PART I

YIPPEE! I'M IN STATISTICS

"Well, if it isn't a significant subset of our study group!"

Not much to shout about, you might say? Let me take a minute and show you how some real-life scientists use this widely used set of tools we call *statistics*:

- Ruth Plackett, a researcher at the University College of London, is interested in the effect, if any, of social media use and our mental health. Dr. Plackett got access to data of more than 3,000 teenagers in the United Kingdom. At age 13 and 14, these kids were asked how many hours a day they used social media like Facebook and other sites. Two years later, they were asked questions about their mental health. Dr. Plackett found that there wasn't really a relationship between the two variables! Though sometimes apparent connections between exposure to social media and well-being in adolescents are observed, and make sense, the research team concluded that most of that apparent correlation is explained by the child's self-esteem and the quality of their friendships.

  Want to know more? Why not read the original work? You can find more about this in Plackett, R., Sheringham, J., & Dykxhoorn, J. (2023). The longitudinal impact of social media use on UK adolescents' mental health: longitudinal observational study. *Journal of Medical Internet Research, 25*, Article e43213.

- Why do some people eat meat and others don't? The traditional explanation has been that meat eaters like the taste of meat. Christopher A. Monteiro at Cornell University wondered if that was so or whether there was more to it than that. Might philosophical beliefs about animals play a role, as well? Dr. Monteiro created a measure to assess the level of two different beliefs: whether it is OK to eat animals and whether it is OK to kill animals if you are going to eat them. He collected data and used statistics to find relationships among the questions on his measure to identify that there seem to be these different attitudes that work with things like "meat tastes good" to explain the practice of meat eating. Interestingly, those who scored the highest on the "OK to kill animals" scale also tended to score the highest on measures of racism and sexism.

  Want to know more? You can read the whole study! Look this up:

  Monteiro, C. A., Pfeiler, T. M., Patterson, M. D., & Milburn, M. A. (2017). The Carnism Inventory: Measuring the ideology of eating animals. *Appetite, 113*, 51–62.

- Do you suffer from migraine headaches? Millions do. Doctors have long been concerned about the risks associated with migraines because there are often serious diseases associated with them. Dr. Tobias Kurth, a German researcher working with Harvard, analyzed data from thousands of women who have migraine headaches and had information about their health across decades. After equalizing the women on a bunch of variables, he discovered that women with migraines were 50% more likely to have heart disease and 62% more likely to have a stroke than those who did not have migraine headaches. The suggestion by Kurth and his coauthors is that women who suffer from migraines get more frequent checkups that include an evaluation for these other risks.

  Want to know more? Find out for yourself how the study was conducted and how statistics were used to get these estimates. Here are the details: Kurth, T., Winter, A. C., Eliassen, A. H., Dushkes, R., Mukamal, K. J., Rimm, E. B., . . .

Rexrode, K. M. (2016). Migraine and risk of cardiovascular disease in women: Prospective cohort study. *BMJ, 353*, Article i2610.

All of these researchers had a specific question they found interesting and used their intuition, curiosity, and excellent training to answer it. As part of their investigations, they used this set of tools we call *statistics* to make sense out of all the information they collected. Without these tools, all this information would have been just a collection of unrelated outcomes. And the research would be meaningless and not really research at all!

Statistics—the science of organizing and analyzing information to make it more easily understood—made these tasks doable. The reason that any of the results from such studies are useful is that we can use statistics to make sense out of them. And that's exactly the goal of this book—to provide you with an understanding of these basic tools and how researchers use them and, of course, how to use them yourself.

In this first part of *Statistics for People Who (Think They) Hate Statistics*, you will be introduced to what the study of statistics is about and why it's well worth your efforts to master the basics—the important terminology and ideas that are central to the field. This part gives you a solid preparation for the rest of the book.

1 STATISTICS OR SADISTICS? IT'S UP TO YOU

Difficulty Scale ☺ ☺ ☺ ☺ ☺ (very easy)

LEARNING OBJECTIVES

- **1.1** Describe the development of statistics as a useful scientific approach to understanding the world.
- **1.2** Define descriptive statistics like averages and explain how scientists use them.
- **1.3** Demonstrate how to compute means.
- **1.4** Demonstrate how to find the median.
- **1.5** Demonstrate how to find the mode.

WHY STATISTICS?

You've heard it all before, right?

"Statistics is hard and so different from anything else I've had to learn."

"I'm not a math person."

"I don't know how to use statistics software."

"What do I need this stuff for?"

"What do I do next?"

And the famous cry of the introductory statistics student: "I don't get it!"

Well, relax. Students who study introductory statistics find themselves, at one time or another, thinking some of these thoughts and quite possibly sharing them with another student, their partner, a colleague, or a friend.

And all kidding aside, some statistics courses can easily be described as *sadistics*. That's often because the books are repetitiously boring, the examples don't seem to apply to real life, and too much math is thrown at you too quickly.

That's not the case for you. The fact that you or your instructor has selected *Statistics for People Who (Think They) Hate Statistics* shows that you're ready to take the right approach—one that is unintimidating, informative, and applied (and even a little fun) and that tries to teach you what you need to know about using statistics as the valuable tool that it is.

If you're using this book in a class, it also means that your instructor is clearly on your side. They know that statistics can be intimidating but have taken steps to see that it is not intimidating for you. As a matter of fact, we'll bet there's a good chance (as hard as it may be to believe) that you'll be enjoying this class in just a few short weeks.

And Why SPSS?

Throughout this book, you'll be shown how to use SPSS, a statistical analysis tool, for the analysis of data. No worries; you'll also be shown much of the time how to do the same analysis by hand in case that's your preferred way to learn.

Why SPSS? Simple. It's one of the most popular, most powerful analytic tools available today, and it can be an exceedingly important and valuable tool for learning how to use basic and some advanced statistics. In fact, many stats courses taught at the introductory level use SPSS as their primary computational tool, and you can look to Appendix A for a summary or refresher on some basic SPSS tasks. Also, the way that technology is advancing, few opportunities to use statistics in research, administration, and everyday work will not require some knowledge of how and when to use tools such as SPSS. That's why we're including it in this book! We will show you how to use it to make your statistics learning experience a better one.

A 5-Minute History of Statistics

Before you read any further, it would be useful to have some historical perspective about this topic called statistics. After all, almost every undergraduate in the social, behavioral, and biological sciences and every graduate student in education, nursing, psychology, social welfare and social services, anthropology, and . . . (you get the picture) are required to take this course. Wouldn't it be nice to have some idea from whence the topic it covers came? Of course it would.

Way, way back, as soon as humans realized that counting was a good idea (as in "How many of these do I need to trade for one of those?" and "Uh oh, there are more of them than there are of us!" and "Yes, little one, you may keep one pet saber-toothed tiger, but not more than one"), collecting information became a useful skill.

If counting counted, then one could define the seasons by how often the sun rose and set, how much food was needed to last the winter, and what amount of resources belonged to whom. So, many of our early words were number words.

That was just the beginning. Once numbers became part of language, it seemed like the next step was to attach these numbers to outcomes. That started in earnest during the 17th century, when the first set of data pertaining to populations was collected. This was when the use of *descriptive statistics* began, which we will talk about later. From that point on, scientists (mostly mathematicians at first but, later, physical and biological scientists, and a little later than that, social scientists) needed to develop specific tools to answer specific questions. For example, Francis Galton (a half-cousin of Charles Darwin, by the way), who lived from 1822 to 1911, was very interested in the nature of human intelligence. (He also speculated that hair loss was due to the intense energy that went into thinking. It's probably not.) To explore one of his primary questions regarding the similarity of intelligence among family members, he used a specific statistical tool called the correlation coefficient (first developed by mathematicians), and then he popularized its use in the behavioral and social sciences. You'll learn all about this tool in Chapter 3.

Interestingly, most of the basic statistical procedures that you will learn about were first developed and used in the fields of agriculture, astronomy, and even politics. Their application to human behavior came much later.

PEOPLE WHO LOVED STATISTICS

Inferential statistics, the use of sample observations or data that we can see to make guesses about the likely characteristics of populations that we cannot see, probably started with Blaise Pascal (1623–1662), a French mathematician and religious philosopher. He developed the mathematical formulas that can predict important things like the probability of dice rolls and the likelihood of flipping a coin three times and having it come up heads each time. He even proved that if the coin almost always comes up heads, someone is almost certainly cheating. The application of these statistical inventions was of immediate practical use to gamblers, and this might be the first time in history that statistics was seen as having practical applications. (Because one could make money by understanding them.) You may notice that Pascal did not live very long. He suffered from various illnesses during the later years of his life, and his cause of death isn't even known for sure, although after his death, he was found to have had stomach cancer and some brain damage. As a deeply religious man, Pascal believed that suffering was necessary for a good life. Thus, he probably would have enjoyed being a stats professor.

The past 100 years have seen great strides in the invention of new ways to use old ideas. The simplest test for examining the differences between the averages of two groups was first advanced during the early 20th century. Techniques that build on this idea were offered decades later and have been greatly refined. And the introduction of personal computers and such programs as spreadsheets like Excel and software applications like SPSS have opened up the use of sophisticated techniques to anyone who wants to explore these fascinating topics.

The introduction of powerful personal computers and access to the internet and artificial intelligence apps (like ChatGPT) has been both good and bad. It's good because most statistical analyses no longer require access to a huge and expensive mainframe computer. Instead, a simple personal computer costing less than $250 or a cloud account or a smartphone can do 95% of what 95% of the people need. On the other hand, less than adequately educated students (such as your fellow students who chose not to take this course!) will take any old data they have and think that by running them through some sophisticated analysis, they will have reliable, trustworthy, and meaningful outcomes—not true. What your professor would say is "Garbage in, garbage out"; if you don't start with data you can trust, what you'll have after your data are analyzed are results you cannot trust.

Today, statisticians in all different areas, from criminal justice to geophysics to psychology to determining whether the "hot" hand really exists in the NBA, find themselves using basically the same techniques to answer different questions. There are, of course, important differences in how data are collected, but for the most part, the analyses (the plural of *analysis*) that are done following the collection of data (the plural of *datum*, which means one piece of information) tend to be very similar, even if called something different. The moral here? This class will provide you with the tools to understand how statistics are used in almost any discipline. Pretty neat, and all for just the cost of a few credit hours (and a book you will keep).

OK. Five minutes is up, and you know as much as you need to know for now about the history of statistics. You'll get some more history here and there as we learn about different procedures. Let's move on to what statistics is (and isn't).

DESCRIPTIVE STATISTICS AND AVERAGES

Statistics for People Who (Think They) Hate Statistics is a book about basic statistics and how to apply them to a variety of different situations, including the analysis and understanding of information, especially when that information is expressed as numbers and quantities.

