2 more than 8 correct, which is 3 more than 5 correct.

Can Anyone Have Nothing of Anything? The Ratio Level of Measurement

Well, here's a little conundrum for you. An assessment tool at the **ratio level of measurement** is characterized by the presence of an absolute zero on the scale. What that zero means is the absence of any of the trait that is being measured. The conundrum? Are there outcomes we measure where it is possible to have nothing of what is being measured? In some disciplines, that can be the case. For example, in the physical and biological sciences, you can have the absence of a characteristic, such as absolute zero (no molecular movement) or zero light. In the social and behavioral sciences, it's a bit harder. Even if you score zero on that spelling test or miss every item of an IQ test (in Russian), that does not mean that you have no spelling ability or no intelligence, right? So, the chance of 0 occurring in real life and in your data may be quite small, but we are talking about the theoretical possibility of a 0 or even values below 0.

In Sum . . .

These scales of measurement, or rules, represent particular levels at which outcomes are measured. And, in sum, we can say the following:

- Any outcome can be assigned to one of four scales of measurement.
- Scales of measurement have an order, from the least precise being nominal to the most precise being ratio.
- The "higher up" the scale of measurement, the more precise the data being collected, and the more detailed and informative the data are. It may be enough to know that some people are rich and some poor (and that's a nominal or categorical distinction), but it's much better to know exactly how much money they make (ratio). We can always make the "rich" versus "poor" distinction if we want to once we have all the information.
- Finally, the more precise scales contain all the qualities of the scales below them; for example, the interval scale includes the characteristics of the ordinal and nominal scales. If you know that the Cubs' batting average is .350, you know it is better than that of the Tigers (who hit .250) by 100 points, but you also know that the Cubs are better than the Tigers (but not by how much) and that the Cubs are different from the Tigers (but there's no direction to the difference).

Okay, we've defined levels of measurement, discussed three different measures of central tendency, and given you fairly clear examples of each. But the most important question remains unanswered: "When do you use which measure?"

In general, which measure of central tendency you use depends on the type of data that you are describing, which in turn means at what level of measurement the data occur. Unquestionably, a measure of central tendency for qualitative, categorical, or nominal data (such as racial group, eye color, income bracket, voting preference, and neighborhood location) can be described using only the mode.

For example, you can't be interested in the most central measure that describes which political affiliation is most predominant in a group and use the mean—what in the world could you conclude, that everyone is half Republican? Rather, saying that out of 300 people, almost half (140) are Independent seems to be the best way to describe the value of this variable. In general, the median and mean are best used with quantitative data, such as height, income level in dollars (not categories), age, test score, reaction time, and number of hours completed toward a degree.

It's also fair to say that the mean is a more precise measure than the median and that the median is a more precise measure than the mode. This means that all other things being equal, use the mean, and indeed, the mean is the most often used measure of central tendency. However, we do have occasions when the mean would not be appropriate as a measure of central tendency—for example, when we have categorical or nominal data, such as hospitalized versus nonhospitalized people. Then we use the mode.

So, here is a set of three guidelines that may be of some help. And remember, there can always be exceptions.

1. Use the mode when the data are categorical in nature and values can fit into only one class, such as hair color, political affiliation, neighborhood location, and religion. When this is the case, these categories are called mutually exclusive.
2. Use the median when you have extreme scores and you don't want to distort the average (computed as the mean), such as when the variable of interest is income expressed in dollars.
3. Finally, use the mean when you have data that do not include extreme scores and are not categorical, such as the numerical score on a test or the number of seconds it takes to swim 50 yards.

USING THE COMPUTER TO COMPUTE DESCRIPTIVE STATISTICS

Visit **edge.sagepub.com/salkindshaw** to watch an R tutorial video on this topic.

If you haven't already, now would be a good time to go back to Chapter 2 to install R and RStudio and Chapter 3 to start practicing with R. Then come back here.

Let's use R to compute some descriptive statistics. The data set we are using is Chapter 4 Data Set 1, named ch4ds1.csv. There is a method to our file names. Let's take a moment to decipher it.

- ch—shorthand for chapter
- 4—the number of the chapter that will use the data set
- ds—shorthand for data set
- 1—the first data set
- .csv—this file extension refers to comma separated values

R can import data from files that have spaces in the filename, but the process is simpler when the filename doesn't have spaces. So in the spirit of keeping things simple, we will use this file-naming convention throughout the book.

The file ch4ds1.csv is a set of 20 scores on a test of prejudice (also listed in Appendix C as Chapter 4 Data Set 1). All of the data sets are available in Appendix C and from the SAGE website **edge.sagepub.com/salkindshaw**. There is one variable in this data set:

Variable	Definition
Prejudice	The value on a test of prejudice as measured on a scale from 1 to 100

In R, we are going to look at one example of each measure of central tendency, using a combination of built-in functions to calculate the mean, and then look at the mean function itself. For median, we already looked at the sort function to put all items in a vector in order from smallest to largest. Now, we will look at the function that will do all of the work for us. Last, we will look at `mode()` and talk about why we get something unexpected. We will then look at another function that will simplify finding the mode.

