Preface

What if we had the tools to make sense of the world around us—to understand the claims we find on TikTok, the clickbait headlines we read, and the research that seems to contradict itself every other week?

What if learning these tools could be unintimidating, realistic, and, dare I say, fun?

Because it can be without sacrificing depth. And I believe that statistical literacy doesn't just help us understand the world, but it's also essential to make it a better place. I believe that if more people understood, rather than feared, statistics that we could reduce—or eliminate—almost all of the world's biggest problems.

Hunger. Homelessness. Suicide. Illiteracy. Poverty. Terrorism.

You name it…

Because the tools are there…

Welcome to statistics… are you ready to save the world?

I am.

The Challenge

Numbers scare people. I get it. Maybe you're not a "numbers person" or you've "never been good at math." That's okay. Statistics isn't just math. It's meaning.

Because of this, statistics are more like a foreign-language class than a math one. The math in this book? Mostly stuff you probably learned a very long time ago. So yes, we may need to review a few basic mathematics skills, but mostly, we're applying those skills to investigate our research questions.

We start this book by doing statistics by hand. Why? Because this helps us to understand what the formulas mean. Sure, a computer can give us results, but if we can't interpret them, those results are meaningless.

The Difference

This book doesn't just teach how to compute statistics but it also teaches us to use statistics to think critically and actually apply them. This book:

- Uses innovative metaphors and analogies to help students understand and remember complex concepts such as cheesecake, dating apps, horror movies, and even Nicolas Cage—yes, really.
- Teaches students to think critically about statistics from the very beginning in Chapter 1.
- Instructs students on how to conduct statistical analyses by hand first and then follows that up with teaching them to use statistical software second, specifically IBM® SPSS® Statistics 31.

- Provides concrete and relatable examples, from cat people to TikTok.
- Reads in a voice that's human, not robotic—and definitely not like a mathematically inclined cyborg. You're welcome!
- Incorporates modern visuals that make statistics easier to learn.
- Embeds "Practice Makes Perfect" questions throughout each chapter so that students can conduct self-checks to gauge if they're reading and retaining the information.
- Includes vital content that's often glossed over in introductory statistics textbooks such as assumptions, outliers, and robustness.

The Content

Statistics With Humans builds upon itself in three connected parts. Each part builds the students' skills to go from foundational to more complex.

Part I: Welcome to Statistics Boot Camp

Teaches the basic information necessary to be successful in conducting inferential statistics.

- **Chapter 1. May I Make an Introduction? To Statistics That Is…**
  Provides a big-picture overview of statistics, research, and levels of measurement.
- **Chapter 2. Measures of Central Tendency: Attraction**
  Explores how and when to use the mean, median, and mode to describe the center of a data set.
- **Chapter 3. Measures of Variability: Tree Branches Scattered Across the Sky**
  Teaches measures of variability such as exclusive and inclusive ranges, mean absolute deviation, variance, and standard deviation to describe the spread of a data set.
- **Chapter 4. Visual Displays and Measures of Distribution: Let's Lay It All Out**
  Instructs how to create visual displays of our data's distribution so that we can get an idea of its shape.
- **Chapter 5. The Standard Normal Distribution: Finding Order in the Chaos**
  Explains how to standardize scores to create a standard normal distribution so that we can find the area under the curve.
- **Chapter 6. Probability and Sampling: It's More Than Sports and Gambling**
  Illustrates how to find probabilities of discrete and continuous data to build a strong foundation for what we do with inferential statistics.

Part II: Welcome to Inferential Statistics

Teaches the recipe for the hypothesis-testing steps so that we know the process for testing our research question with quantitative data.

- **Chapter 7. Hypothesis Testing and Statistical Significance: Putting It All Together**
  Combines everything we've learned so far to teach us the hypothesis-testing steps to try to find an answer to our research question.

- **Chapter 8. One-Sample *z*-Test: Are NBA Players Taller Than the Average Male?**
  Is the mean height of the 2020 Milwaukee Bucks players significantly greater than the mean of all males? Use the one-sample *z*-test to find out if the sample mean is significantly different from the population mean.
- **Chapter 9. Independent Samples *t*-Test: Who Owns More Shoes?**
  Do people who identify as women have significantly more shoes than people who identify as men? We use the independent samples *t*-test to determine if there are significant differences between two independent groups.
- **Chapter 10. Dependent Samples *t*-Test: Does Dry January Actually Work?**
  Do people drink significantly less alcohol the week after Dry January than the week before it? Investigate whether there are significant differences between two related groups using the dependent samples *t*-test.
- **Chapter 11. The ANOVA: Who Is the Biggest Diva?**
  Are there significant differences in the self-reported diva levels among frontpeople, lead guitarists, bassists, and drummers of rock bands? The ANOVA can help us determine whether there are significant differences among three or more groups.
- **Chapter 12. The Pearson Product–Moment Correlation: Are TikTok and Reading Times Related?**
  Is there a significant relationship between time spent reading books and time spent on TikTok? The Pearson product–moment correlation can help find out how strong the relationship is.
- **Chapter 13. Simple Linear Regression: Does Time Spent on Dating Apps Predict Dating Bitterness?**
  Does the amount of time spent on dating apps predict levels of dating bitterness? Linear regression makes prediction models to help us find out.

Part III: Welcome to Nonparametric Analyses

What do we do when the data aren't normal? This part of the book introduces nonparametric analyses and teaches students to conduct chi-square tests.

- **Chapter 14. Chi-Square for Goodness of Fit Test: Are People Terrified by AI More Than We'd Expect?**
  Are the proportions of people who are intrigued by, indifferent to, or terrified by artificial intelligence significantly different than expected? The chi-square test helps us test significant differences among proportions.

Pedagogy

- **Learning Objectives:** Each chapter begins with a list of learning objectives that are clear, measurable goals that outline what students should be able to do by the end of the chapter, providing structure and focus to enhance student learning.
- **Practice Makes Perfect:** End-of-section targeted practice questions embedded throughout each chapter to reinforce student comprehension of the concepts immediately after they're introduced.

- **Chapter Summary:** A summary that reviews the major concepts covered.
- **Symbol Guide:** A quick reference list of the symbols introduced in each chapter.
- **Formula Guide:** A list of the formulas introduced in each chapter, clearly labeled to reinforce students' understanding of the new formulas they've learned.
- **Terms to Know:** A list of the key terms introduced in the chapter.
- **Putting in the Work:** End-of-chapter review and applied questions that provide students with ample opportunity to practice the statistical calculations they learned.

Sage Vantage Features

This text is available in Sage Vantage—an intuitive learning platform you and your students will actually love.

9 out of 10 of your colleagues using Sage Vantage would recommend it!

Sage Vantage for *Statistics With Humans* integrates all the textbook content with assignable multimedia activities and autograded assessments to drive student engagement and ensure accountability. Unparalleled in its ease of use and built for dynamic teaching and learning, Vantage offers customizable LMS integration and best-in-class support.

Vantage Learning Platform ISBN: 9781071891087

See why you'll love Sage Vantage at **collegepublishing.sagepub.com/vantage**

Faculty and Students Approve

"Student performance after implementing Vantage is incredible….
We know that reading increases performance, but this is a really tangible way to see that."
—*Karen Deeming, Instructor, Merced College*

"With Vantage, **I set up my course in under 10 minutes…** I love Vantage!"
—*Karin Machluf, Instructor, Penn State Scranton*

"Interactive, engaging and simple to use."
—*Alyssa Salazar, Student, University of North Texas*

"It really **helps you wrap your head around what you're learning** and come to a complete understanding of the course material."
—*Aubrey Akins, Student, St. Bonaventure University*

"This is the **smoothest student onboarding process** I have ever experienced."
—*Echo Leaver, Instructor, Salisbury University*

Analytic Skill-Building

SPSS Walkthrough Videos available only in Vantage align with learning objectives, reinforcing fundamental concepts in every chapter.

- ***SPSS Walkthrough Videos*** help students set up and run each statistical test in SPSS and can be viewed at home to save instructional time in class.

Videos for *Statistics With Humans*

Chapter 1: Setting Up SPSS Data Files

Chapter 2: Finding Measures of Central Tendency

Chapter 3: Finding Measures of Variability

Chapter 4: Calculating Measures of Distribution

Chapter 5: Calculating *z*-Scores

Chapter 9: Running the Independent Samples *t*-Test

Chapter 10: Running the Dependent Samples *t*-Test

Chapter 11: Running the One-Way ANOVA

Chapter 12: Running the Pearson Product-Moment Correlation

Chapter 13: Running Simple Linear Regression

Chapter 14: Running the Chi-Square Goodness-of-Fit Test

Additional Teaching Resources

Visit **collegepublishing.sagepub.com** and navigate to the Resources tab on your book's page to find the teaching materials designed to accompany this textbook. On this site, you will find an array of materials that will save you time and help you keep students engaged, including:

- **Learning management system** cartridges that easily integrate with your course management system so that student test results and graded assignments seamlessly flow into your gradebook;
- **Test banks**, aligned to Bloom's Taxonomy, that provide a diverse range of test items, including multiple-choice, true/false, and essay questions;
- **Lecture notes** that provide an outline and the key concepts in each chapter to aid in lecture preparation;
- **PowerPoint® slides** that offer a flexible, accessible, and customizable solution for creating multimedia lectures;
- **Tables and figures from the book** are available to support lecture preparation and class discussions.
- **Tables** to help students keep work organized. Teaching students to organize their work becomes particularly important as they learn to master hand calculations;
- **Sample course syllabi** include suggested models for structuring your course; and
- **Worksheets** that students can use as they work through textbook examples to help them check their own understanding.

Join me.

Chapter 1

May I Make an Introduction? To Statistics, That Is...

iStock.com/Bim

What do your BFF and your rusty, old car have to do with statistics? Read on to find out...

Learning Objectives

1.1 Explain how data relates to statistics.

1.2 Distinguish between descriptive and inferential statistics.

1.3 Differentiate the different types of sampling methods through explaining their advantages and disadvantages.

1.4 Explain and apply measurement fundamentals to statistics.

1.5 Apply research methods fundamentals using the applicable terminology to research studies.

A Brief Introduction to Statistics

Welcome to statistics. You're excited, right? If we're real about it, you're probably not very excited to be in stats class. That's okay. When I was in your place, I wasn't excited either. I thought I would take the class once in college and never see that information again. The joke was on me, though.

When I got my very first teaching schedule after graduating from grad school, I read the list of classes I would teach that fall. Most of them made sense to me—general psychology, lifespan development—until I got to…statistics. I was pretty sure that they were hazing me, because who really wanted to teach statistics? Somehow, I had made it through four different statistics courses and still had no idea what I was doing. Now, I would be responsible for other people learning statistics. Great (insert sarcastic voice here)!