In the most general sense, **statistics** describes a set of quantitative tools and techniques that are used for describing, organizing, and interpreting information or data. Those data might be the scores on a test taken by students participating in a special math curriculum, the speed with which problems are solved, the number of side effects when patients use one type of drug rather than another, the number of errors in each inning of a World Series game, or the average price of a dinner in an upscale restaurant in Santa Fe, New Mexico (not cheap).

In all these examples, and the million more we could think of, data are collected, organized, summarized, and then interpreted. In this section, you'll learn about collecting, organizing, and summarizing data as part of descriptive statistics. And in the next section, you'll learn about interpreting data when you learn about the usefulness of inferential statistics.

Descriptive statistics are used to organize and describe the characteristics of a collection of data. The collection is sometimes called a **data set** or just **data**. Scientists would say that descriptive statistics describe a *sample*—a collection of data that you have in front of you. Some common ways to describe data are to average a bunch of values, summarize the frequency at which each value occurs, and quantify how much each score varies from the other. Here, we will explore averages (there are different kinds), and we will look at other descriptive statistics in Chapter 2.

An **average** is the one value that best represents an entire group of scores. It doesn't matter whether the group of scores represents the number correct on a spelling test for 30 fifth graders or the typical batting percentage for all the baseball players on the Kansas City Royals or how voters feel about a congressional candidate. In all of these examples, a big group of data can be summarized using an average. You can usually think of an average as the "middle" space or as a fulcrum on a seesaw. It's the point in a range of values that seems to most fairly represent all the values.

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 so popular that scientists sometimes sloppily treat the word *average* as if it means *mean* when it only sometimes means *mean*. The mean 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.

We are about to show a formula or equation for the first time in this book. Don't panic. Equations are just statements or sentences that use symbols instead of words. We will always tell you what words the symbols stand for. The formula for computing the mean is shown in Formula 1.1:

$$\bar{X} = \frac{\sum X}{n} \tag{1.1}$$

where

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

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. Compute the sum or total of all the values.
3. Divide the total or sum 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 (Table 1.1).

TABLE 1.1 ■ Shopper Data

Location	Number of Shoppers Last Year
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 1.2 shows how this average was computed using the formula you saw in Formula 1.1:

$$\bar{X} = \frac{\sum X}{n} = \frac{2{,}150 + 1{,}534 + 3{,}564}{3} = \frac{7{,}248}{3} = 2{,}416 \qquad (1.2)$$

Or, if you needed to compute the mean number of students in each grade in a school building, you would follow the same procedure (Table 1.2).

TABLE 1.2 ■ Student Data

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 1.3 shows how this average was computed using the formula you saw in Formula 1.1:

$$\bar{X} = \frac{\sum X}{n} = \frac{18 + 21 + 24 + 23 + 22 + 24 + 25}{7} = \frac{157}{7} = 22.43 \qquad (1.3)$$

See, we told you it was easy. No big deal. By the way, when you calculated that mean just now, you may have gotten a number with lots more digits in it: 22.42857143 or something like that. Statisticians are usually OK with you shortening numbers to just a couple digits past the decimal. So, we felt fine reporting the mean as 22.43 (rounding up for that last digit).

- The mean is sometimes represented by the letter *M* and is also called the typical, average, or most central 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. Technically, that capital letter *M* is used when you are talking about the mean of the larger population represented by the sample in front of you. Those sorts of distinctions aren't important right now but might be interesting later on.
- 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. Notice, as with the capital *M* when talking about the mean of a population,

statistical types often capitalize a letter symbol to refer to a population and keep the letter as lowercase when talking about samples.

- 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.

COMPUTING THE MEDIAN

The median is also an average but of a very different kind. The **median** is defined as the midpoint in a distribution or 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 not really a formula for computing the median but instead a set of steps.

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 annual incomes from five different households:

$135,400

$45,500

$62,456

$54,365

$37,668

Here is the list ordered from highest to lowest:

$135,400

$62,456

$54,365

$45,500

$37,668

There are five values. The middle-most value is $54,365, and that's the median.

Now, what if the number of values is even? An even number of scores means there is no middle value. Let's add a value ($64,500) to the list so there are six income levels. Here they are sorted with the largest value first:

$135,400

$64,500

$62,456

$54,365

$45,500

$37,668

When there is an even number of values, there is no middle score! In that case, the median is simply the mean of the two middle values. In this case, the middle two cases are $54,365 and $62,456. The mean of those two values is $58,410.50. That's the median for that set of six values.

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 rehabilitating an injury is 32.

If you know about medians, you might also be interested in **percentile ranks**. Percentile ranks are used to define the percentage of cases equal to or 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 at or 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 not sensitive to extreme scores! When you have a set of scores in which one or more scores are extreme, the median better represents the center-most 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. Later in life, when you do more sophisticated statistical analyses, you might call these occasional extreme scores **outliers**.

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 closer to most of the scores, and we like to think about "the average" 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. Think of those super-rich billionaires. As an example, the median annual family income in the United States for 2022 (the most recent year we could find data for) was about $90,000, while the mean annual family income was about $123,000. Which is closer to your family's income?

> 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:

- We use the word *median* in the real world to describe other middle things, like the median on a highway, that stripe down the middle of a road.
- Because the median is based on how many cases there are, and not the values of those cases, extreme scores only count a little.

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 sample of scores. The **mode** is the value that occurs most frequently. Like the median, there is no formula for computing the mode; rather, there's a procedure.

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 distribution of scores shown in Table 1.3.

TABLE 1.3 ■ Political Party Data

Party Affiliation	Number or Frequency
Democrats	90
Republicans	70
Independents or None	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 wanted to use scores to represent those three categories, we might arbitrarily label them as 1 = Democrats, 2 = Republicans, and 3 = Independents. If we did that, the mode would be *3*.

If we were looking at the modal response on a single question on a multiple-choice test, we might find that the correct answer (in this case, option A) was chosen more frequently than any other. The data might look like Table 1.4.

Want to know the easiest and most common mistake when identifying 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 or 3 (in our first example), it's easy for someone to conclude the mode is 140. Why? Because they are looking at the number of times the value occurred, not the value that occurred most often! This is a simple mistake to make, so be on your toes when you are asked about these things.

TABLE 1.4 ■ Multiple-Choice Question Data

Answer Option Selected	A*	B	C	D
Number of times	57	20	12	11

*Correct answer option

Inferential Statistics: The Other Kind of Statistics

Inferential statistics are often (but not always) the next step after you have collected and summarized data. Inferential statistics are when you use information about a sample to make guesses about a population. You might look at a sample mean and infer that it is a good estimate of the population mean. That's why we call it inferential statistics.

A smaller group of data is usually called a **sample**, which is a portion, or a subset, of a **population**. For example, all the fifth graders in Newark, New Jersey (Neil's fair city of origin), would be a population. The population is the large group of people or things you want to describe. In this case, it's all the children in fifth grade and attending school in Newark. We can't observe or measure a whole population, so we'd select a smaller number of students. That smaller number (30 children, 50, 150, whatever) is a sample. If we think this sample represents the population really well, we can make guesses about the population. That seem reasonable?

Let's look at another example. Your marketing agency asks you (a newly hired researcher) to determine which of several names is most appealing for a new brand of potato chip. Will it be Chiperinos? SuperFunChips? Crunchies? As a statistics pro (we know we're moving a bit ahead of ourselves, but keep the faith), you need to find a small group of potato chip eaters who are representative of all potato chip eaters and ask these people to tell you which one of the three names they like the most. Then, if you do things right, you can easily extrapolate the findings to the huge group of all potato chip eaters that you are trying to sell your potato chips to.

Or let's say you're interested in the best treatment for a particular type of disease. Perhaps you'll try a new drug as one alternative, a placebo (a substance that is known not to have any effect) as another alternative, and nothing as the third alternative to see what happens. Well, maybe you find out that more patients get better when no action is taken and nature just takes its course! If so, you might conclude that your drug does not have any effect. (You can't be sure, of course, unless you had the most perfectly designed study ever!)

Then, with that information, you can guess that if a larger group of patients who suffer from the disease took your drug, it also wouldn't work. This is an inference and does more than just describe your sample.

Inferring from a sample to a population makes a lot of sense, especially when you are sure the sample represents the population. That's why, as you'll see later, scientists spend a lot of effort getting a representative sample. And, as you might guess, much of the rest of this book is about the rules and math involved in making that guess as accurately as possible.

In Other Words . . .

Statistics is a tool that helps us understand the world around us. It does so by organizing information we've collected and then letting us make certain statements about how

characteristics of those data are applicable to new settings. Descriptive and inferential statistics work hand in hand, and which statistic you use and when you use it depends on the question you want answered and how you happened to measure your variables.

Today, a knowledge of statistics is more important than ever because it provides us with the tools to make decisions that are based on empirical (observed) evidence and not our own biases or beliefs. Want to know whether early intervention programs work? Then test whether they work and provide that evidence to the court that will make a ruling on the viability of a new school bond issue that could pay for those programs.

WHAT AM I DOING IN A STATISTICS CLASS?

You might find yourself using this book for many reasons. You might be enrolled in an introductory statistics class. Or you might be reviewing for your comprehensive exams. Or you might even be reading this on summer vacation (horrors!) in preparation for a more advanced class.

In any case, you are a statistics student, whether you have to take a final exam at the end of a formal course or you're just in it of your own accord. But there are plenty of good reasons to be studying this material—some fun, some serious, and some both.

Here's the list of some of the things that our students hear at the beginning of our introductory statistics courses:

1. Statistics 101 or Statistics 1 or whatever it's called at your school looks great listed on your transcript. But, also, this may be a required course for you to complete your major. Even if it is not required, having these skills is definitely a big plus when it comes time to apply for a job or for further schooling. And with more advanced courses, your résumé will be even more impressive.

2. If this is not a required course, taking basic statistics sets you apart from those who do not. It shows that you are willing to undertake a course that is above average with regard to difficulty and commitment. And, as the political and economic (and sports!) worlds become more "accountable," more emphasis is being placed on analytic skills. Who knows, this course may be your ticket to a job!

3. Basic statistics is an intellectual challenge of a kind that you might not be used to. There's a good deal of thinking that's required, a bit of math, and some integration of ideas and application. The bottom line is that all this activity adds up to what can be an invigorating intellectual experience because you learn about a whole new area or discipline.

4. There's no question that having some background in statistics makes you a better student in the social or behavioral sciences, because you will have a better understanding not only of what you read in journals but also of what your professors and colleagues may be discussing and doing in and out of class. You will be amazed the first time you say to yourself, "Wow, I actually understand what they're talking about." And it will happen over and over again, because you will have the basic tools necessary to understand exactly how scientists reach the conclusions they do.

5. If you plan to pursue a graduate degree in education, anthropology, economics, nursing, sociology, or any one of many other social, behavioral, and biological pursuits, this course will give you the foundation you need to move further.