But first, let's read in our data with the read.csv function.

```
> ch4ds1 <- read.csv(file.choose())
>
```

This command launches a dialog box. In this example, I have all of my Chapter 4 data sets in the same folder as my R syntax file. Putting everything for this chapter in the same folder saves time. We then selected ch4ds1.csv and hit Open. Next, let's look at the data set we imported. You can double-click on ch4ds1 in the Global Environment (top left). In the R Console, RStudio will now show this command:

```
> View(ch4ds1)
```

By clicking on the ch4ds1 object in the Global Environment, RStudio is prompted to send the `View()` function to the R Console.

Calculating the Mean

Earlier in the chapter, we used R to compute the mean much like you would with a calculator. Let's now use some built-in functions to fill in the blanks for Formula 4.1. Starting with the numerator—the sum of all *X*s—we will use `sum()` to add up the Prejudice scores.

```
> sum(ch4ds1$Prejudice)
[1] 1694
>
```

The sum function adds up all items in the object, in this case our vector of Prejudice scores. Our total is 1,694. For the denominator, we need to know how many scores we have. And R has a function for that!

```
> length(ch4ds1$Prejudice)
[1] 20
>
```

Filling in Formula 4.1, we will divide 1,694 by 20, which can be entered directly in the R Console, or you could put both functions on the same line.

```
> sum(ch4ds1$Prejudice) / length(ch4ds1$Prejudice)
[1] 84.7
>
```

R returns 84.7. Now let's use the built-in function, mean(), to quickly get our answer.

```
> mean(ch4ds1$Prejudice)
[1] 84.7
>
```

You can use R to calculate the mean by filling in the values for the formula, something we encourage as you are learning this simple formula, or go straight to the function to easily obtain the answer.

Finding the 50th Percentile: The Median

To find the middle score by hand, we can ask R to display all scores.

```
> ch4ds1$Prejudice
[1] 87 99 87 87 67 87 77 89 99 96 76 55 64 81 94 81 82 99 93 94
>
```

Not so helpful. We could sort these numbers automatically like we did several pages ago. Or, we can just use the built-in function, median().

```
> median(ch4ds1$Prejudice)
[1] 87
>
```

Recall, because we have an even number of observations, R needed to find the middle two scores and calculate the mean of those two numbers. In this example, the 10th and 11th scores were both 87, so the median is simply 87.

A Rose by Any Other Name: The Nominal Level of Measurement

Try this command for mode:

```
> mode(ch4ds1$Prejudice)
```

This time, we didn't get what we expected. R told us that the mode is "numeric." The function `mode()` in R returns the type of object; in this case, the Prejudice vector contains numbers. But what we really want to know is the most frequent response. Let's use `summary()` and `as.factor()`.

```
> summary(as.factor(ch4ds1$Prejudice))
55 64 67 76 77 81 82 87 89 93 94 96 99
 1  1  1  1  1  2  1  4  1  1  2  1  3
>
```

Starting with the vector:

- `ch4ds1$Prejudice`—the vector of scores
- `as.factor`—treat the vector of scores as a nominal vector. Up to now, we have been treating our scores like a ratio-level set of values.
- `summary`—summarize the object in the parentheses, in this case nominal Prejudice scores

By inspecting the results, we can see that one number occurred four times, more frequent than any other score. And that Prejudice score is 87. We can figure out the mode with our Prejudice scores, but the mode isn't really that helpful as a measure of central tendency. More useful is the mean or the median with our scores because of the level of our data.

R output can be full of information or just give you the basics. It all depends on the type of analysis that you are conducting. In the above examples, we have just the basics and, frankly, just what we need. Throughout *Statistics for People Who (Think They) Hate Statistics Using R,* you will be seeing output and then learning about what it means, but in some cases, discussing the entire collection of output information is far beyond the scope of the book. We'll focus on output that is directly related to what you learned in the chapter.

There are other ways to compute the central tendency:

Package	Function	What It Tells You
Base	summary	Minimum, 25th percentile, 50th percentile (median), mean, 75th percentile, and maximum

```
> summary(ch4ds1$Prejudice)
Min.  1st Qu.  Median  Mean  3rd Qu.  Max.
55.0    80.0     87.0  84.7    94.0   99.0
>
```

REAL-WORLD STATS

Few applications of descriptive statistics would make more sense than using them in a survey or poll, and there have been literally millions of such surveys (as in every U.S. presidential election). In an article, Roger Morrell and his colleagues examined Internet use patterns in 550 adults in several age groups, including middle aged (ages 40–59), young-old (ages 60–74), and old-old (ages 75–92). With a response rate of 71%, which is pretty good, they found a few very interesting (but not entirely unexpected) outcomes:

- There are distinct age and demographic differences among individuals who use the Internet.
- Middle-aged and older web users are similar in their use patterns.
- The two primary predictors for not using the Internet are lack of access to a computer and lack of knowledge about the Internet.
- Old-old adults have the least interest in using the Internet compared with middle-aged and young-old adults.
- The primary content areas in learning how to use the Internet are using electronic mail and accessing health information and information about traveling for pleasure.