So how did I make it through four statistics courses and still have no idea what I was doing? My theory is that traditionally, really smart people wrote the statistics textbooks using really fancy words and examples that normal people couldn't relate to. These authors presented the information in a dry way without explaining why that information was important. Then, bless my statistics teachers' hearts, they didn't have access to textbooks that students could connect to. The ones they used often relied on statistical jargon without really explaining what that jargon meant.

Even though I was initially terrified to teach statistics, after teaching it for some time now, it has become my favorite course of all. It's my favorite because my students typically hold such low expectations for it, and then, by the end of the course, they understand what they're doing. After such a rewarding teaching experience, it became my mission to write a book that would help students everywhere understand and relate to statistics. So let's get started!

In this chapter, we get the grand introduction. We'll discuss data and how it relates to statistics. Then, we'll talk about the two major branches of statistics, descriptive and inferential. We will end the chapter with learning about the research design basics necessary for us to be successful in statistics. In the first six chapters of this textbook, we will go through what I've dubbed statistics boot camp. We will learn all the prerequisite information to understand and conduct the different types of inferential statistics that we'll learn about in the rest of the book.

Welcome to statistics boot camp!!

Statistics Is All About Data

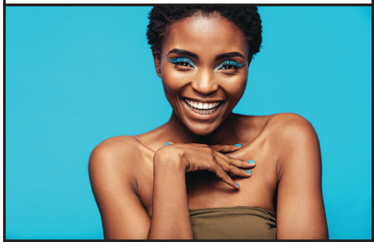
At the basic level, statistics is all about data. In short, **data** is information. Generally, we think about data as either qualitative or quantitative. Researchers use both qualitative and quantitative data to try to explain a vast variety of phenomena in the world such as mental illness, love, homelessness, and almost anything else you can think of. Although both types of data provide valuable information, researchers tend to align with either qualitative or quantitative research in their work. However, I contend that both quantitative and qualitative data provide equal value to research and to the broader society.

Qualitative data is information that is represented by words. For example, if we asked our friend to explain what it felt like to fall in love for the first time, we might hear about the feeling of electricity on a first date, grabbing hands and impulsively dashing into traffic with a new love without even looking. Or if you asked someone what it was like to lose a friend to suicide, they might lament about the intense guilt they felt that they couldn't prevent their friend's death. Both responses provide details about these experiences. Often, qualitative researchers want to make sense of human experiences. Although qualitative research can supply the juicy details along with a human perspective, critics of qualitative research argue that it can also be biased and subjective.

Quantitative data is information that is represented by numbers. In statistics, we are primarily concerned with quantitative data. For example, if we want to know how old someone is, we expect them to give us a number representing their age in years. Or if we wanted to know how much weight we gained over spring break (because of all those delicious street tacos), we would probably step onto a scale to find our weight in pounds or kilograms. In statistics, we use a collection of quantitative data, also known as a **data set**, for analysis using (1) descriptive statistics and (2) inferential statistics. Although quantitative research is often more objective than qualitative research, critics of quantitative research argue that it fails to provide the context for the research.

Find an example of a data set in Figure 1.1.

Figure 1.1 ▪ Example Data Set

Participant	How many cups of coffee did you have yesterday?
	0 cups
	6 cups
	2 cups
	1 cup

iStock.com/jacoblund; iStock.com/shapecharge; iStock.com/digitalskillet; iStock.com/ajijchan

Because of the limitations of qualitative and quantitative research, many researchers opt to combine both to address a research question. This is known as **mixed-methods research**.

Practice Makes Perfect

1. If someone asked how old people were when they had their first kiss, would that be qualitative or quantitative data?
2. What is data?
3. What type of research examines data represented by both words and numbers?
4. If someone asked what a person's first kiss felt like, would the response be qualitative or quantitative data?

Answers

1. quantitative data
2. information
3. mixed-methods research
4. qualitative data

Descriptive and Inferential Statistics

As we chatted about in the last section, in statistics, we are primarily concerned with quantitative rather than qualitative data. In the social and behavioral sciences, researchers are interested in learning more about different topics related to people, such as what motivates people to run a marathon or defining the differences in the parenting styles of people who struggle with depression versus those who do not. Researchers call the people who take part in research **participants**. For example, in my research lab, we study issues surrounding suicide. In one study, we asked college students tons of very personal questions about mental illness and suicidal thoughts and behaviors. The students who answered those very personal questions were the participants in our study. In statistics, the symbol X represents the data from our participants.

Let's take the pretend data set presented in Figure 1.1 and put it into the format we'll most likely use in statistics, a table (see Table 1.1). The left column will usually be labeled participant or ID, arbitrary numbers used to represent each of the participants instead of their names. For example, Participant 1 might be named Seyko, Participant 2, Anthony, Participant 3, Leilani, and Participant 4, Rushil. When doing research, we won't know our participants' names but will have a code to identify each one. In the right column, labeled X, we will have the data from each of our participants. In this example, we asked our participants how many cups of coffee they had yesterday. Table 1.1 lets us know that Participant 1 consumed 0 cups, Participant 2 downed 6 cups, Participant 3 drank 2 cups, and Participant 4 sipped on only 1 cup of coffee yesterday.

Table 1.1 ▪ Cups of Coffee Example Data Set

Participant	X
1	0
2	6
3	2
4	1

Throughout this textbook, I will introduce a number of important symbols. To make your life easier, please make it a priority to learn the symbols we discuss. For example, you might create flashcards of the symbols for review. As we move further into the course content, we will need to use the symbols on the spot. Statistics is more of a language than it is a mathematics course. While you aren't likely to learn any new math in this book, you will use the math you already know in a new way to investigate a variety of research questions. To succeed in statistics, treat it like a foreign language course and learn the symbols with fluency.

iStock.com/SolStock

Typically, we split statistics into two major branches, descriptive and inferential. **Descriptive statistics** is just what it sounds like—that is, when we use one or more numbers to *describe* our data set. You may already be familiar with descriptive statistics such as the mean and range. For example, if we went into a bar and wanted to know the mean number of drinks people consumed there, we would find out how many drinks each person had, add those numbers together, and divide that sum by the number of people in the bar. Some people might have 10 drinks, while others won't drink a single thing. Let's raise our glasses to the designated drivers who have been keeping the roads safe since 1885. The mean attempts to describe the center of the data set in a single number. We will learn about three major types of descriptive statistics including measures of central tendency, measures of variability, and measures of distribution.

Although we use descriptive statistics to describe our data set, we use inferential statistics to try to answer research questions. Often, researchers will identify a certain population that they want to study. A **population** consists of every single case with a certain characteristic. For example, if we wanted to study the Australian population, the population would consist of every single Australian. We use the symbol capitalized N to refer to the population size (i.e., how many participants there are in the population we're studying).

At this point, you might be thinking it would be really hard to get data from every single Australian, especially because people are being born and dying every single moment. This is where inferential statistics comes in. Because it can be really hard to study an entire population, researchers often study a smaller portion of a population (called a **sample**) to make guesses about what's happening in a broader population. We use the symbol lowercase n to represent the sample size (i.e., how many participants are in our sample).

Now that we have the vocabulary for population and sample, we can formally define inferential statistics. **Inferential statistics** is when we take a sample of participants to make conclusions about a larger population. When we apply our findings from a sample back to the overall population, we call that **generalization**.

Practice Makes Perfect

5. Imagine we wanted to find out if there was a significant relationship between the number of books children in Mississippi read before going to kindergarten and how many letters they could identify in 1 minute on the last day of school their kindergarten year. To do this, we would take a sample of children born in 2020 and have their parents track the number of books their children read until they

started kindergarten. On the last day of kindergarten, we would test each child on how many letters they could identify in 1 minute. What type of statistics would we be conducting if we applied what we learned from that sample to the entire population of children born in Mississippi in 2020?

6. What type of statistics would it be if we found the mean number of vinyl records people bought at the South by Southwest (SXSW) Music Festival last year?

Answers

5. inferential statistics
6. descriptive statistics

Sampling

As we learned in the last section, inferential statistics uses a sample to make guesses about what's going on in the overall population. So how do we get that sample? **Sampling** is the process of selecting a sample from a population. There are several types of sampling procedures including random sampling, convenience sampling, and snowball sampling.

Random sampling is a sampling method in which each member of the population has an equal chance of being selected as part of the sample. For example, if I wanted to study how frequently students at my school experienced suicidal thoughts, it's likely I would not be able to ask every single student at my school. Instead, I might contact the Office of Institutional Research (OIR) to take a random sample of students for me to use for my research study. To take a random sample, OIR would get a list of every single student at my school and then randomly select my sample from the list of all students. This would be like getting the biggest hat ever, writing each student's name on a tiny piece of paper, putting all those slips of paper into that enormous hat, and drawing a certain number of those slips of paper to create my sample.

Later in this book, we will discuss random sampling with relation to probability and assumptions of statistical tests. This doesn't mean much yet, but it will. What is important to remember is that many statistical analyses rely on the assumption that we randomly selected the sample we're analyzing in inferential statistics. Researchers regard random sampling as the gold standard of sampling methods because if we randomly select the participants, the sample is more likely to represent the population we want to study. We will discuss this more after we learn about the next method of sampling, convenience sampling.

Although random sampling is the gold standard in research, in the real world, convenience sampling occurs more frequently. **Convenience sampling** is exactly what it sounds like: a sampling method in which the researcher selects the sample based on easy access or because it is convenient. Many college professors use convenience samples of college students at their own universities because they can easily access participants for their studies. We typically find this in college psychological science programs, in which professors often give their students extra credit to participate in research studies.

Although this makes it easier for professors to recruit participants for their studies, it limits what generalizations can be made about the results. Whenever we use a convenience sample, there are typically factors that will connect those participants. For example, if we wanted to study the frequency of suicidal thoughts in the students at my school, but we only surveyed the students in my class, we would not be able to apply our results from our sample back to the entire university.

But why?

First off, the classes I teach often include people in helping professions such as psychology, nursing, health studies, pre–physical therapy, pre–occupational therapy, athletic training, and social work. The students may be drawn to these specific areas of study for underlying reasons that we can't necessarily detect from looking at them. Students in political science, music, English, or another major may not experience suicidal thoughts at the same rates as the students in my classes. This means that if we use a convenience sample (such as using just my class to answer this survey), we cannot apply those findings to the overall population (e.g., all students from my university) because of the possible commonalities of students drawn to helping professions. However, with a random sample, we can generalize the findings back to the larger population.