6. There are many different ways of thinking about, and approaching, different types of problems. The set of tools you learn about in this book (and this course) will help you look at interesting problems from a new perspective. And, while the possibilities may not be apparent now, this new way of thinking can be brought to new situations.

7. Finally, you can brag that you completed a course that everyone thinks is the equivalent of building and running a nuclear reactor.

TEN WAYS TO USE THIS BOOK (AND LEARN STATISTICS AT THE SAME TIME!)

Yep. Just what the world needs—another statistics book. But this one is different. It is directed at the student, is not condescending, is informative, and is as basic as possible in its presentation. It makes no presumptions about what you should know before you start and proceeds in slow, small steps, which lets you pace yourself.

However, there has always been a general aura surrounding the study of statistics that it's a difficult subject to master. And we don't say otherwise, because parts of it are challenging. On the other hand, millions and millions of students have mastered this topic, and you can, too. Here are 10 hints to close this introductory chapter before we move on to our first topic:

1. **You're not dumb.** That's true. If you were, you would not have gotten this far in school. So, treat statistics as you would any other new course. Attend the lectures, study the material, do the exercises in the book and from class, and you'll do fine. Rocket scientists know statistics, but you don't have to be a rocket scientist to succeed in statistics.
2. **How do you know statistics is hard?** Is statistics difficult? Yes and no. If you listen to friends who have taken the course and didn't do well, they'll surely volunteer to tell you how hard it was and how much of a disaster it made of their entire semester, if not their lives. And let's not forget—we always tend to hear from complainers. So, we'd suggest that you start this course with the attitude that you'll wait and see how it is and judge the experience for yourself. Better yet, talk to several people who have had the class and get a good idea of what they think. Don't base your expectations on just one spoilsport's experience. Get a bigger sample!
3. **Don't skip lessons—work through the chapters in sequence.** *Statistics for People Who (Think They) Hate Statistics* is written so that each chapter provides a foundation for the next one in the book. When you are all done with the course, you will (I hope) continue to use this book as a reference. So if you need a particular value from a table, you might consult Appendix B. Or if you need to remember how to compute the standard deviation, you might turn to Chapter 2. But for now, read each chapter in the sequence that it appears. It's OK to skip around and see what's offered down the road. Just don't study later chapters before you master earlier ones.
4. **Form a study group.** This is a big hint and one of the most basic ways to ensure some success in this course. Early in the semester, arrange to study with friends or classmates. If you don't have any friends who are in the same class as you, then make some new ones or offer to study with someone who looks as happy to be there as you are. Studying with others allows you to help them if you know the material better or to

benefit from those who know some material better than you. Set a specific time each week to get together for an hour and go over the exercises at the end of the chapter or ask questions of one another. Take as much time as you need. Studying with others is an invaluable way to help you understand and master the material in this course.

5. **Ask your teacher questions, and then ask a friend.** If you do not understand what you are being taught in class, ask your professor to clarify it. Have no doubt—if you don't understand the material, then you can be sure that others do not as well. More often than not, instructors welcome questions. And especially because you've read the material before class, your questions should be well informed and help everyone in class to better understand the material.

6. **Do the exercises at the end of a chapter.** The exercises are based on the material and the examples in the chapter they follow. They are there to help you apply the concepts that were taught in the chapter and build your confidence at the same time. If you can answer these end-of-chapter exercises, then you are well on your way to mastering the content of the chapter. Correct answers to each exercise are provided in Appendix D.

7. **Practice, practice, practice.** Yes, it's a very old joke:

   Q How do you get to Broadway?

   A Practice.

Well, it's no different with basic statistics. You have to use what you learn and use frequently to master the different ideas and techniques. This means doing the exercises at the end of Chapters 1 through 17 and Chapter 19, as well as taking advantage of any other opportunities you have to understand what you have learned.

8. **Look for applications to make it more real.** In your other classes, you probably have occasion to read journal articles, talk about the results of research, and generally discuss the importance of the scientific method in your own area of study. These are all opportunities to see how your study of statistics can help you better understand the topics under class discussion as well as the area of beginning statistics. The more you apply these new ideas, the fuller your understanding will be.

9. **Browse.** Read over the assigned chapter first; then go back and read it with more intention. Take a nice leisurely tour of *Statistics for People Who (Think They) Hate Statistics* to see what's contained in the various chapters. Don't rush yourself. It's always good to know what topics lie ahead as well as to familiarize yourself with the content that will be covered in your current statistics class.

10. **Have fun.** This might seem like a strange thing to say, but it all boils down to you mastering this topic rather than letting the course and its demands master you. Set up a study schedule and follow it, ask questions in class, and consider this intellectual exercise to be one of growth. Mastering new material is always exciting and satisfying—it's part of the human spirit. You can experience the same satisfaction here—just keep your eye on the ball and make the necessary commitment to stay current with the assignments and work hard.

ABOUT THE BOOK'S FEATURES

Throughout the book, there are short biographies of People Who Loved Statistics. All the statistical tricks and procedures we will discover in this book were invented by real people, and it's good to realize that they were just like you and me!

In some places, you'll find a small-steps icon like the one you see here. This indicates that a set of steps is coming up that will direct you through a particular process. Sometimes you will use SPSS to do these steps. These steps have been tested and approved by whatever federal agency approves these things.

Real-World Stats

Real-World Stats will appear at the end of most chapters as appropriate and, it is hoped, will provide you with a demonstration of how a particular method, test, idea, or some aspect of statistics is used in the everyday workplace of scientists, physicians, policy makers, government folks, and others. That sort of thing began this chapter, but we will add one more.

Let's look at a very short paper in which the author recalls and shares the argument that the National Academy of Sciences (first chartered in 1863, by the way!) "shall, whenever called upon by any department of the Government, investigate, examine, experiment, and report upon any subject of science or art." This charter, some 50 years later in 1916, led to the formation of the National Research Council, another federal body that helped provide information that policy makers need to make informed decisions. And often this "information" takes the form of quantitative data—also referred to as statistics—that assist people in evaluating alternative approaches to problems that have a wide-ranging impact on the public. So, this article, as does your book and the class you are taking, points out how important it is to think clearly and use data to support your arguments.

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

Cicerone, R. (2010). The importance of federal statistics for advancing science and identifying policy options. *Annals of the American Academy of Political and Social Science, 631*, 25–27.

Appendix A contains an introduction to SPSS. Working through this appendix is all you really need to do to be ready to use SPSS. If you have an earlier version of SPSS (or the Mac version), you will still find this material to be very helpful. In fact, the latest Windows and Mac versions of SPSS are almost identical in appearance and functionality.

Appendix B contains important tables you will learn about and need throughout the book.

In working through the exercises in this book, you will use the data sets in Appendix C. In the exercises, you'll find references to data sets with names like "Chapter 2 Data Set 1,"

and each of these sets is shown in Appendix C. You can either enter the data manually or download them from the publisher's site at: go to **collegepublshing.sagepub.com** and search for this book, then click on the Resources tab.

Appendix D contains answers to end-of-chapter questions.

Appendix E contains a primer on math for those who could use a refresher.

Appendix F contains the most helpful hints for gathering your own data.

And Appendix G offers Neil's long-sought-after brownie recipe (yes, you finally found it).

KEY TO DIFFICULTY ICONS

To help you along a bit, we placed a difficulty index at the beginning of each chapter. This adds some fun to the start of each chapter, but it's also a useful tip to let you know what's coming and how difficult chapters are in relation to one another. Because the index uses smiley faces, the more smiles, the merrier!

- ☺ (very hard)
- ☺ ☺ (hard)
- ☺ ☺ ☺ (not too hard, but not easy either)
- ☺ ☺ ☺ ☺ (easy)
- ☺ ☺ ☺ ☺ ☺ (very easy)

SUMMARY

That couldn't have been that bad, right? We talked about statistics in general but also covered and actually computed by hand (with a calculator, or maybe even just your brain and a pen or pencil!) three important ways to average or fairly represent a bunch of values! That might have been easy or mildly stressful for you. However you felt, please continue reading and don't worry about what's difficult or time-consuming or too complex for you to understand and apply. Just take one chapter at a time, as you did this one.

KEY TERMS

Average
Data
Data set
Descriptive statistics
Inferential statistics:
Mean
Measures of central tendency
Median
Mode
Outliers
Percentile rank
Population
Sample
Skew
Statistics

ACTIVITIES

Because there's no substitute for the real thing, most chapters will end with a set of Activities, Review Questions, and Critical Thinking Questions that will help you review the material that was covered in the chapter. The activities are to give you some experience applying the ideas in each chapter or to talk about in class. They have no right answers. But the questions do have right answers and those answers can be found near the end of the book in Appendix D.

For example, here is the first set of Activities:

1. Interview someone who uses statistics in their everyday work. It might be your advisor, an instructor, a researcher who lives on your block, a health care analyst, a marketer for a company, a city planner, or someone representing almost any profession. Ask the person what their first statistics course was like. Find out what the person liked and didn't like. See if this individual has any suggestions to help you succeed. And most important, ask the person about how they use these new-to-you tools at work.
2. We hope that you are part of an in-person study group or, if that is not possible, that you have more than one online study buddy and, of course, plenty of texting and Instagram friends. Talk to your group or a fellow student in your class about similar likes, dislikes, fears, and so on about the statistics course. What do you have in common? Not in common? Discuss with your fellow student strategies to overcome your fears.
3. Search through hard-copy or online magazines or news sites and find the results of a survey or interview about any topic. Summarize the results and do the best job you can describing how the researchers who were involved, or the authors of the survey, came to the conclusions they did. Their methods and reasoning may or may not be apparent. Once you have some idea of what they did, try to speculate as to what other ways the same information might be collected, organized, and summarized.
4. Find a copy of a journal article in your own discipline. This can be at the library or online or in your professor's office or whatever. Then, go through the article and highlight the section (usually the "Results" section) where statistical procedures were used to organize and analyze the data. You don't know much about the specifics of this yet, but how many different statistical procedures (such as *t* test, mean, and calculation of the standard deviation) can you identify? Can you take the next step and tell your instructor how the results relate to the research question or the primary topic of the research study?

REVIEW QUESTIONS

1. What measure of central tendency do most people mean when they say "average"?
   - **A.** Mean
   - **B.** Median
   - **C.** Mode

2. Would a teacher probably use descriptive or inferential statistics to summarize how their class did on a quiz?
   A. Descriptive
   B. Inferential
   C. Both
   D. Neither

3. If you wanted to estimate the attitude of a population of people, would you use descriptive or inferential statistics?
   A. Descriptive
   B. Inferential
   C. Both
   D. Neither

4. Which is largest: a distribution of scores, a sample, or a population?
   A. A distribution
   B. A sample
   C. A population

5. Why do we call averages measures of central tendency?
   A. They are the most important descriptive statistic.
   B. They identify the most representative value.
   C. They are in the middle of a distribution.
   D. They describe trends over time.