This survey, which primarily used descriptive statistics to reach its conclusions, was done almost 20 years ago, and many, many things have changed since then. But the reason we are including it in this example of real-world statistics is that it is a good illustration of a historical anchor point that can be used for comparative purposes in future studies—a purpose for which descriptive statistics and such studies are often used.

Want to know more? Go online or go to the library and read . . .

Morrell, R. W., Mayhorn, C. B., & Bennett, J. (2000). A survey of world wide web use in middle-aged and older adults. *Human Factors, 42,* 175–182.

Summary

No matter how fancy schmancy your statistical techniques are, you will almost always start by simply describing what's there—hence the importance of understanding the simple notion of central tendency. From here, we go to another important descriptive construct: variability, or how different scores are from one another. That's what we'll explore in Chapter 5!

Time to Practice

1. By hand, compute the mean, median, and mode for the following set of 40 chemistry final scores.

93	85	99	77
94	99	86	76
95	99	97	84
91	89	77	87
97	83	80	98
75	94	81	85
78	92	89	94
76	94	96	94
90	79	80	92
77	86	83	81

2. Compute the mean, median, and mode for the following three sets of scores in Chapter 4 Data Set 2 (ch4ds2.csv). Do it by hand or R. Show your work, and if you use R, print out a copy of the output from the R Console.

Score 1	Score 2	Score 3
3	34	154
7	54	167
5	17	132
4	26	145
5	34	154
6	25	145
7	14	113
8	24	156
6	25	154
5	23	123

3. Compute the means for the following set of scores in Chapter 4 Data Set 3 (ch4ds3.csv) using R. Print out a copy of the output from the R Console.

Number of Beds (Infection Rate)	Infection Rate (per 1,000 Admissions)
234	1.7
214	2.4
165	3.1
436	5.6
432	4.9
342	5.3
276	5.6
187	1.2
512	3.3
553	4.1

4. You are the manager of a fast-food restaurant. Part of your job is to report to the boss at the end of each day which special is selling best. Use your vast knowledge of descriptive statistics and write one paragraph to let the boss know what happened today. Here are the data. Don't use R to compute important values; rather, do it by hand. Be sure to include a copy of your work.

Special	Number Sold	Cost
Huge Burger	20	$2.95
Baby Burger	18	$1.49
Chicken Littles	25	$3.50
Porker Burger	19	$2.95
Yummy Burger	17	$1.99
Coney Dog	20	$1.99
Total specials sold	119	

5. Imagine you are the CEO of a huge corporation and you are planning an expansion. You'd like your new store to post similar numbers as the other three that are in your empire. By hand, provide some idea of what you want the store's financial performance to look like.

(Continued)

(Continued)

And remember that you have to select whether to use the mean, the median, or the mode as an average. Good luck, young Jedi.

Average	Store 1	Store 2	Store 3	New Store
Sales (in thousands of dollars)	323.6	234.6	308.3	
Number of items purchased	3,454	5,645	4,565	
Number of visitors	4,534	6,765	6,654	

6. Here are ratings (on a scale from 1 through 5) for various Super Bowl party foods. You have to decide which food is rated highest (5 is a winner and 1 a loser). Decide what type of average you will use and why. Do this by hand or use R.

Snack Food	North Fans	East Fans	South Fans	West Fans
Loaded Nachos	4	4	5	4
Fruit Cup	2	1	2	1
Spicy Wings	4	3	3	3
Gargantuan Overstuffed Pizza	3	4	4	5
Beer Chicken	5	5	5	4

7. Under what conditions would you use the median rather than the mean as a measure of central tendency? Why? Provide an example of two situations in which the median might be more useful than the mean as a measure of central tendency.

8. Suppose you are working with a data set that has some very "different" (much larger or much smaller than the rest of the data) scores. What measure of central tendency would you use and why?

9. For this exercise, use the following set of 16 scores (ranked) that consists of income levels ranging from about $50,000 to about $200,000. What is the best measure of central tendency and why?

$199,999
$98,789
$90,878
$87,678
$87,245
$83,675
$77,876
$77,743

$76,564
$76,465
$75,643
$66,768
$65,654
$58,768
$54,678
$51,354

10. Use the data in Chapter 4 Data Set 4 (ch4ds4.csv) and, manually, compute the average attitude scores (with a score of 10 being positive and 1 being negative) for three groups (little experience, moderate experience, and lots of experience) of individuals' attitudes reflecting their experience with urban transportation.

11. Take a look at the following number of pie orders from the Lady Bird Diner and determine the average number of orders for each week.

Week	Chocolate Silk	Apple	Douglas County Pie
1	12	21	7
2	14	15	12
3	18	14	21
4	27	12	15

Student Study Site

Get the tools you need to sharpen your study skills! Visit **edge.sagepub.com/salkindshaw** to access practice quizzes and eFlashcards, watch R tutorial videos, and download data sets!