The last type of sampling method discussed in this book is snowball sampling. **Snowball sampling** is when participants in the research study recruit additional participants. This type of sampling method works really well for hard-to-reach populations (e.g., people who are addicted to opioids, people who lost a family member to gun violence). Many times, researchers will get permission from support groups to recruit participants in sensitive studies, and in turn, the members of those support groups can recruit participants who meet the characteristics within their support network. The snowball sampling method allows researchers to find participants that they might otherwise struggle to access.

Practice Makes Perfect

Match the sampling method with the most appropriate example.

7. Snowball sampling	**a.** Celeste wants to study the relationship between the number of hours students at her college work and the GPA of those same students. To find her sample, she went to her college's Office of Institutional Research and had their computer randomly select 100 students from the entire student body.
8. Random sampling	**b.** Laurynn wants to determine if there is a relationship between the number of tattoos people have and how many times those same people have been arrested. To find her sample, she posted a Facebook status update and asked her Facebook friends to comment on it.
9. Convenience sampling	**c.** Jay wants to study the perceptions of people who are unhoused on what the state political leaders could do to help people without secure housing. First, Jay recruited participants from a local shelter, who then helped Jay find additional qualifying participants for the study.

Answers

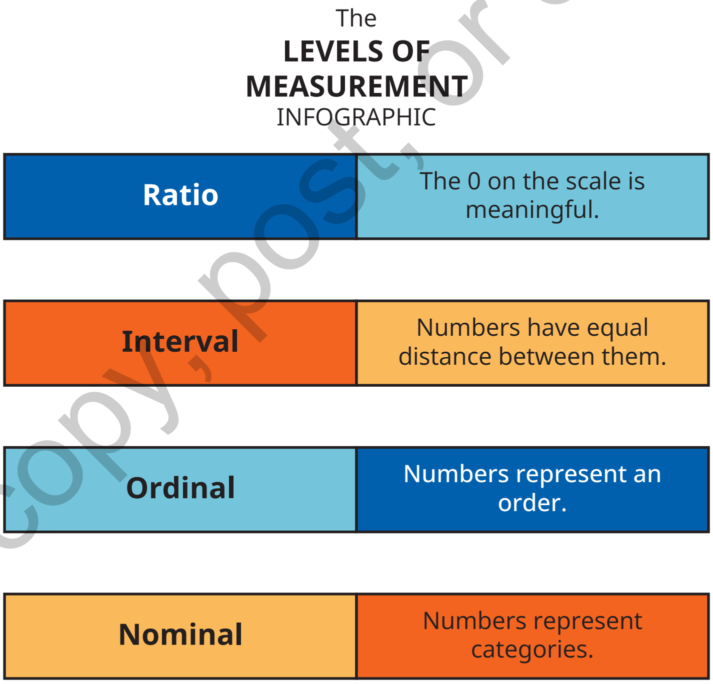
7. c
8. a
9. b

Measurement

If we remember from the beginning of the chapter, in statistics, we are primarily concerned with quantitative data. However, there may be times that we want to know more information about a phenomenon that is not represented as a number. In these cases, we use measures to get our data in numerical form. A **measure** is any tool we use to quantify information. If we use a survey, this measure might include questions that ask participants to numerically gauge how much of an attribute they possess. If we want to know how much we weigh, our measure is the scale we step onto to find out how many pounds or kilograms we are.

The process of turning information into numbers is known as **measurement**. If we want to measure how well someone can spell, how do we translate that into a number? In schools, teachers often do this with a spelling test that yields a score. There are several additional concepts related to measurement we need to understand before moving forward: levels of measurement, reliability, validity, and error.

Figure 1.2 ▪ The Levels of Measurement Infographic



Levels of Measurement

One of the most important things we will learn in this chapter is that not all measurement is created equal. In fact, there are several levels of measurement. **Levels of measurement** refer to the four main types of labels given to quantitative data—nominal, ordinal, interval, ratio—based on the amount of information it provides. Another way of saying this is that the level of measurement describes how much information the data we collected tells us.

In Figure 1.2, the Levels of Measurement Infographic shows the four different levels of measurement. Because the levels of measurement are hierarchical, the level of measurement that provides us the least amount of information falls at the bottom of the infographic, and the level of measurement that provides us the most amount of information resides at the top of it. The higher we get on the Levels of Measurement Infographic, the more information we know about the data. Keep this in mind as we go through each of the levels of measurement individually.

Sometimes, we will hear levels of measurement referred to as scales of measurement, but I avoid this terminology because of the specific meaning of the word "scale" within statistics and research design.

Levels of measurement are important, so really pay attention here. We will continue to use the levels of measurement throughout the rest of the book to think critically about things we will learn later.

Nominal is the most basic level of measurement. Here, we use numbers to represent categories. For example, say we wanted to separate people based on whether we thought they were awesome or not. To do this, we could assign a 1 to all the people who we thought were awesome and a 2 to all the people we thought weren't. Although that would provide us some information about the two basic categories of people, it would not quantify how awesome we thought each of these people were. A common example of nominal data includes the survey question "What is your biological sex," where 1 = female, 2 = intersex, 3 = male, and 4 = prefer not to respond. Other prime examples of nominal data include political affiliation, race/ethnicity, and religious affiliation.

Ordinal is the next level on the Levels of Measurement Infographic. In the ordinal level of measurement, we assign a number to represent some sort of order. For example, if we recruited three people to race each other in high heels down the hall, we would assign 1 to the person who crossed the finish line first, a 2 to the person who crossed the finish line second, and a 3 to the person who crossed the finish line third. Ordinal data gives us information about the order of the data but does *not* tell us how big the distances between the data points are.

Photo courtesy of Katie Jacobs

So meet Adam (on the left), Mike (in the middle), and Shannon (on the right). These terribly brave people volunteered to race for me down the hallway in heels just so that I could teach about ordinal data. I know what you're thinking: Did they really do this? Yes! And it. Was. Glorious!

As they lined up, several students and people who work for my university peeked their heads out to watch them race. No doubt, they were thinking, *What is Jacobs up to now?*

Adam looked unsure in his 5-inch wedges. Shannon started looking a little fierce, and because of her regular stiletto usage, she was considered the favorite. Mike smiled, wondering what he had gotten himself into. I stood at the finish line.

"Runners, on your mark," I called out as I raised my hands in the air. "Get set. Goooooooo!" I yelled down the hall as my hands dropped to my sides to start the race.

Photo courtesy of Katie Jacobs

Adam got off to the best start, pumping his arms back and forth. Mike propelled himself forward but got his hand caught in his wheels and slowed down temporarily. Shannon saw Adam racing ahead, and we could see the determination on her face as she tried to run faster.

Adam and Shannon looked like they were neck and neck, but within moments, Adam crossed the finish line first, Shannon sped across the line in a close second, and Mike came in last. If we were coding the data using an ordinal level of measurement, Adam would have been assigned the number 1, Shannon the number 2, and Mike the number 3. That is all the information that we would possess. This is because with ordinal data, we know the order, but we have no idea about the amount of time or distance between the racers. For example, it might have taken Adam 30 seconds to get to the finish line, Shannon 31 seconds, and Mike 35 seconds. With ordinal data, we would never know.

Interval is the next level of measurement and is the most commonly misunderstood. In the interval level of measurement, there is an equal distance between the data points, but the 0 is an arbitrary point on the scale. In other words, zero does not actually mean zero or a lack of the quality being measured. The most common example of an interval level of measurement is temperature in degrees Fahrenheit or Celsius. If we take degrees Celsius, for instance, there is the same distance between 1°C and 2°C as there is between 2°C and 3°C.

Although there is an equal distance between degrees on the Celsius scale, 0° does *not* represent the complete lack of heat. In fact, 0°C represents the freezing point of water. In this case, the 0 is just a placeholder for where water freezes. We also know that 0°C does *not* mean an absolute lack of heat, because we can have negative degrees Celsius, which still indicates some level of heat. Typically, if we have a negative number, it is generally on an interval scale, although there are a few exceptions. Another way statisticians say this is, "The 0 is not meaningful."

Table 1.2 ▪ Newborn Baby Shoe Sizes

U.S. Shoe Size	Length (cm)
0	7.5
1	8.5
2	9.5
3	10.5
4	11.5
5	12.5
6	13.5
7	14.5
8	15.5

Another example of interval data is shoe size. In Table 1.2, Newborn Baby Shoe Sizes, we find that a size 0 does *not mean* that there is *not* a shoe. The 0 here is not meaningful; it is just a placeholder. In this shoe size chart, the size 0 represents a shoe length of 7.5 centimeters. We also notice in this example that there are is a 1-cm difference between each of the shoes sizes, an equally distanced interval for each size (e.g., 1 cm between sizes 0 and 1, 1 cm between sizes 1 and 2, 1 cm between sizes 2 and 3).

Data for our final level of measurement, **ratio**, provides us with the most amount of information. Much like interval data, the distances between data points on a ratio level of measurement are equally spaced. However, on a ratio scale, 0 means the total lack of the attribute or quality we are measuring. In other words, this means that 0 means 0.

Another way that statisticians say this is that the 0 is meaningful for data on a ratio level of measurement. For example, if we had 0 street tacos to eat today, that means we've had no tacos, and it's a sad day! Common examples of ratio data include height and weight. If we are 0 inches tall, that means we have no height. If we weigh 0 pounds, that means that we weigh nothing. The zero means the complete lack of the thing that's being measured.

Although sometimes, it is difficult to tell the difference between ratio and interval levels of measurement, with respect to statistics, it is usually more important to be able to tell if the data is nominal, ordinal, or scale. A **scale** is data that is on an interval or ratio level of measurement. Another way to define this is that a scale is data with equally spaced intervals. This is why earlier in this chapter, I recommended that we avoid referring to levels of measurement as scales of measurement. For this book, the term "scale" will have a very specific and important definition.

Practice Makes Perfect

Identify the levels of measurement for questions 10 through 13.

10. Number of classes you've skipped in the last year
11. Class rank
12. The year
13. Your favorite type of ice cream
14. What levels of measurement are considered scale?
15. What is the process of converting information into numbers called?

Answers

10. ratio
11. ordinal
12. interval
13. nominal
14. interval and ratio
15. measurement

Credit: iStock.com/jarih

Reliability

In addition to the levels of measurement, reliability plays a huge part in measurement. So now, let's get to what our BFF and our old car has to do with statistics.

Everyone has a best friend. Let's call ours Miranda. Any time we need her, she's got our back. Every. Single. Time. She doesn't even care if it's our mama she has to put into her place. That is because our best friend is reliable; she's consistent.

Then, there's our old, rusty car. We're thankful for it, but sometimes it decides to start and other times we hear that click, click, click noise and know we're going to have to get a jumpstart. The old, rusty car is not consistent; it's *not* reliable.