6. Which of the following is an example of **descriptive statistics**?
   A. Predicting future population growth based on past trends
   B. Calculating the average test score of a class
   C. Testing whether a new drug is more effective than an old one
   D. Determining if there is a relationship between two variable

7. Why do scientists use **measures of central tendency** like the mean, median, and mode?
   A. To summarize large amounts of data in a meaningful way
   B. To establish cause-and-effect relationships
   C. To manipulate data and get the results they want
   D. To determine the statistical significance of an experiment

8. If a teacher wants to calculate the **mean** test score for a class, what steps should they take?
   A. Identify the most common score that appears in the data set.
   B. Add up all the test scores and divide by the number of students.
   C. Arrange all test scores in order and find the middle number.
   D. Subtract the lowest score from the highest score.

9. A data set consists of the following numbers: **4, 7, 9, 10, 12**. What is the **median** of this data set? _____

10. A data set contains the values 3, 5, 7, 7, 10, 12, 12, 12, 15. What is the mode? _____

CRITICAL THINKING QUESTIONS

1. Which is probably easiest to calculate?
   - **A.** The mean of a sample
   - **B.** The mean of a population

2. In inferential statistics, what is always being used to infer to a population?
   - **A.** The highest score
   - **B.** The lowest score
   - **C.** A sample
   - **D.** A measure of central tendency

2 WHAT DO YOUR DATA LOOK LIKE?

Summarizing and Picturing Distributions

Difficulty Scale ☺ ☺ ☺ ☺
(moderately easy but not a cinch)

LEARNING OBJECTIVES

- **2.1** Compare and contrast the four levels of measurement.
- **2.2** Compare and contrast three measures of variability—the range, the variance, and the standard deviation.
- **2.3** Correctly compute the standard deviation.
- **2.4** Use SPSS to compute descriptive statistics.
- **2.5** Compare and contrast histograms, polygons, bar charts, line charts, and pie charts.
- **2.6** Graph data using SPSS to create histograms, polygons, bar charts, line charts, and pie charts.

Last chapter, we started talking about what statistics is (or are?) and we even began calculating some popular descriptive statistics like averages (measures of central tendency—the mean, median and mode). In this chapter, we will learn descriptive statistics that summarize the variability of a distribution (how much the scores vary from each other) and also how to use graphs to draw a picture of the "shape" of a distribution. Most excitingly, though, we will actually begin using SPSS to do all of this. First, though, let's talk more about levels of measurement. Because the level of measurement of our data tells us which average, which type of variability, and which graph we should use.

HOW MUCH INFORMATION IS IN YOUR VARIABLE?

You might remember from Chapter 1 that which measure of central tendency or type of average you use depends on certain characteristics of the data you are working with. Basically, what we said, using different words, was

- If your scores just represent categories, use the mode.
- If your scores are spread out, but not evenly spaced, use the median.
- If your scores are spread out *and* evenly spaced, use the mean.

These three types of data are called **levels of measurement** and are actually in order of how much information is in the distribution. There's even a fourth higher level of measurement when your scores represent a count or frequency of something.

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. That's pretty straightforward and simple. Notice this definition includes values, which are numbers, and involves giving meaning to those numbers. The result of measurement is *outcomes*, a group of scores that provide different amounts of information about things and concepts like hair color, gender, knowledge, or height.

Measurement experts and statisticians have identified four different ways these scores can provide information. Each way is called a level, and each level allows for more information than the one below it. The four levels of measurement are called nominal, ordinal, interval, and ratio. Let's get to know these four levels.

A Rose by Any Other Name: Nominal Level

The **nominal level of measurement** describes situations where the numbers are used only as names for different categories. The Latin root *nom* means "name." For example, which country you are from. If you are from Canada, we might code that as a 1 in our data. Some other country would get another number. One category isn't more or less than another. The numbers are just used as labels.

Which Came First, the Chicken or the Egg? Ordinal Level

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

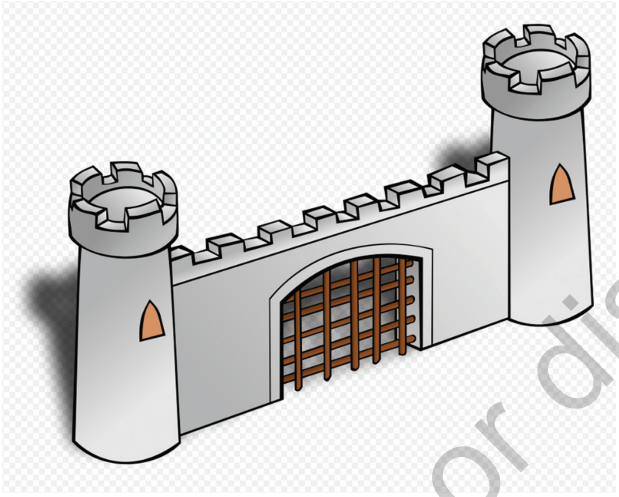
Equal Space: Interval Level

Now we're getting somewhere. When we talk about the **interval level of measurement** we are thinking about the distances or intervals between any two adjacent scores anywhere along the scale. If all those intervals are "equal" along the whole range of possible scores, we call that interval level. This is one of the more difficult ideas in this chapter, so bear with us. But here's what we mean.

Think of a Celsius thermometer, the difference between 48 degrees and 47 degrees is one degree of heat (whatever that means), and the difference between 34 degrees and 33 degrees is also one degree of heat. Everywhere along the scale, there is an equal amount of difference in the variable (heat) between any two scores that are side by side. In other words, a mathematical difference between two scores is the same as the quantitative difference between units of the variable.

It might help to look at the architectural origins of the word *interval*. Interval, or *inter val*, means "between walls." It describes the top of towers and such on old castles like in this picture. Notice those little stone protective barriers at the top of the towers? Well, a well-designed fortress like a castle needed protection for the guards at the gate shooting

arrows, but there also had to be openings to shoot those arrows at the bad guys. The best design had equal spacing between those barriers all the way around because guards never knew where the enemy would be coming from. So those intervals have equal spacing, like an interval-level scale!



Much Ado About Nothing: Ratio Level

Here's a riddle for you. Whatever variable you want your scores to represent, can there be zero of it? Scores at the **ratio level of measurement** are characterized by the presence of an absolute zero on the scale. On ratio-level scales, there is nothing less than zero, that is, no negative numbers.

In the social and behavioral sciences, we usually measure concepts that everyone has some of. Even if you score zero on that spelling test or miss every item of an intelligence test (written in Martian), that does not mean that you have no spelling ability or no intelligence, right? Even many physical scales, like the thermometer we talked about earlier, may have a 0 on them, but you can still get scores below zero. And that 0 degrees Fahrenheit does not mean there is no heat!

In the field you are likely studying, the only time you will be at the ratio level is probably if you are counting things. How many children does each respondent have? Those variables will be at the ratio level.

It's called the ratio level because it's the only level where talking about ratios or percentages makes sense. For example, if it's 60 degrees Fahrenheit today and was 30 degrees yesterday, we don't say, "It's twice as hot today!" Because that scale is only at the interval level. But if you have four cats, and Bruce has two cats, we would say you have twice as many cats as Bruce. Because that variable is at the ratio level. Sometimes it's hard to decide if a variable is interval level or ratio level. For statistical purposes, it doesn't really matter much because if you are at least as high as interval level in your scaling, you use the same statistics as you would at the ratio level. It's more of a theoretical conversation than a practical one.

Okay, we've defined four levels of measurement. The reason we did is that for the rest of the book, starting with this chapter's extension of descriptive statistics, whenever you have a choice of which statistic or analysis to use, the answer will almost always depend

on which level of measurement you are at! We will start a table here (Table 2.1) to match the different averages we've learned already with their level of measurement. Then, we will add to it later as we cover measures of variability and types of graphs. (The way the empty rows are split for measures of variability and graphs is a little foreshadowing to add suspense!)

TABLE 2.1 ■ Level of Measurement and Measures of Central Tendency

Level of Measurement	Average	Measure of Variability	Graph
Nominal	Mode		
Ordinal	Mode Median		
Interval	Mode Median Mean		
Ratio			

VIVE LA DIFFÉRENCE! UNDERSTANDING VARIABILITY

We know that a great way to describe a distribution using only one number is to use an average, like the mean. But when it comes to describing the characteristics of a distribution, averages are only half the story. We also need to use some number that describes variability.

In the simplest of terms, **variability** reflects how much scores differ from one another. For example, the following set of scores differ from each other and has a mean of 4.

7, 6, 3, 3, 1

And this set of scores also differ from each other and has a mean of 4:

3, 4, 4, 5, 4

And finally, here's a third set of scores that do not differ from each other at all and has a mean of 4:

4, 4, 4, 4, 4

Maybe you can see the problem! Summarizing those three very different distributions with just the mean of 4 doesn't give the whole picture. And it is pretty misleading, as well, right? Because the amount of variability is substantially different in each distribution.

How data points differ from one another is a central part of understanding and using basic statistics. But when it comes to differences between individuals and groups (a mainstay of most social and behavioral sciences), the whole concept of variability becomes really important. Sometimes it's called fluctuation, or error, or one of many other terms, but the fact is, variety is the spice of life, and what makes people different from one another also makes understanding them and their behavior all the more challenging (and interesting). Without variability in a set of data or between individuals and groups, things are just boring.

Three measures of variability are commonly used to reflect the degree of variability, spread, or dispersion in a group of scores. These are the range, the standard deviation, and the variance. Let's take a closer look at each one and how each one is used.

The Range

The **range** is the simplest measure of variability and kind of intuitive. It is the distance of the biggest score from the smallest score. The range is computed simply by subtracting the lowest score in a distribution from the highest score in the distribution.

In general, the formula for the range is

$$r = l - s \tag{2.1}$$

where

- r is the range,
- l is the largest score in the data set, and
- s is the smallest score in the data set.

Take the following set of scores, for example (shown here in descending order):

98, 86, 77, 56, 48

In this example, 98 – 48 = 50. The range is 50. Even if there was a set of 500 scores, where the largest is 98 and the smallest is 48, the range would still be 50.

In real life, the range is hardly ever reported. But it sometimes gives some interesting information.

The Variance

Here comes another measure of variability. The **variance** is the average amount of distance of each score from the mean. (The "average" we are talking about here is the mean, but we didn't want to have to say mean twice in the same sentence!) In other words, you take each score, subtract the mean from it to create a **difference score**, and then you add up all those difference scores. Finally, you calculate the mean of those distance scores. Actually, it's not quite that simple. The problem is that about half of those difference scores will be negative, and if you add them all up they will equal zero! In fact, that's the defining quality of the mean as an average, but in this case it is a problem. So, to eliminate that problem, statisticians square each difference score before adding them up. (Multiplying any number by itself makes it positive.) All these complications result in a moderately frightening looking equation:

$$s^2 = \frac{\sum(X-\bar{X})^2}{n-1} \tag{2.2}$$

where

- s^2 is the variance,
- $\sum$ is a symbol that tells you to find the sum of what follows,
- X is each individual score,

- $\bar{X}$ is the mean of all the scores, and
- n is the sample size.