In statistics, **reliability** simply means consistency of our measures. If the measures we use to quantify our data set aren't reliable, it undermines the integrity of our data analysis. For example, if we go on a diet to lose some weight, we are going to want a scale that measures our weight consistently. We would expect that if we stepped on a scale, stepped off for 1 minute, and then stepped back on, our weight would be the same. However, if we stepped on the scale and saw 225 pounds, stepped off for 1 minute, then stepped back on and saw 160 pounds, we would suspect something wrong with the scale because it was unreliable. Yes! Diet over... time for doughnuts and pizza and street tacos!

Validity

Not only do we want our measures to be reliable, but we also want them to be valid. Within statistics, **validity** means that a measure quantifies what it intended to assess. In other words, the measure does what it said it was going to do. When we were in grade school, if our teacher said, "Hey, on Tuesday, we're going to have a spelling test," what would we expect to be on the test? We would envision getting a list of words to study and then, on Tuesday, to the teacher makes us spell those words. That would be a valid spelling test. Conversely, if the teacher said, "Hey, on Friday, we're going to have a spelling test," we would think it absolutely ridiculous if they broke out a stopwatch and counted how many push-ups we could do in 1 minute. That would be because the spelling test would not assess what it was supposed to. That "spelling test" would *not* be valid.

Error

Even when we have reliable and valid measures, when we measure something, there will always be a degree of error in the score that we obtain. The **obtained score** is the actual score that we get from using a measure. Some people use the term "observed score" as another way to say this. Any obtained score will always include the true score (i.e., the actual amount of the quality) plus **error** (i.e., fluctuations in the measured score not due to the actual ability or quality). Error can come from several sources.

First, we might encounter measurement error. **Measurement error** is the error introduced into the obtained score just because we measured the attribute or quality. Another source of error is **sampling error**, or the error introduced into our data set because we used a sample instead of the overall population. The third source of error we discuss in this book is **random error**, which is unpredictable error introduced into our data set.

To recap, for every score that we get from measuring some quality (i.e., **obtained score**), there is the actual amount of the quality (i.e., **true score**) and fluctuations in the obtained score not due to the actual ability or quality (i.e., **error**). This idea of error in our obtained scores plays into the rationale for conducting statistical analyses.

Obtained Score = True Score + Error

Basically, this means that if we get a 99% on our first statistics exam and our best friend gets a 98%, we cannot brag about how much smarter we are and that we know so much more. Remember, the obtained score is the actual score we got on the test, while the true score is the actual reflection of our real ability. Let's break this down a little bit more in Figure 1.3.

Figure 1.3 ▪ Test Score Example of Obtained Score = True Score + Error

You				
Obtained score	=	True score	+	Error
99	=	96	+	3
BFF				
Obtained score	=	True score	+	Error
98	=	100	+	(–2)

If we got a better score on the exam, how is it possible that our best friend actually knows more than we do? Well, there could be several reasons. We could have gotten lucky and guessed three of the answers correctly. Our best friend could have stayed up way too late the night before and gotten no sleep because they went to the Radiohead concert and had difficulty retrieving answers they actually knew. The fundamental concept of statistical error provides a rationale for why we need statistics. Just because scores appear to be different, it doesn't mean they actually are.

Practice Makes Perfect

16. What is in every obtained score?
17. What term is used to describe that your survey on music preference actually measures someone's music preference?
18. What term is used to describe that your depression inventory consistently measures depression?

Answers

16. true score and error
17. validity
18. reliability

Research Methods Fundamentals

In addition to basic measurement, to be successful in statistics, it's also important to know some of the research methods fundamentals. These would include the types of variables and four basic types of research design.

Variables

There are three types of variables we need to understand to be successful in statistics: dependent, independent, and confounding variables. To understand independent and dependent variables, let's imagine a pretend research study. Pretend we were curious to see whether traditional weight training or CrossFit would result in more gains in the gym. To find out which one resulted in more gains in amount of muscle, there are several things we would have to do. First, we would need to identify the independent variable. The **independent variable** (symbolized by **IV**) is the grouping variable. In this example, the IV would be types of workouts. Then, the levels of the independent variable (symbolized by ***k***) are the number of groups within the grouping variable. In this example, our independent variable, types of workouts, has two levels: (1) traditional weight training and (2) CrossFit. Sometimes, people refer to the levels of the independent variable as conditions.

In addition to establishing the IV, we also need to identify the dependent variable. The **dependent variable** (symbolized by **DV**) is the variable that we want to measure in our study. Since we are interested in finding out if there are differences in gains in the amount of muscle between the traditional weight-training and CrossFit conditions, our DV is muscle gain. To measure that muscle gain, we could use what they call a dual-energy x-ray absorptiometry (DEXA) scan that can provide us information about not only muscle gain but also body fat percentage and bone density.

When we conduct a research study, many times, we want to see if different conditions of the IV in the study result in different amounts of the DV. If we apply this to the traditional weight-training versus CrossFit example, we are trying to determine if the different conditions (traditional weight training and CrossFit) result in significantly different levels of the DV (muscle gain as measured by the DEXA scan). Significantly different means that we believe that these are real differences, not differences due to chance or error. When we design a study, we want to do so in a way that minimizes the impact of confounding variables on our results as much as we can.

A **confounding variable** is any variable that impacts the DV other than the IV. As a researcher, we want to control for confounding variables as much as we can. For this pretend study, there could be several potential confounding variables. One confounding variable could be the number of calories that a person consumes each day above what they would normally need. Another potential confounding variable could be the typical amount of protein a person consumes each day. A third potential confounding variable could be how much physical activity a person engages in outside of the traditional weight-training or CrossFit workout regimens. If we did this study, we would want to design it so that we eliminated as many other factors, or confounding variables, as possible that might impact the dependent variable other than our independent variable.

Basic Research Designs

Throughout this book, we will distinguish among four basic research designs, including (1) between-groups, (2) within-group, (3) matched-groups, and (4) bivariate designs. The first three types of basic research design deal with the participants who are in each level (i.e., conditions) of the independent variable. A **between-groups design** is when there are different participants in each level of the IV. For example, if I wanted to test the differences between my face-to-face and online statistics students on their final exam grades, it would be a between-groups design. A **within-group design** is when we have the same participants in each level of the independent variable. Sometimes, we will hear within-group design also called **repeated measures design**. An example of a within-group design would be any type of pretest–posttest study in which a teacher gives a class a pretest and then compares its scores to the same class's posttest scores to see if there were gains in the students' learning.

The most complicated of these basic research designs is the matched-groups design. In a **matched-groups design**, we measure all the participants on some potential confounding variable and then try to make sure that participants in each level of the independent variable have approximately the same amount of it. We want to make certain that the confounding variable is not the reason for any changes in the dependent variable. One benefit of matched-groups design is that it allows for the groups to be close to equivalence at the beginning of the study in at least one characteristic.

To demonstrate one way to use matched-groups design, let's go back to our traditional weight-training and CrossFit example from our discussion of confounding variables. In this example, we may want to have each of the conditions include people with about the same amount of body fat percentages (i.e., lean-to-fat ratios) at the beginning of the study. To do this, we would first measure the body fat percentages of all our participants. Then, we would put the participants in order from highest to lowest body fat percentage, take the first two participants and randomly assign them to groups, and then continue this process for all the rest of the participants. Sometimes people refer to matched-groups design as matched-subjects design.

Step-by-Step Example of the Ranking Method of Matched-Subjects Design

Step 1 Use the DEXA scan to obtain the starting body fat percentages for all the participants in the study.

Table 1.3 ▪ Participant Assignment to Conditions

Participant	Body Fat Percentage
1	21
2	23
3	28
4	25
5	32
6	33
7	35
8	34

Step 2 Put the participants in order from highest to lowest body fat percentage.

Table 1.4 ▪ Body Fat Percentage Data Set in Order of Magnitude

Participant	Body Fat Percentage
7	35
8	34
6	33

(*Continued*)

Table 1.4 ▪ Body Fat Percentage Data Set in Order of Magnitude (*Continued*)

Participant	Body Fat Percentage
5	32
3	28
4	25
2	23
1	21

Step 3 Take the two participants with the highest body fat percentages and randomly assign one of the participants to one of the conditions (traditional weight training or CrossFit) using a random number generator and then place the other participant in the other condition.

Within this example, we would match Participants 7 and 8 together because they were the two participants with the highest body fat percentages. This means that we would use a random number generator to randomly assign Participant 7 to one of the conditions and then place Participant 8 in the other condition.

Let's say that when we used a random number generator, we assigned 1 to equal the traditional weight-training condition and 2 to equal the CrossFit condition. We would take Participant 7 and use the random number generator to assign it a 1 or 2. Let's pretend that the random number generator output the number 1. This outcome would mean that we would assign Participant 7 to the traditional weight-training condition, and then we would assign Participant 8 to the other condition (i.e., the CrossFit condition), as demonstrated in Table 1.5.

Table 1.5 ▪ Matching Participants 7 and 8 and Assigning Them to Conditions

Traditional Weight Training	CrossFit
Participant 7	Participant 8

Step 4 Repeat this process until all participants have been assigned to either the traditional weight-training or CrossFit conditions. For example, this means that we would continue with this process for Participants 6 and 5, Participants 4 and 3, and Participants 2 and 1, until all the participants were assigned to their conditions (refer to Table 1.6).

Table 1.6 ▪ Participant Assignment to Conditions

Traditional Weight-Training Condition	CrossFit Condition
Participant 7	Participant 8
Participant 5	Participant 6

Traditional Weight-Training Condition	CrossFit Condition
Participant 4	Participant 3
Participant 2	Participant 1

*This would be a matched-groups design, because the groups are related to one another based on body fat percentages.

The last of the four types of basic research design we learn about in this book includes the bivariate research design. Unlike the other three types, a **bivariate research design** is when we ask participants to provide us data on two variables. For example, if we wanted to examine if there was a relationship between the number of stickers someone had on their water bottle and their level of creativity, we might design a study with a bivariate research design. To accomplish this, we would take a sample of participants and ask each of the participants how many stickers they had on their water bottle. We would also measure the creativity levels of each of the same participants.

Practice Makes Perfect

If I wanted to see if this chapter helped students learn, I might randomly assign half the students in this class to read the chapter and the other half of the students in this class to *not* read this chapter (lucky). Then, I could compare how well the two groups did over a test over the chapter material. In this example, answer the following questions:

19. What is the dependent variable?
20. What is the independent variable?
21. How many levels are there to the independent variable?
22. Which one of the three basic research designs would this be?