Other than having to square each distance, there is another weirdness in this equation. To get the mean, notice we divide the sum of all those difference scores by $n - 1$, the number of scores minus one, not by n, the number of scores. This is because in inferential statistics, we usually are using the sample to estimate the larger population, and it turns out that this "subtract 1 from n before dividing" adjustment makes the population estimate more accurate.

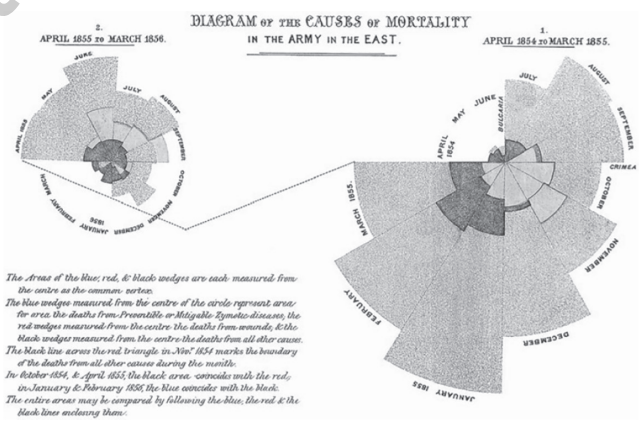
You are not likely to see the variance mentioned by itself in a journal article or see it used as a descriptive statistic (so we won't bother to calculate one here). This is because the variance is a difficult number to interpret and get useful meaning from. After all, squaring those difference scores changes the original scale dramatically.

But the variance is important because it is used as a key component in other statistical formulas, such as that of the standard deviation, which we will learn about next.

PEOPLE WHO LOVED STATISTICS

You may know about the famous English nurse Florence Nightingale (1820–1910), who managed and trained nurses in the Crimean War and developed standards and procedures to make nursing a profession. Also, though, she was a gifted mathematician and was influential in promoting the collection and analysis of statistical data. She was one of the first medical researchers to use graphs and charts to display the variability of data.

Nightingale used these visual forms of communication to examine causes of death for soldiers and to convince administrators that environmental conditions, like unsanitary water, affected death rates. Her application of statistics to real-world problems of life and death, such as sanitation and other issues, is believed to have dramatically increased life expectancy throughout English towns in the latter half of the 19th century.



THE STANDARD DEVIATION

Now we get to the most frequently used measure of variability, and one that is actually reported in scientific papers—the standard deviation. Just think about what the term implies—it's a deviation from something (guess what?) that has been standardized. Actually, the **standard deviation** (sometimes abbreviated as *SD*, sometimes *s*) represents the average amount of variability in a set of scores. In practical terms, it's the average distance of each score from the mean. The larger the standard deviation is, the larger the average distance each data point is from the mean of the distribution and the more variety there is in the set of scores.

But wait! Isn't that the same definition as variance? Not quite. Remember the variance is the average value of the squared differences. The standard deviation unsquares everything at the end so we get back to the original scale of the scores. It is so much more meaningful that way.

Here's the formula for computing the standard deviation:

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{n - 1}}, \tag{2.3}$$

where

- s is the standard deviation;
- $\sum$ is sigma, which tells you to find the sum of what follows;
- X is each individual score;
- $\overline{X}$ is the mean of all the scores; and
- n is the sample size.

This formula finds the difference between each individual score and the mean $(X - \overline{X})$, squares each difference, and sums them all together. Then, it divides the sum by the size of the sample (minus 1) and takes the square root of the result.

Here are the data we'll use in the following step-by-step explanation of how to compute the standard deviation:

5, 8, 5, 4, 6, 7, 8, 8, 3, 6

1. List each score. It doesn't matter whether the scores are in any particular order.
2. Compute the mean of the group.
3. Subtract the mean from each score.
4. Square each individual difference. The result is the column marked $(X - \overline{X})^2$.
5. Sum all the squared deviations about the mean. As you can see, the total is 28.
6. Divide the sum by n – 1, or 10 – 1 = 9, so then 28/9 = 3.11.
7. Compute the square root of 3.11, which is 1.76 (after rounding). That is the standard deviation for this set of 10 scores.

Here's what we've done so far (Table 2.2), where $(X - \bar{X})$ represents the difference between the actual score and the mean of all the scores, which is 6. Table 2.3 shows more standard deviation calculations.

TABLE 2.2 ■ Calculating Standard Deviation

X	$\bar{X}$	$(X - \bar{X})$
8	6	8 – 6 = +2
8	6	8 – 6 = +2
8	6	8 – 6 = +2
7	6	7 – 6 = +1
6	6	6 – 6 = 0
6	6	6 – 6 = 0
5	6	5 – 6 = –1
5	6	5 – 6 = –1
4	6	4 – 6 = –2
3	6	3 – 6 = –3

TABLE 2.3 ■ More Standard Deviation Calculations

X	$(X - \bar{X})$	$(X - \bar{X})^2$
8	+2	4
8	+2	4
8	+2	4
7	+1	1
6	0	0
6	0	0
5	–1	1
5	–1	1
4	–2	4
3	–3	9
Sum	0	28

What we now know from these results is that each score in this distribution differs from the mean by an average of 1.76 points. How does one interpret that? "Most scores in this distribution are within a couple of points of the mean."

Looking at the complexity of the equation for the standard deviation, it's useful to remember three reasons why it is so convoluted:

1. We square the distances to get rid of negative values.
2. We divide by $n - 1$ instead of n because that gives us a better estimate of the population's standard deviation.
3. We take the square root of the whole thing to bring the values back to their original size.

Real-World Stats

If you were like a mega stats person, then you might be interested in the properties of measures of variability for the sake of those properties. That's what mainline statisticians spend lots of their time doing—looking at the characteristics and performance and assumptions (and the violation thereof) of certain statistics.

But we're more interested in how these tools are used, so let's take a look at a study that actually focused on variability as an outcome. And, as you read earlier, variability among scores is interesting for sure, but when it comes to understanding the reasons for variability among substantive performances and people, then the topic becomes really interesting.

This is exactly what Nicolas Stapelberg and his colleagues in Australia did when they looked at variability in heart rate as it related to coronary heart disease. Now, they did not look at this phenomenon directly, but they entered the search terms *heart rate variability, depression*, and *heart disease* into multiple databases and found that decreased heart rate variability is found in conjunction with both major depressive disorders and coronary heart disease.

Why might this be the case? The researchers think that both diseases disrupt control feedback loops that help the heart function efficiently. This is a terrific example of how looking at variability can be the focal point of a study rather than an accompanying descriptive statistic.

Want to know more? Read . . .

Stapelberg, N. J., Hamilton-Craig, I., Neumann, D. L., Shum, D. H., & McConnell, H. (2012). Mind and heart: Heart rate variability in major depressive disorder and coronary heart disease—A review and recommendations. *Australian and New Zealand Journal of Psychiatry, 46*, 946–957.

USING SPSS TO COMPUTE DESCRIPTIVE STATISTICS

If you haven't already, now would be a good time to check out Appendix A so you can become familiar with the basics of using SPSS. Then come back here.

Let's use SPSS to compute some descriptive statistics. The data set we are using is named Chapter 2 Data Set 1, and it is a set of 20 scores on a test of prejudice. All of the data sets are available in Appendix C and from the SAGE website **edge.sagepub.com/salkindfrey8e**. For at least the first couple of times that you use SPSS to analyze data in this course, we recommend you enter data by hand. That is a good skill to have for the real-life data collecting you'll do one day. After you have a little experience with that, feel free to download or copy and paste the data directly. There is one variable in this data set (Table 2.4).

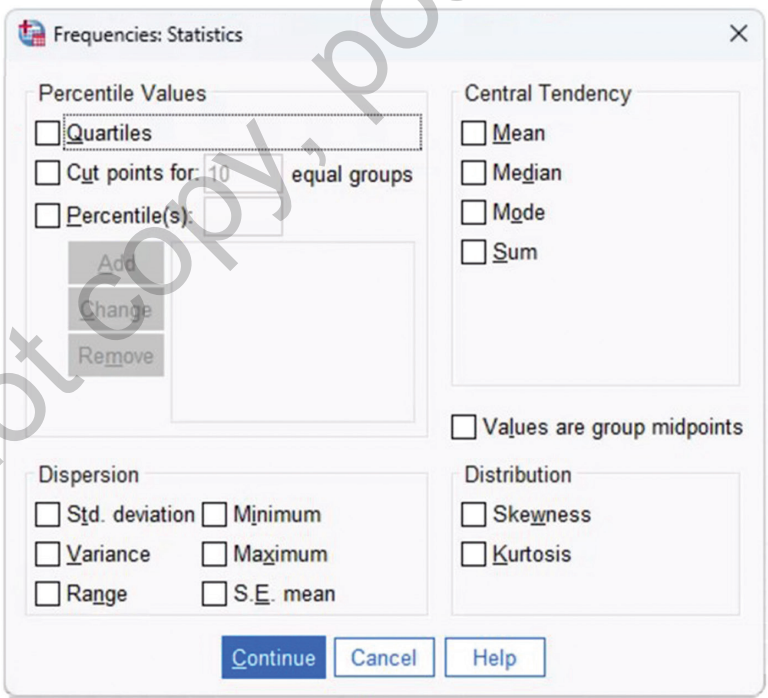
TABLE 2.4 ■ Operationalizing Prejudice

Variable	Definition
Prejudice	The value on a test of prejudice as measured on a scale from 1 to 100. Higher scores mean greater prejudice.

Here are the steps to compute several descriptive statistics that summarize these data. Follow along and do it yourself. With this and all exercises, including data that you enter or download, we'll assume that the data set is already open in SPSS.

1. Click Analyze → Descriptive Statis → Frequencies . . .
2. Double-click on the variable named Prejudice to move it to the Variable(s) box.
3. Click Statistics, and you will see the Frequencies: Statistics dialog box shown in Figure 2.1.
4. Under Central Tendency, check the Mean, Median, and Mode boxes.
5. Under Dispersion, check the Std. deviation, Variance, and Range boxes.
6. Click Continue.
7. Click OK.

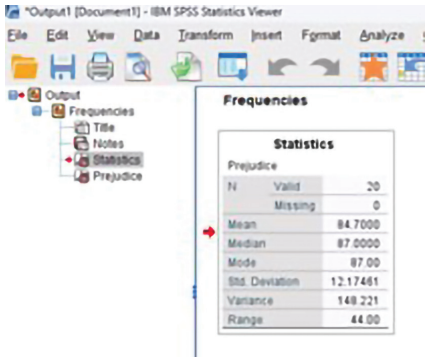
FIGURE 2.1 ■ The frequencies: Statistics dialog box from SPSS



The SPSS Output

Figure 2.2 shows you selected output from the SPSS procedure for the variable named Prejudice.