Answers

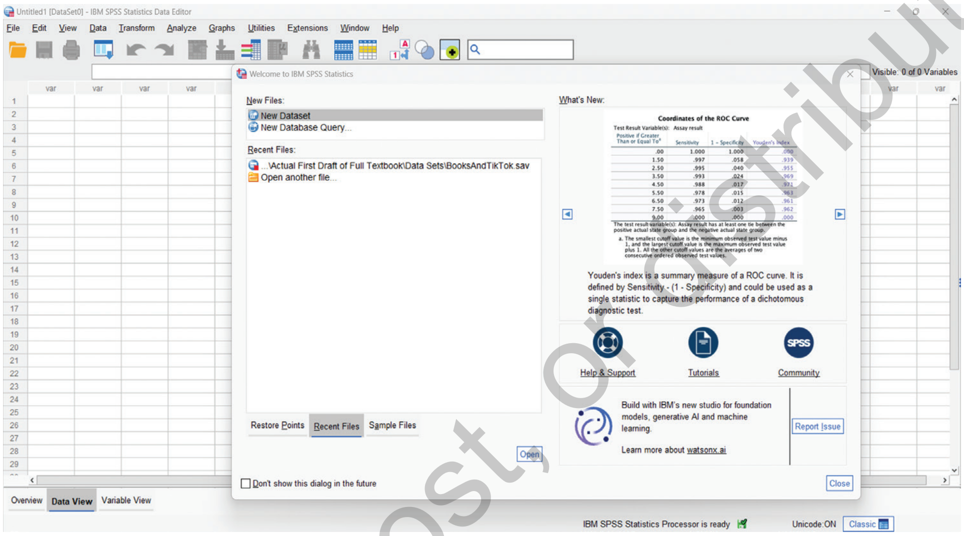
19. chapter test scores
20. reading assignment condition
21. 2
22. between-groups design

SPSS Bonus Content Unlocked: Setting Up SPSS Data Files

Although this book primarily teaches how to conduct statistical analyses by hand like most introductory statistics courses, in the real world you'll likely use statistical software like SPSS to analyze data sets. Because of this, this book provides the opportunity to learn how to conduct statistical analyses in both ways. To use SPSS properly we need to learn how to set up our data sets, which we'll learn about here. Setting up the data set file correctly will help prevent problems later when using it to analyze the data set.

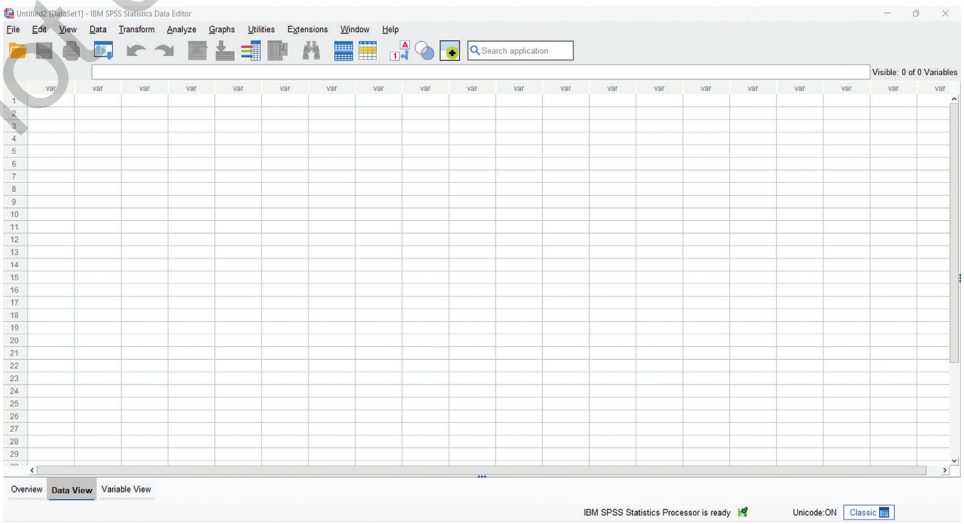
When we first open SPSS, it shows us a "Welcome to IBM SPSS Statistics" dialog box where we can select "New Dataset" to create a new data set like we find in Figure 1.4.

Figure 1.4 ▪ Welcome to the SPSS Environment



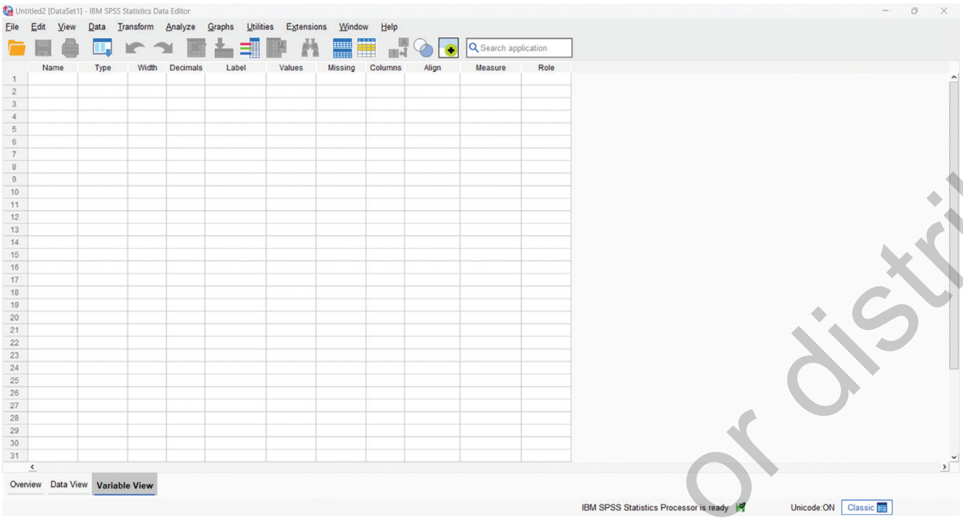
After we double click the "New Dataset" button, SPSS leads us to the "Data View" like we find in Figure 1.5. The **"Data View"** in SPSS shows the data set in a spreadsheet format. The variable "Name" is at the top of each column, with the data for that variable going down its corresponding column.

Figure 1.5 ▪ Empty Data View



To get the variable "Name" to populate across the top row in the "Data View", we must navigate to the "Variable View" to give SPSS more information about the variable (refer to Figure 1.6).

Figure 1.6 ▪ SPSS Variable View



The first row of the SPSS spreadsheet in "**Variable View**" includes the headers for each of the columns. We find brief descriptions of these headers in Table 1.7.

Table 1.7 ▪ SPSS Variable View Header Descriptions

Header Title	Description
Name	Allows the user to create a name for the variable.
Type	Allows the user to identify the variable's broad type of data (e.g., numeric, string).
Width	Allows the user to customize how many characters the column will display in the "Data View."
Decimals	Allows the user to identify the number of digits displayed to the right of the decimal for numerical values.
Label	Allows the user to create a brief label for the variable name. If there's a label, SPSS will use this in place of the variable name in any output.
Values	Allows the user to assign meaning to numerical values (e.g., 1 = female, 2 = male, 3 = intersex, 4 = prefer not to respond) to enhance the readability of SPSS Output.
Missing	Allows the user to designate numbers for missing values in the data set (e.g., 999 could mean that the participant didn't provide an response).
Columns	Allows the user to customize the column length in the "Data View."
Align	Allows the user to customize if the data are aligned left, center, or right in the "Data View."
Measure	Allows the user to identify whether the data is nominal, ordinal, or scale.
Role	Allows the user to identify the role of the variable (e.g., input, target).

Name

The first column "**Name**" is where we put the name of the variable. There are several rules we must follow when naming our variables. If we don't follow the rules, it will result in error warnings like we find in Figures 1.7.

Figure 1.7 ▪ SPSS Naming Rules and Error Messages

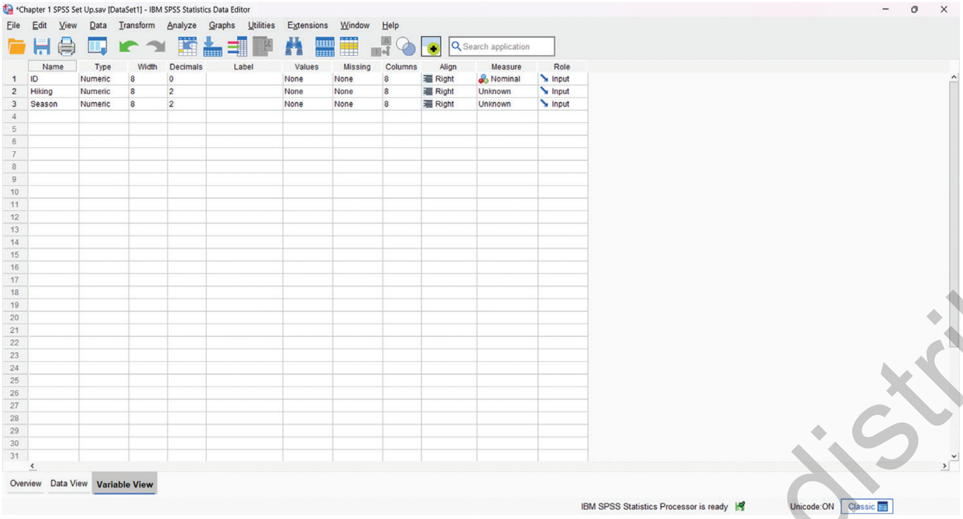
Naming Rule	Error Message
Refrain from including a space in the variable name.	IBM SPSS Statistics – Variable name contains an illegal character. OK
Refrain from ending the variable name in a period or underscore.	IBM SPSS Statistics – Variable name contains an illegal last character. OK
Refrain from using reserved words in the variable name, such as AND, BY, EQ, WITH, NE, LT, GT, GE.	IBM SPSS Statistics – Name contains a reserved word. OK
Variable name must be less than 33 characters long.	IBM SPSS Statistics – The variable name is too long. Please reduce the length. OK
Each variable name must be unique.	IBM SPSS Statistics – This variable name duplicates an existing variable name. OK

To "Name" our variables, we first need to navigate to the "Variable View." In the first column, we can name each of our three variables: ID, Hiking, and Season. When we insert a "Name" for a variable, it automatically fills out some of the information for the variable in its corresponding row that we may need to edit depending on the specific variable (refer to Figure 1.8).