FIGURE 2.2 ■ Descriptive statistics from SPSS



In the Statistics part of the output, you can see that the mean, the median, and the mode are all computed along with the sample size (n) and the fact that there were no missing data. You also asked for the standard deviation, variance, and range. SPSS does not use symbols such as X in its output. Below this table, there is another large one that shows frequencies of each value and the percentage of times they occurred. (Strangely, we have to go through this Frequencies option to get all three averages, instead of the Descriptives option!)

Understanding the SPSS Output

This SPSS output is pretty straightforward and easy to interpret.

The mean score for the 20 scores is 84.70 (and remember that the span of possible scores is from 0 to 100). The median, or the point at which 50% of the scores fall above and 50% fall below, is 87 (which is pretty close to the mean), and the most frequently occurring score, or the mode, is 87.

Of the three measures of variability (or *dispersion*, as SPSS says), the one you would likely report is the standard deviation of 12.17, which means that most scores are within about 12 points of the mean (84.7).

> SPSS output can be full of information or give you just the basics. It all depends on the type of analysis that you are conducting. In the preceding example, we have just the basics and, frankly, just what we need. Throughout this book, 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 focus on output that is directly related to what you learned in the chapter.

No matter how fancy-schmancy your statistical techniques are, you will almost always start by simply describing what's there—hence the importance of examining the basic descriptive statistics of averages and measures of variability. To really understand your data, though, you might want to be able to draw a picture of what they look like. That's what we do next.

SHAPING THINGS UP

So far, we've learned about the two most important types of descriptive statistics—averages (measures of central tendency) and summaries of variability. Both of these provide us with the best numbers for describing a group of data (the average) and reflecting how diverse, or different, scores are from one another (variability).

What we did not do, and what we will do here, is examine how differences in these two measures result in different "shapes" or distributions of groups of scores. Numbers alone (such as $\mu = 3$ and $s = 3$) may be important, but a visual representation is a much more effective way of examining the characteristics of a distribution as well as the characteristics of any set of data.

So, now we'll learn how to visually represent a distribution of scores as well as how to use different types of graphs to represent different types of data.

> Sometimes it is a good idea to graph your data first and then do whatever calculations or analysis is called for. By first looking at the data, you may gain insight into the relationship between variables, what kind of descriptive statistic is the right one to use to describe the data, and so on. This extra step might increase your insights and the value of what you are doing.

Ten Hints for Making a Picture Worth a Thousand Words

Whether you create illustrations by hand or use a computer program, the same principles of decent design apply. Here's a good rule book to use:

1. **Minimize the junk.** "Chart junk" happens when you use every function, every graph, and every feature a computer program has to make your charts busy, full, and uninformative. With graphs, more is definitely less.
2. **Plan out your chart before you start creating the final copy.** Use graph paper even if you will be using a computer program to generate the graph. Actually, why not just use your computer to generate and print out graph paper (try www.printfreegraphpaper.com).
3. **Say what you mean and mean what you say—no more and no less.** There's nothing worse than a cluttered (with too much text and fancy features) graph to confuse the reader.
4. **Label everything so nothing is left to confuse your audience.**
5. **A graph should communicate only one idea—a comparison of values or a demonstration of a relationship.**
6. **Keep things balanced.** When you construct a graph, center titles and labels.
7. **Maintain the scale in a graph.** "Scale" refers to the proportional relationship between the horizontal and vertical axes. This ratio should be about 3 to 4, so a graph that is 3 inches wide will be about 4 inches tall. (Do the figures in this book follow this rule?)
8. **Simple is best and less is more.** Keep the chart simple but not simplistic. Convey one idea as straightforwardly as possible, with distracting information saved for the accompanying text.
9. **Limit the number of words you use.** Too many words or words that are too large, in terms of both physical size and ideas, can detract from the visual message your chart should convey. (This is a good rule for PowerPoint slides, too!)

10. **A chart alone should convey what you want to say.** Remember, a chart or graph should be able to stand alone, and the reader should be able to understand the message. If it doesn't, go back to your plan and try it again.

The most basic way to illustrate data is by showing a **frequency distribution**, a method of tallying and representing how often certain scores occur in a set of scores. In the creation of a frequency distribution, scores are usually grouped into class *intervals*, or ranges of numbers.

And a common chart that shows those intervals is a histogram.

PEOPLE WHO LOVED STATISTICS

Helen M. Walker (1891–1983) began her college career studying philosophy and then became a high school math teacher. She got her master's degree, taught mathematics at the University of Kansas (your authors' favorite college) where she was tenured, and then studied the history of statistics (at least up to 1929, when she wrote her doctoral dissertation at Columbia). Dr. Walker's greatest interest was in the teaching of statistics, and many years after her death, a scholarship was endowed in her name at Columbia for students who want to teach statistics! Her publications included a whole book teaching about the best way to show statistics using tables. Oh, and along the way, she became the first woman president of the American Statistical Association. All this achievement from someone who actually loved teaching statistics. Just like your professor!

Histograms

Histograms are graphs of frequency distributions where the frequencies are represented by bars. And the width of the bars represent the intervals of the scores when put in order.

By the way, with these sorts of charts that show two variables, one horizontally along the bottom and one vertically along the side, the horizontal line is called the *x*-axis and the vertical line is the *y*-axis. (The memory trick we learned in statistics kindergarten was that a *Y* looks like you're reaching your arms to the sky vertically!)

Depending on the book or journal article or report you read and the software you use, visual representations of data are called graphs (such as in *SPSS*) or charts (such as in the Microsoft spreadsheet *Excel*). It really makes no difference. All you need to know is that a graph or a chart is the visual representation of data.

To create a histogram, do the following:

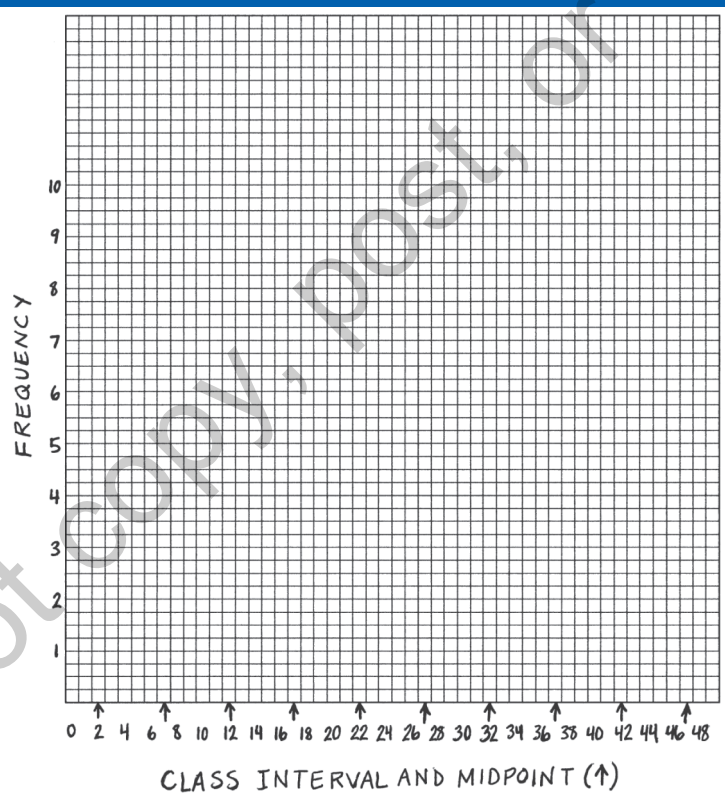
1. Using a piece of graph paper, place values at equal distances along the *x*-axis, as shown in Figure 2.3. These are called class intervals, and they basically turn nice interval data into ordered chunks. Now, identify the midpoint of each class interval, which is the middle point in

the interval. It's pretty easy to just eyeball, but you can also just add the top and bottom values of the class interval and divide by 2. For example, the midpoint of the class interval 0–4 is the average of 0 and 4, or 4/2 = 2.

2. Draw a bar or column centered on each midpoint that represents the entire class interval to the height representing the frequency of that class interval. Bars go across horizontally and columns go up and down vertically. For example, in Figure 2.4, you can see that in our first entry, the class interval of 0–4 is represented by the frequency of 1 (representing the one time a value between 0 and 4 occurs). Continue drawing bars or columns until each of the frequencies for each of the class intervals is represented. Figure 2.4 is a nice hand-drawn (really!) histogram for the frequency distribution of a set of 50 scores.

Notice that each class interval is represented by a range of scores along the *x*-axis.

FIGURE 2.3 ■ Class intervals along the *x*-axis



Polygons

Creating a histogram wasn't so difficult, and the next step (and the next way of illustrating data) is even easier. We're going to use the same data—and, in fact, the histogram that you just saw created—to create a frequency polygon. (*Polygon* is a word for shape.) A **frequency polygon** is a continuous line that represents the frequencies of scores within a class interval, as shown in Figure 2.5.

FIGURE 2.4 ■ A hand-drawn histogram

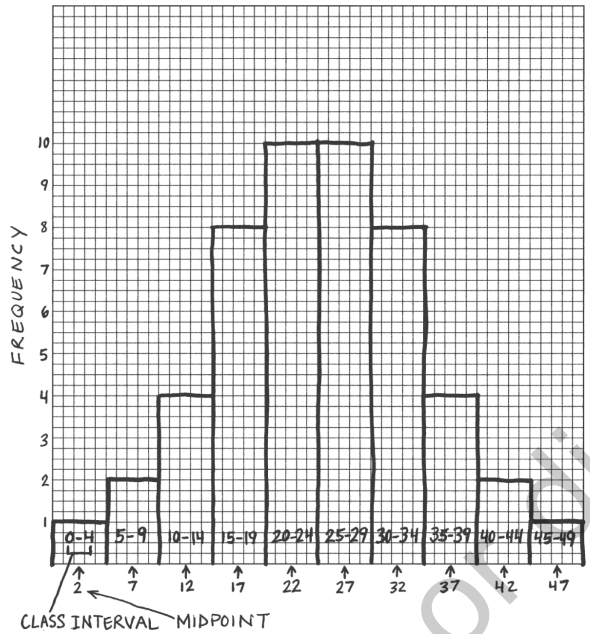
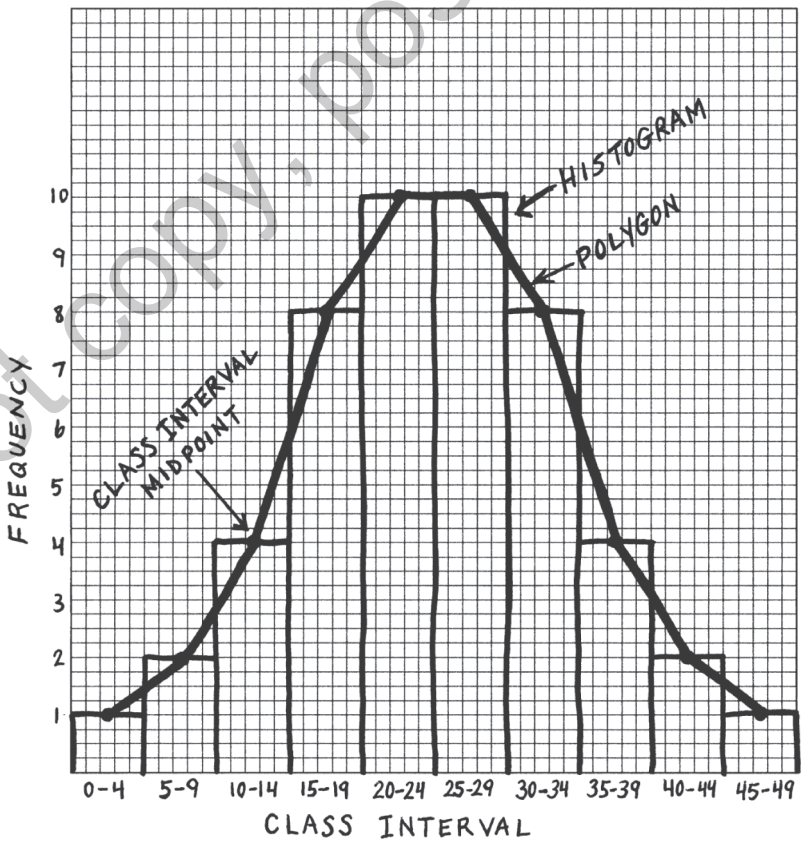


FIGURE 2.5 ■ A hand-drawn frequency polygon



How did we draw this? Here's how:

1. Place a midpoint at the top of each bar or column in a histogram (see Figure 2.4).
2. Connect the lines and you've got it—a frequency polygon!

Note that in Figure 2.5, the histogram on which the frequency polygon is based is drawn using vertical and horizontal lines, and the polygon is drawn using curved lines. In real life, though, you usually aren't shown the underlying histogram.

Why use a frequency polygon rather than a histogram to represent data? The use of a continuous line suggests that the variable represented by the scores along the *x*-axis is a theoretically continuous, interval-level measurement as we talked about at the start of this chapter. (To purists, the fact that the bars touch each other in a histogram suggests the interval-level nature of the variable, as well.)

Bar Charts

A bar chart should be used when you want to compare the frequencies of different categories with one another. You'll recall that when scores represent categories, they are at the nominal level of measurement. With most **bar charts**, categories are organized vertically on the *y*-axis, and values are shown horizontally on the *x*-axis. (You can also show the bars going up and down, if you'd like, with the categories along the *x*-axis and the scores along the side.) Here are some examples of when you might want to use a bar chart:

- Number of participants in different water exercise activities at your local pool
- The sales of three different types of products
- Number of children in each of six different grades

Figure 2.6 shows a graph of number of participants in different water activities.

Line Charts

A **line chart** should be used when you want to show change. This sort of graph is often used when the *x*-axis represents time. Here are some examples of when you might want to use a line chart:

- Number of cases of mononucleosis (mono) per season among college students at three state universities
- Toy sales for the T&K company over four quarters
- Number of travelers on two different airlines for each quarter

In Figure 2.7, you can see a chart of sales in units over four quarters.

FIGURE 2.6 ■ A bar chart that compares different water activities

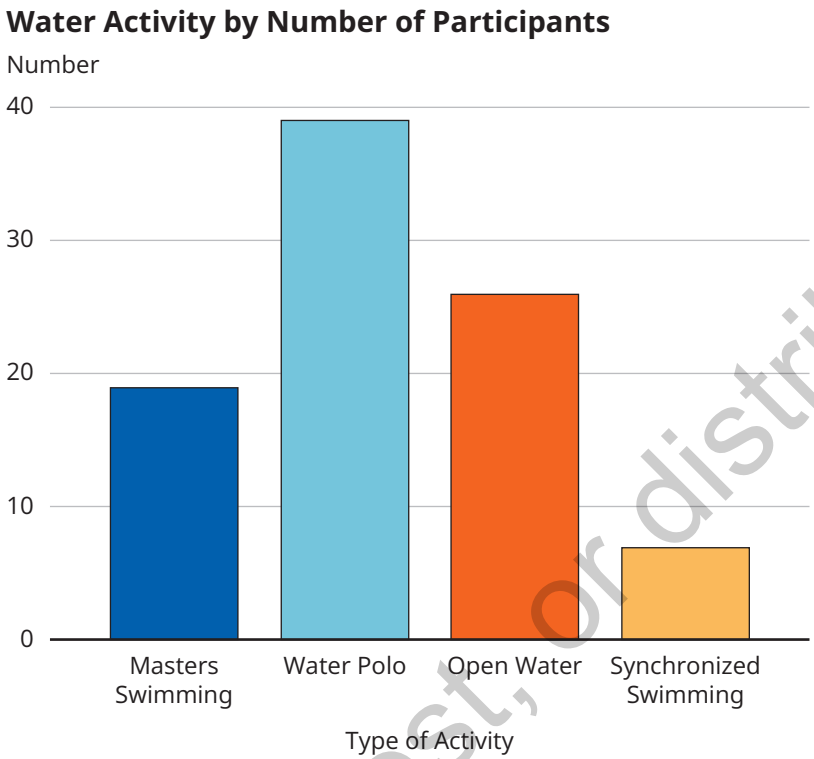
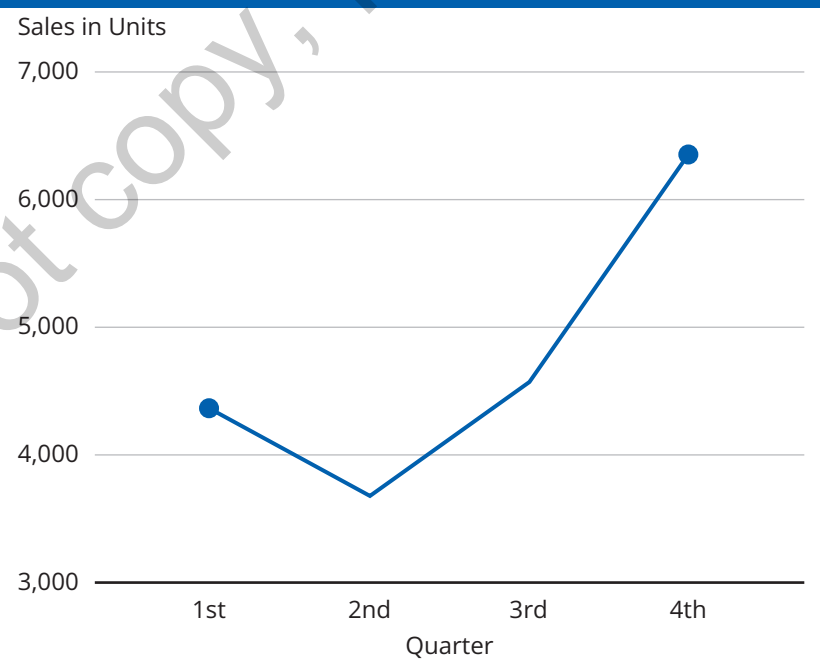


FIGURE 2.7 ■ A line chart that shows change over time



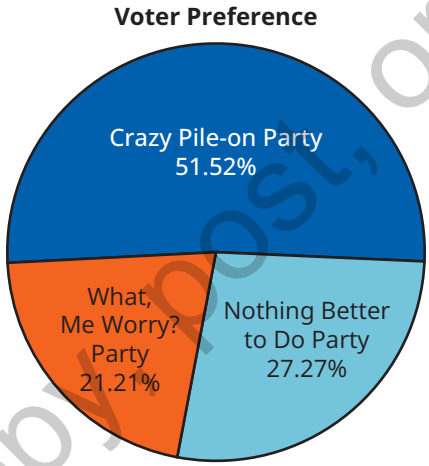
Pie Charts

A **pie chart** should be used when you want to show the proportion or percentage of people or things that are in each category of a nominal-level variable. The rule is that the percentages in each "slice" must add up to 100%, to make a whole "pie." Here are some examples of when you might want to use a pie chart:

- Of children living in poverty, the percentage who represent various ethnicities
- Of students enrolled, the proportion who are in night or day classes
- Of participants, the percentage from various cities in your state

In Figure 2.8, you can see a pie chart of voter preference. And we did a few advanced things, such as separating and labeling the slices.

FIGURE 2.8 ■ A pie chart illustrating the relative proportion of one category compared to others



USING THE COMPUTER (SPSS, THAT IS) TO ILLUSTRATE DATA

Now let's use SPSS and go through the steps in creating some of the charts that we explored in this chapter. First, here are some general SPSS charting guidelines.

1. Although there are a couple options, we will use the Chart Builder option on the Graphs menu. This is the easiest way to get started and well worth learning how to use.
2. In general, you click Graphs → Chart Builder, and you see a dialog box from which you will select the type of graph you want to create.
3. Click the type of graph you want to create and then select the specific design of that type of graph.

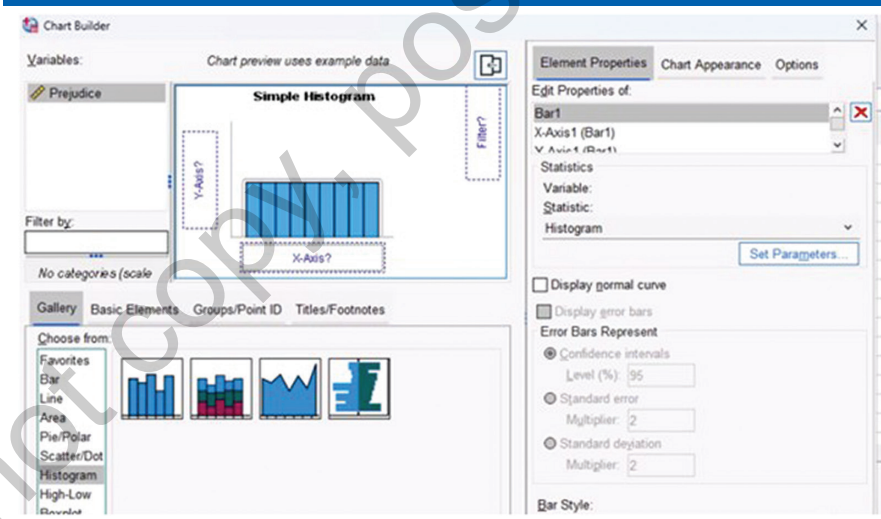
4. Drag the variable names to the axis where each belongs.
5. Click OK, and you'll see your graph.