Type

After we give the variables a "Name," we can identify the "Type" of data for each variable. In introductory statistics, we'll typically identify the data as "String" or "Numeric." We identify data as "**String**" when we want SPSS to treat the data as text rather than numbers. For example, every time I create a new data set, I like to create an ID variable to have a number to identify

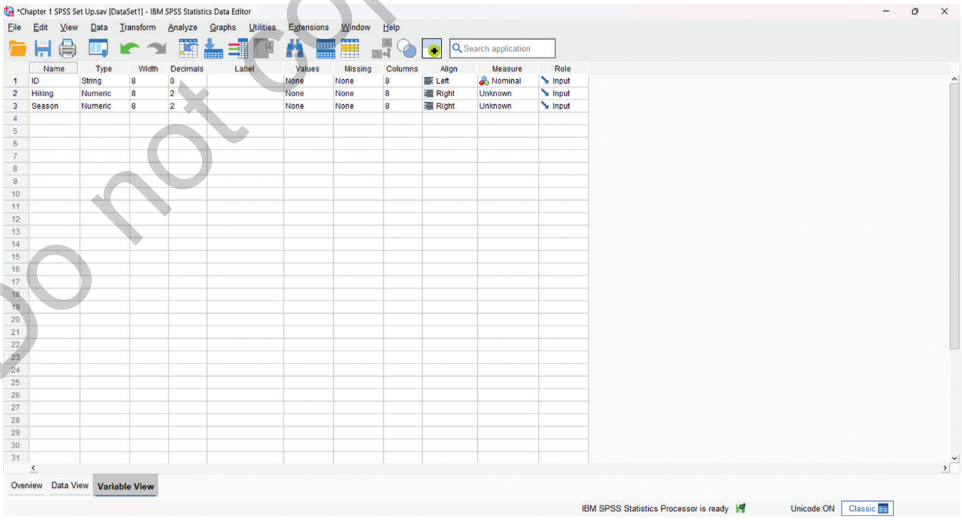
Figure 1.8 ▪ "Name" in Variable View for Each Variable



each of the participants. We would identify this variable as "String" because we'll use this information to keep track of which participant is which (refer to Figure 1.9).

In SPSS, we identify the data as "**Numeric**" when the data are numerical numbers (e.g., number of times we went hiking last summer, favorite season). By identifying the data "Type" as numeric, it'll allow us to identify the data as nominal, ordinal, or scale data under the "Measure" column. For example, if we wanted collect data on how many times the participants in our sample went hiking last summer or what their favorite seasons are, we would identify the data "Type" as numeric.

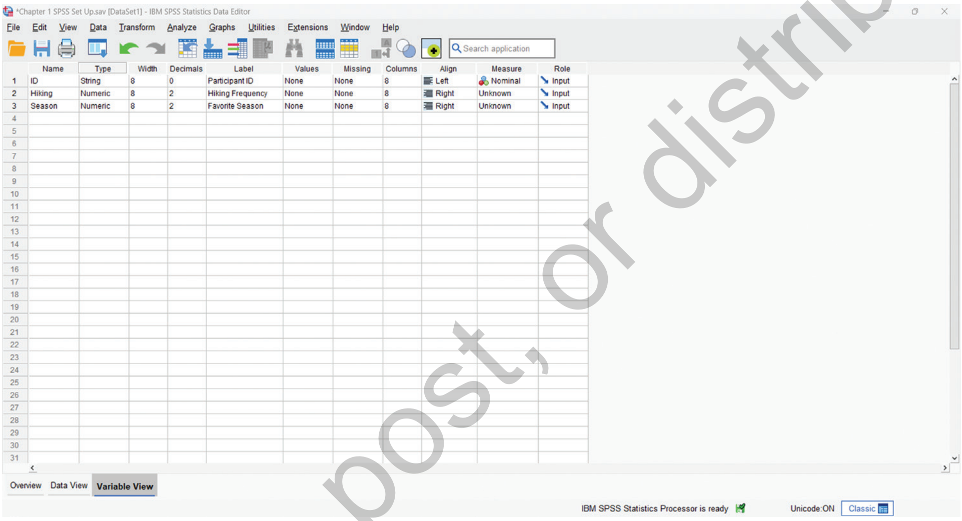
Figure 1.9 ▪ Completed "Type" in Variable View for Each Variable



Label

Under "**Label**" we can write a better variable name for the variable "Name" that had to follow very specific SPSS formatting guidelines. If we provide labels in the variable view, then SPSS will use those for the names rather than the variable "Name" that we identified for each variable in the first column in any output. The "Label" for the variable doesn't have to follow the same formatting rules as the variable "Name." For example, we could label "ID" as "Participant ID," "Hiking" as "Hiking Frequency," and "Season" as "Favorite Season" like we find in Figure 1.10.

Figure 1.10 ▪ Completed "Label" in Variable View for Each Variable



Values

We use the "Values" column when participants answer demographic questions such as biological sex or Likert-scale data. For example, if we asked participants about their favorite season where 1= winter, 2= spring, 3= summer, and 4= winter, we could identify what each of the numbers represent. To identify what each of the numbers mean, we click the "Values" cell for the variable, and it'll lead us to a "Value Labels" dialog box like we find in Figure 1.11.

Once we're into the "Value Labels" dialog box, we click the plus symbol that's surrounded by a blue box. That will allow us to put a number under "Value" and it's label under "Label." Once we provide the label for each variable, we click "OK" (refer to Figure 1.12).

After we click "OK" we have finished entering the labels for each of the values for the variable like we find in Figure 1.13.

Figure 1.11 ▪ "Value Labels" Dialog Box

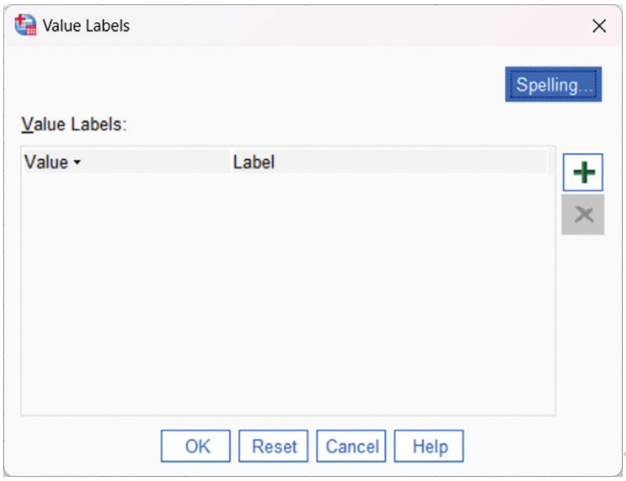


Figure 1.12 ▪ Completed "Value Labels"

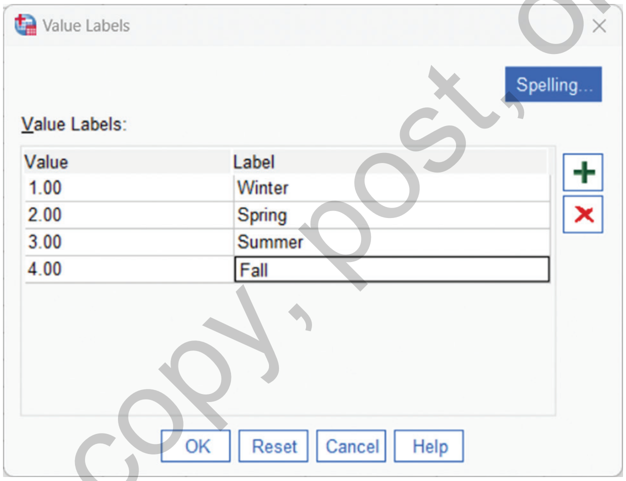
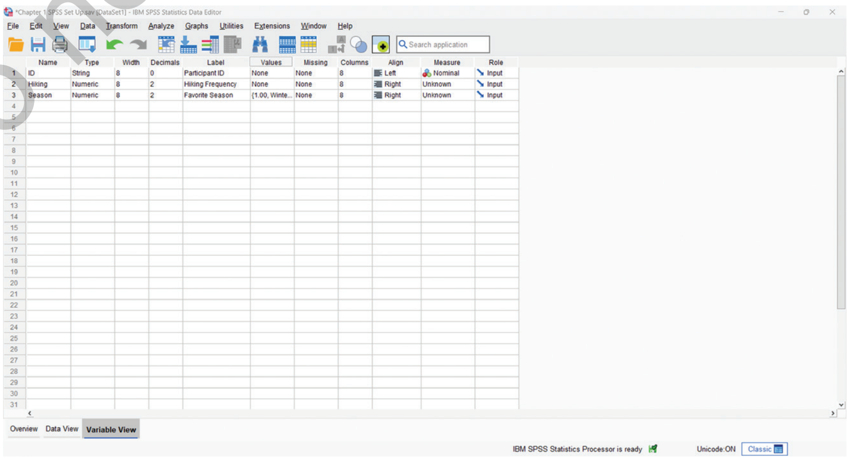


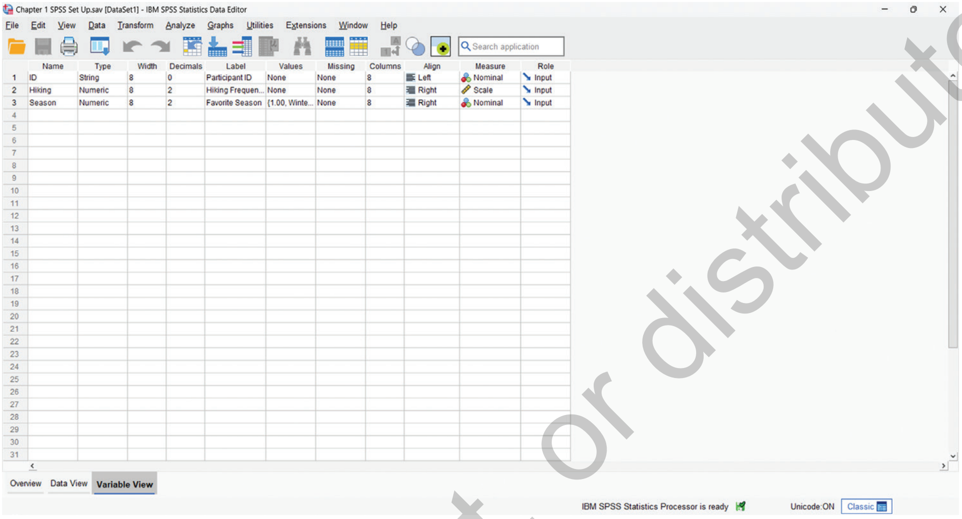
Figure 1.13 ▪ Completed "Values" in Variable View



Measure

Under the "Measure" column we identify if the data is nominal, ordinal, or scale, like we did for the "ID," "Hiking," and "Season" (refer to Figure 1.14).

Figure 1.14 ▪ Completed "Measure" Column



Pro Tip: If the SPSS won't let you identify the data as "Scale" under "Measure" and it's scale data, double-check that the data "Type" is listed as "Numeric" rather than "String."