Let's practice. We will focus on just two types of graphs—histograms and line charts. We bet you can figure out how to do the other ones on your own.

Creating a Histogram

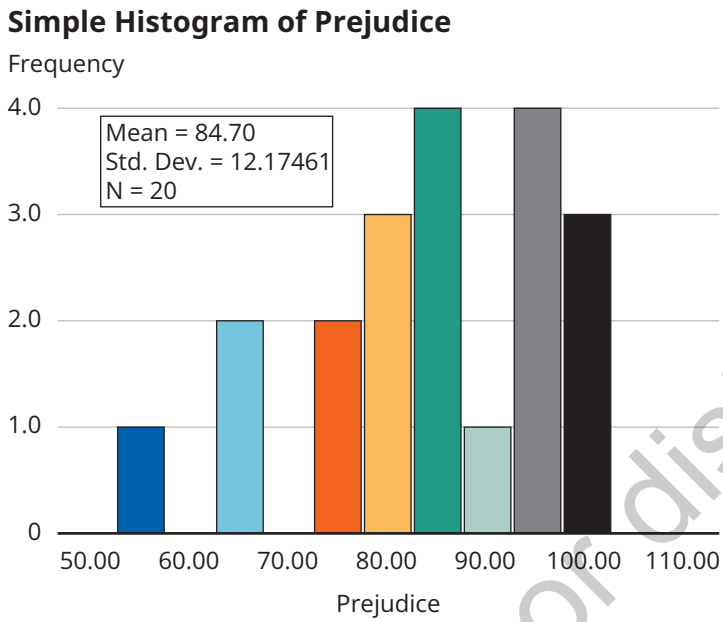
1. Enter the data you want to use to create the graph. Chapter 2 Data Set 1 in Appendix C provides 20 scores measuring prejudice. Use these!
2. Click Graphs → Chart Builder and you will see the Chart Builder dialog box, as shown in Figure 2.9.
3. Click the Histogram option in the Choose from: list, and double-click the first image or drag it to *Chart preview* box.
4. Drag the variable you wish to graph to the "*x*-axis?" location in the preview window.
5. Click OK and you will see a histogram, as shown in Figure 2.10.

FIGURE 2.9 ■ The Chart Builder dialog box



The histogram in Figure 2.10 looks a bit different from the hand-drawn one representing the same type of data shown earlier in this chapter, in Figure 2.4 The difference is that SPSS defines class intervals using its own kind of weird method. SPSS took as the middle of a class interval the bottom number of the interval (such as 10) rather than the midpoint (such as 12.5). Consequently, scores are allocated to different groups. The lesson here? How you group data makes a big difference in the way they look in a histogram. And, once you get to know SPSS well, you can make all kinds of fine-tuned adjustments to make graphs appear exactly as you want them.

FIGURE 2.10 ■ A histogram created using the Chart Builder



Creating a Line Graph

To create a line graph, follow these steps:

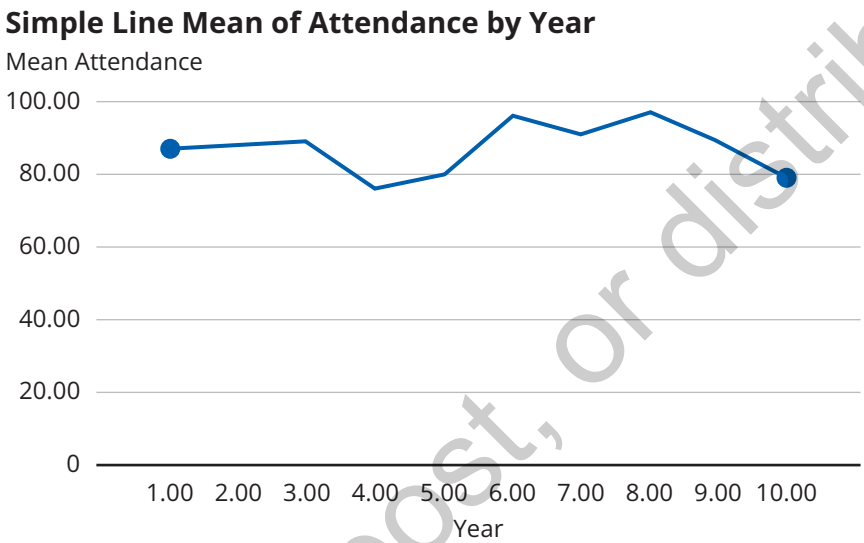
1. Enter the data you want to use to create the graph. In this example (Table 2.5), we will be using the percentage of the total student body who attended the first day of classes each year over the duration of a 10-year program. Here are the data. You can type them into SPSS exactly as shown here, with the top row being the names you will give the two variables (columns).

TABLE 2.5 ■ Program Attendance by Year

Year	Attendance
1	87
2	88
3	89
4	76
5	80
6	96
7	91
8	97
9	89
10	79

2. Click Graphs → Chart Builder and you will see the Chart Builder dialog box.
3. Click the Line option in the *Choose from:* list, and double-click the first image.
4. Drag the variable named Year to the *x*-axis? location in the preview window.
5. Drag the variable named Attendance to the *y*-axis? location.
6. Click OK, and you will see the line graph, as shown in Figure 2.11.

FIGURE 2.11 ■ A line graph created using the Chart Builder



SUMMARY

In this chapter we used descriptive statistics and charts to visualize a set of data. These are invaluable tools for understanding your "results." One piece of advice: SPSS will gladly show you all three types of averages (mean, median and mode) and all three ways to show variability (range, variance, and standard deviation). And it will gladly make all sorts of charts for all your variables! But don't report all of that stuff. Use the right statistic and the right graph for the level of measurement of your variables. Here's that table we showed earlier in this chapter that aligned averages with levels of measurement. We've filled in the rest of it for measures of variability and types of charts (Table 2.6). Do you agree with our choices?

TABLE 2.6 ■ Summarizing Data Based on Level of Measurement

Level of Measurement	Average	Measure of Variability	Graph
Nominal	Mode		Pie Chart Bar
Ordinal	Mode Median	Range	Histogram Polygram
Interval	Mode Median Mean	Range Variance Standard Deviation	Bar Line
Ratio			

There is a measure of variability for nominal level of data, but we don't cover it in this book. In fact, we had to do a little research to find it because it's rarely used. It is called the *variation ratio*. The variation ratio is the proportion of scores that are not in the most popular category. In other words, it is the proportion of scores that are not the mode.

KEY TERMS

Bar chart
Class interval
Difference score
Frequency distribution
Frequency polygon
Histogram
Interval level of measurement
Levels of measurement
Line chart
Measurement
Midpoint
Nominal level of measurement
Ordinal level of measurement
Pie chart
Range
Ratio level of measurement
Standard deviation
Variability
Variance

ACTIVITIES

1. For this distribution of seven scores, calculate the three averages and three measures of variability (Table 2.7). Use SPSS to calculate the variance and standard deviation, but figure out the rest by hand.

TABLE 2.7 ■ Calculating Averages and Variability

Score	Mean	Median	Mode	Range	Variance	Standard Deviation
12						
92						
42						
8						
40						
38						
40						

2. Table 2.8 shows a frequency distribution. Create a histogram by hand or by using SPSS. On SPSS, can you figure out how to create the class intervals you want?

TABLE 2.8 ■ Frequency Distribution

Class Interval	Frequency
261–280	140
241–260	320
221–240	3,380
201–220	600
181–200	500
161–180	410
141–160	315
121–140	300
100–120	200

3. For each of the following, indicate whether you would use a pie, line, or bar chart and why.

a. The proportion of undergraduates who are freshmen, sophomores, juniors, and seniors at your college

b. Change in temperature over a 24-hour period

c. Number of applicants for four different jobs

d. Percentage of test takers who passed

REVIEW QUESTIONS

1. Which distribution has the smallest standard deviation?

A	B.	C.	D.
1, 5, 9	2, 2, 5, 8, 8	1, 1, 5, 9, 9	4, 4, 4, 5, 5, 6, 6, 6

A. A
B. B
C. C
D. D

2. If you had a variable that was the number of cavities that the average dental patient has, what level of measurement would you be at?

A. Nominal
B. Ordinal
C. Interval
D. Ratio

3. Which type of graph doesn't have an *x*- or *y*-axis?
   - **A.** Pie chart
   - **B.** Bar chart
   - **C.** Histogram
   - **D.** Polygon

4. Why is the standard deviation usually smaller than the variance?
   - **A.** Sample size affects the variance.
   - **B.** Sample size affects the standard deviation.
   - **C.** The standard deviation is the square root of the variance.
   - **D.** The variance is the square root of the standard deviation.

5. Imagine a pie chart with three sections or slices. There are percentages in each section. What is the total of the three sections?
   - **A.** The variance
   - **B.** The standard deviation
   - **C.** 0
   - **D.** 100

6. When using **SPSS to compute descriptive statistics,** which menu path should you follow?
   - **A.** Analyze → Compare Means → One-Sample *t*-Test
   - **B.** Analyze → Descriptive Statistics → Frequencies
   - **C.** Data → Compute Variable → Descriptive Statistics
   - **D.** Graphs → Legacy Dialogs → Descriptive Charts

7. Which of the following correctly compares histograms and bar charts?
   - **A.** Histograms display continuous data with touching bars, while bar charts display categorical data with separated bars.
   - **B.** Bar charts always display numerical data, while histograms display only ordinal data.
   - **C.** Histograms are used only in inferential statistics, while bar charts are used in descriptive statistics.
   - **D.** Bar charts always have equal-width bars, while histogram bars vary in width based on frequency.

8. In SPSS, which type of graph is best used to show changes in a variable over time?
   - **A.** Pie chart
   - **B.** Line chart
   - **C.** Histogram
   - **D.** Bar

9. Which of the following is true about pie charts when compared to other graph types?
   - **A.** They are best used for showing relationships between multiple continuous variables.
   - **B.** They are ideal for comparing proportions of a whole but can be misleading with too many categories.

C. They provide a more accurate representation of data than histograms and bar charts.

D. They are primarily used to show frequency distributions of continuous variables.

10. Which graph type is most effective for showing the relative number of people in each of several categories?

A. **Pie** chart

B. Bar graph

C. Line graph

D. Scatter **plot**

CRITICAL THINKING QUESTIONS

1. Which characteristic of a distribution contributes the most to the size of a standard deviation?

A. The mean

B. How similar scores are to each other

C. How many scores there are in a distribution

D. The difference between the sample estimate of variance and the population's actual variance

2. The word *histogram* basically means to draw a picture of the masts (*histos* or poles) on a ship. Why do you think that word was chosen to describe this type of graph?

A. Because the vertical bars look like masts on a ship

B. Because the up and down pattern of the chart looks like waves in the ocean

C. Because the slices look like parts of a pie

D. Because early statisticians were interested in navigation