Role

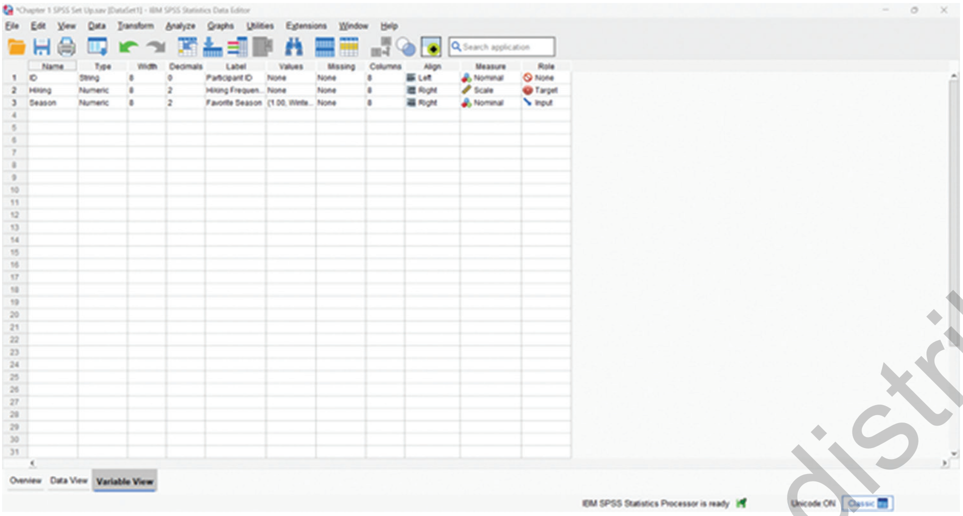
The "**Role**" defines what each of the variables will do in the statistical analyses. Find a list of the relevant roles in Table 1.8.

Table 1.8 ▪ SPSS Role Descriptions

Role	Description
Input	Independent or predictor variable
Target	Dependent or outcome variable
Both	Used as both independent and dependent variables
None	Has no role
Partition	Splits the data set into separate samples

SPSS automatically defaults the role for every variable as "Input." We can change the role to what we're using the variables for, but that won't change what we can do with the data. For our example, we could identify the role for "ID" as "None," "Hiking" as "Target," and "Season" as "Input" like we find in Figure 1.15.

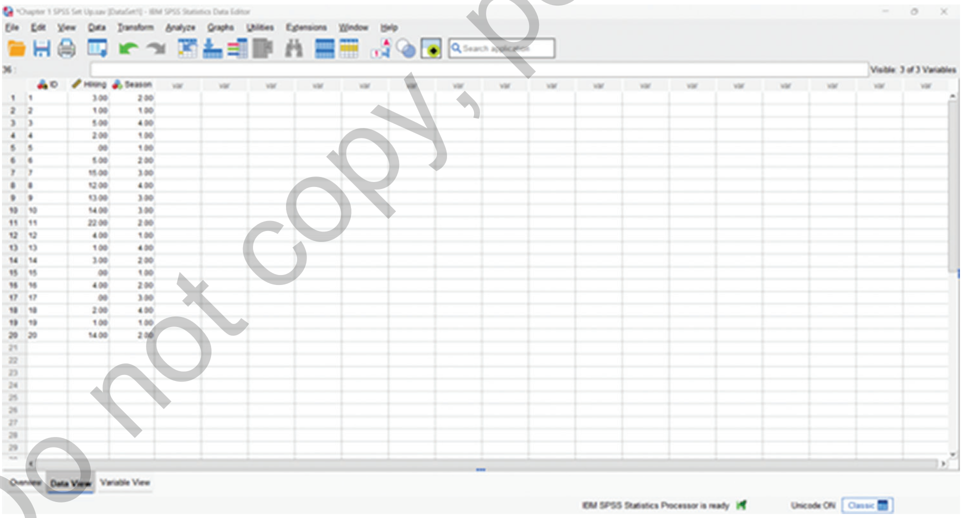
Figure 1.15 ▪ Complete “Role” in Variable View



Data View

After we fill out the information for each of our variables in “Variable View” we can navigate to the “Data View” to insert the data into the spreadsheet like we find in Figure 1.16.

Figure 1.16 ▪ Completed “Data View”



Pro Tip: Click the toggle button (refer to Figure 1.17) to go back and forth between the numerical values and their corresponding “Value Labels” like we find in Figure 1.18.

Figure 1.17 ▪ Toggle Button

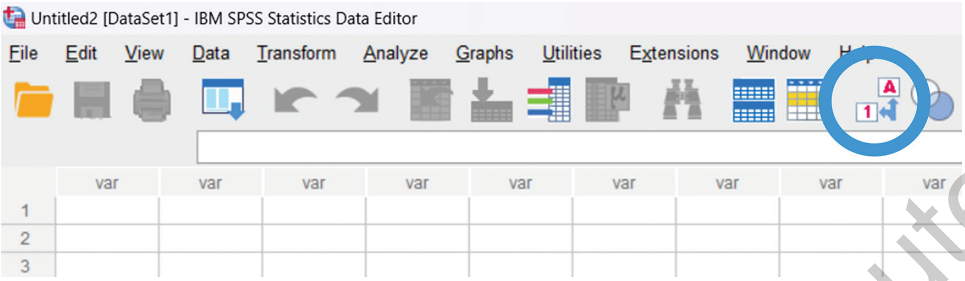
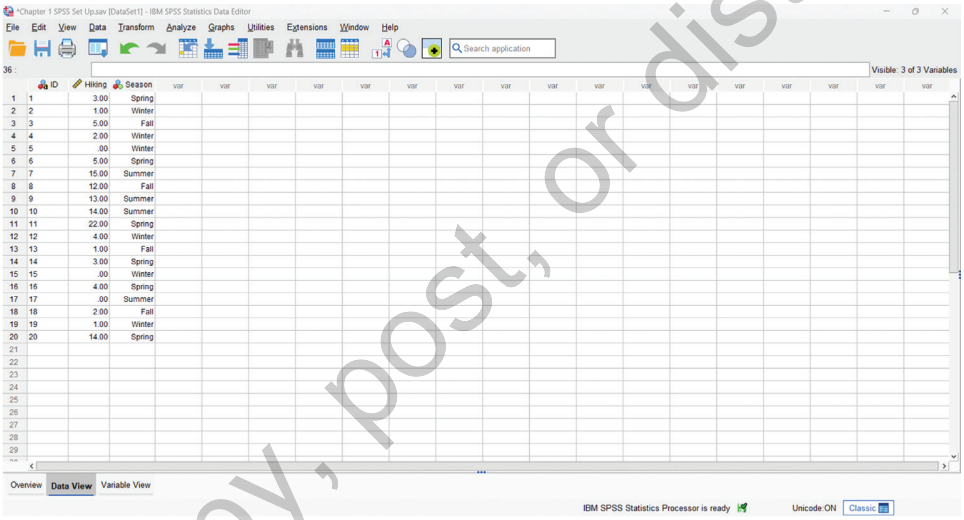


Figure 1.18 ▪ Toggling Between Numbers and Labels



Summary

This chapter provided the foundation for the entire book by giving the background and context necessary to not only prepare us to conduct a variety of statistical analyses but prepare us to solve problems in the real world. We learned that data breaks off into two major branches: qualitative and quantitative. We learned that in statistics, we use quantitative data to make summaries (i.e., descriptive statistics) and inferences (i.e., inferential statistics) about data. To begin to understand how to conduct quality statistical analyses in real life, we learned several closely related research methods concepts that include levels of measurement, reliability, validity, dependent variables, independent variables, basic research designs, sampling techniques, and error.

Symbol Guide

X	data from participants
N	population size

n sample size

DV dependent variable

IV independent variable

k levels of the independent variable

Terms to Know

Between-groups design
Bivariate research design
Confounding variable
Convenience sampling
Data
Data set
Dependent variable
Descriptive statistics
Error
Generalization
Independent variable
Inferential statistics
Interval
Levels of measurement
Matched-groups design
Measure
Measurement
Measurement error
Mixed-methods research
Nominal
Obtained score
Ordinal
Participant
Population
Qualitative data
Quantitative data
Random error
Random sampling
Ratio
Reliability
Repeated-measures design
Sample
Sampling
Sampling error
Scale
Snowball sampling
True score
Validity
Within-group design

Putting in the Work

1. True or False. Quantitative data is data in number form.
2. True or False. In statistics, we analyze qualitative data.
3. What general branch of statistics uses a number or numbers to represent an entire data set?
   a. inferential statistics
   b. descriptive statistics
   c. measures of central tendency
   d. measures of distribution
4. What type of statistics would it be if we used data from Ms. Garrett's third-grade class to make guesses about what was going on in the entire third grade at Bellview Elementary School?
   a. inferential statistics
   b. descriptive statistics
   c. measures of central tendency
   d. measures of distribution

5. What is a key feature of random sampling?
   a. It always represents the population that we took the sample from perfectly.
   b. Every member of the population has an equal chance of being selected into the sample.
   c. It uses current research participants to help recruit additional participants for the study.
   d. It breaks down populations into subgroups to ensure each subgroup is adequately represented in the sample.

6. What is the process of turning information into numbers called?
   a. reliability
   b. measure
   c. measurement
   d. validity

7. What level of measurement would your favorite type of doughnut be if 1 = glazed, 2 = jelly, 3 = traditional, 4 = Boston cream, and 5 = long john?
   a. nominal
   b. ordinal
   c. interval
   d. ratio

8. What level of measurement would the weight of your frozen yogurt be?
   a. nominal
   b. ordinal
   c. interval
   d. ratio

9. What level of measurement is your cousin's ACT Score (the scores range from 1 to 36)?
   a. nominal
   b. ordinal
   c. interval
   d. ratio

10. What level of measurement would be the top five rankings of contestants in the Miss Universe pageant?
   a. nominal
   b. ordinal
   c. interval
   d. ratio

11. What does it mean that a measure is reliable?
   a. It accurately measures what it is supposed to.
   b. It consistently measures the same thing over time.
   c. It resulted in statistically significant results.
   d. It is approved as ethical by the Institutional Review Board.

12. Which of the following is NOT a major type of descriptive statistics?
   a. measures of distribution
   b. measures of variability

  c. measures of variance
  d. measures of central tendency

13. Which of the following options demonstrate a valid measure?
  a. The number of sit-ups completed in a minute for a fitness test
  b. A bathroom scale that shows the same weight every time it gets used
  c. Using the number of fruits someone eats as a happiness measure
  d. A measure approved by the Institutional Review Board

14. What is *not* a potential source of error?
  a. sampling error
  b. random error
  c. measurement error
  d. true error

15. What is in every obtained score?
  a. *z*-score
  b. error
  c. difference score
  d. deviation score

16. Which of the following are measured on a scale?
  a. places in a race
  b. number of piercings
  c. military rank
  d. favorite type of flower

17. What would the dependent variable be in a study that examined the differences in the number of tattoos among people who live in urban, suburban, and rural communities?
  a. community type
  b. between-group design
  c. number of tattoos
  d. three

18. What type of basic research design would it be if we asked participants to tell us how many times they went to the dentist in the past year and how much money they each made in the same year?
  a. between-group design
  b. within-group design
  c. bivariate design
  d. matched-groups design

19. What is the independent variable in the following example? Researchers examined if there were significant differences in the amount of irritability participants reported when listening to music in rap, country, or pop conditions.
  a. irritability levels
  b. music condition
  c. three
  d. within-group design

20. Match the symbol to the correct meaning.

N	data from participants
IV	sample size
k	population size
n	independent variable
DV	dependent variable
X	levels of the independent variable

Chapter 2

Measures of Central Tendency: Attraction

iStock.com/LeoPatrizi

How does rating people on how attractive they are on a scale of 1 to 10 help you learn more in depth about measures of central tendency? Read on to find out…

Learning Objectives

2.1 Define the three measures of central tendency in your own words.

2.2 Calculate the mean of a data set.

2.3 Calculate the mode of a data set.

2.4 Calculate the median of a data set.

2.5 Determine what measure of central tendency can be calculated with data on a nominal level of measurement.

2.6 Determine what measures of central tendency can be calculated with data on an ordinal level of measurement.

2.7 Determine what measures of central tendency can be calculated with data on a scale with at least one outlier in the data set.

An Introduction to Descriptive Statistics

If we remember from the last chapter, statistics splits into two major branches: descriptive statistics and inferential statistics. To review, the goal of descriptive statistics is to summarize our data set in one or a few numbers. For example, if we collected a data set with the number of times participants ordered food delivery (e.g., Uber Eats, DoorDash) last week, we might want to find one number to try to represent what's going on with our entire sample (reference Table 2.1). What would be one way that we could represent the data from our sample in one number?

iStock.com/miodrag ignjatovic

Table 2.1 ▪ Food Delivery Data Set

Participant	*X*
1	1
2	3
3	0
4	2
5	0

There are many ways we could represent this data set in one or a few numbers, but some examples could include the mean, median, mode, and range. Each of these are types of descriptive statistics. Many researchers break descriptive statistics down further into three major branches including measures of central tendency, measures of variability, and measures of distribution. In this chapter, we will learn more about the measures of central tendency.

Measures of central tendency identify the center of a data set in one or a few numbers. Although there are several types of measures of central tendency, in this chapter, we will specifically learn about the mean, median, and mode. We will then use our knowledge of levels of measurement from the last chapter to identify which types of measures of central tendency are appropriate and when. In future chapters, we will learn more about using measures of variability and measures of distribution, but for now, we will need only to understand their basic definitions. **Measures of variability** provide us with information on how the data in our data set varies within the set, while **measures of distribution** attempt to tell us about how frequently each value occurs in our data set.

We can use descriptive statistics for both samples and populations. When we calculate a descriptive statistic for a sample, we call that number a **statistic**. Then, when we calculate a descriptive statistic for a population, we call that number a **parameter**.

Practice Makes Perfect

1. What are the three major branches of descriptive statistics?
2. What is the goal of measures of central tendency?
3. Would it be a statistic or parameter if we found the mean time it took the entire population of climbers to climb the Manitou Incline today?

Answers

1. measures of central tendency, measures of variability, and measures of distribution
2. to describe the center of the data set in one or a few numbers
3. parameter

Mean

One of the best-known types of measures of central tendency is what people typically call the average or **mean**. However, there is some inconsistency on how people use the word "average" and even the word "mean." When we use the word "mean" in this book, we refer specifically to the **arithmetic mean**. The arithmetic mean is when we add up all the data from the participants in our data set and then divide that sum by the number of participants. Let's practice calculating the mean with the "How Much Did You Complain in a 24-hour Period? Data Set" (consult Figure 2.1).

For this example, let's pretend that we went onto social media and asked our friends to report how often they complained in a 24-hour period. Then, we compiled this data into Figure 2.1, How Much Did You Complain in a 24-Hour Period? Data Set.

To calculate the mean, we add up all of the times the participants complained in a 24-hour period and then divide that sum by the number of participants. Here, the first participant reported that he complained 5 times, the second participant also reported that she complained 5 times, the third participant reported that she complained 10 times (don't judge her, at least she's honest), and the last participant reported complaining 2 times. If we add up the number of times the participants reported that they complained, we get a total of 22. Then we divide that sum by the number of participants (which in this case is 4) to get the mean. The mean in this example equals 5.50 complaints.

Figure 2.1 ▪ How Much Did You Complain in a 24-Hour Period? Data Set

Participant	How much did you complain in a 24-hour period?
	5 times
	5 times
	10 times
	2 times

iStock.com/Antonio_Diaz; iStock.com/Valeriy_G; iStock.com/recep-bg; iStock.com/Igor Alecsander

Although we were able to calculate the sample mean by adding up all the participant data in our data set and dividing that sum by the number of participants in the sample, statistics uses very specific formulas and symbols for calculating the sample and population means. For example, we calculate the mean of a sample using Formula 2.1.

Formula 2.1 Sample Mean

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

This formula uses a couple of symbols we learned in the last chapter such as X (participant data) and n (sample size), along with new symbols including $\bar{X}$ and Σ. The symbol $\bar{X}$ (called X-Bar) represents the sample mean. The summation symbol (Σ) is a mathematical operator that tells us to add up everything that follows it. Now that we know what the summation symbol means, what does $\sum X$ tell us to do?

$\sum X$ tells us to add all the participant data in our data set together. Now, let's take our new knowledge of the sample mean formula and take the information from Figure 2.1 and put it into a standard table (consult Table 2.2). After that, let's use it to solve for $\overline{X}$.

Table 2.2 ▪ How Much Did You Complain in a 24-Hour Period? Data Set in Proper Form

Participant	X
1	5
2	5
3	10
4	2
$n = 4$	$\sum X = 22$

$$\sum X = 5 + 5 + 10 + 2 = 22$$

$$\overline{X} = \frac{\sum X}{n} = \frac{22}{4} = 5.50$$

In a similar way, the population mean has its own formula with its own symbols. Statisticians use the lowercase Greek letter mu (***μ***) to represent the population mean. Unlike some of the other formulas that we'll learn, the process for calculating the sample and population means is the same; the difference between the formulas is the symbols.

Formula 2.2 Population Mean

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

Formula 2.2 shows us that to get the population mean, we add all the data from our participants in the entire population ($\sum X$), and then we divide that sum by the population size (N). Although we use $\overline{X}$ to symbolize the sample mean and μ to stand for the population mean, sometimes, researchers also use the symbol ***M*** to represent both sample and population means.

Practice Makes Perfect

4. What is the symbol for sample mean?
5. What does $\sum$ tell us to do?
6. How do we calculate the mean?

Answers

4. $\overline{X}$
5. Tells us to add up everything that follows it
6. We add up all the data from the participants and divide that sum by the number of participants.

Mode

The **mode** tells us the value(s) that occur most frequently in our data set. Often, statisticians use the symbol ***Mo*** to represent the mode. Let's take our data from when we asked our participants how often they complained in a 24-hour period and organize it into an ungrouped frequency distribution table to help us easily find the mode of our data set (consult Table 2.3). A **frequency distribution table** shows us how frequently each value occurs in our data set. (We will learn to properly format these in Chapter 4.)

Table 2.3 ▪ How Much Did You Complain in a 24-Hour Period? Ungrouped Frequency Distribution Table

Number of Complaints (*X*)	Frequencies
10	1
9	0
8	0
7	0
6	0
5	2
4	0
3	0
2	1

In Table 2.3, we find that the $Mo = 5$ because it was the most frequently occurring value in our data set. Two people reported complaining 5 times in the last 24-hour period, whereas only one person reported complaining 10 times and another twice in the same time period. Although this data set only had one mode, sometimes, data sets will have more than one. If the previous data set had been 1, 0, 2, 10, 2, and 1, the data set would have had two modes (1 and 2). If a data set has two modes, we say that it is **bimodal**. If there is more than one mode, we can also describe the data set as **multimodal**.

Practice Makes Perfect

7. What's the *Mo* of the following data set? 10 15 5 7 8 10
8. What's the *Mo* of the following data set? 5 8 5 7 8 3 1

Answers

7. $Mo = 10$
8. This distribution is bimodal, so it has two modes: 5 and 8.

Median

This brings us to the last measure of central tendency of the chapter, the median. The **median**, symbolized by ***Mdn***, is the number that describes the very middle value of the data set when we line it up from the smallest to largest values. The way we find the median depends on if there is an odd or even number of participants in our data set.

If there is an odd number of values in the data set, it will be the very center value when we line the data up in size from smallest to largest. To practice this, let's pretend we asked five people their age, and they told us that they were 36, 12, 14, 8, and 55 years old.

Steps to Finding the Median of a Data Set With an Odd Number of Participants

Step 1 Put the data in order from least to greatest.

8 12 14 36 55

Step 2 Determine whether there is an odd or even number of participants.

In this data set, there is an odd number of participants.

Step 3 For a data set with an odd number of participants, find the center value to find the median.

8 12 14 36 55

Here, we find a median of 14. However, if there is an even number of participants in our sample, the median will be the mean of the two center values. To illustrate this, let's pretend we asked six people how many pets they had, and they told us 2, 0, 2, 3, 0, and 1 pets.

Steps to Finding the Median of a Data Set With an Even Number of Participants

Step 1 Put the data in order from least to greatest.

0 0 1 2 2 3

Step 2 Determine whether there is an odd or even number of participants.

In this data set, there is an even number of participants.

Step 3 Find the two center values in the data set.

0 0 1 2 2 3

Step 4 For a data set with an even number of participants, find the mean of the two center numbers to find the median.

$$Mdn = \frac{1+2}{2} = \frac{3}{2} = 1.50$$

After we complete these steps, we find that the median for this data set equals 1.50.

Now that we know how to calculate the median of a data set, let's go back to our data from Table 2.2 to find the median number of complaints our sample reported in a 24-hour period.

Step 1 Put the data in order from least to greatest.

2 5 5 10

Step 2 Determine whether there is an odd or even number of participants.

In the Table 2.2 data set, we had an even number of participants.

Step 3 Find the two center values in the data set.

2 5 5 10

Step 4 For a sample with an even number of participants, find the mean of the two center values.

$$Mdn = \frac{5+5}{2} = \frac{10}{2} = 5.00$$

Here, we find that the median number of complaints our participants reported in 24 hours was 5.00.

Practice Makes Perfect

Examine the following data set in Table 2.4 to answer questions 9 through 11 (remember to round to two places past the decimal):

Table 2.4 ▪ Phone Obsession Data Set

Participant	How many times did you check your phone in the last hour? (X)
1	20
2	8
3	1
4	10
5	1
6	1

9. What is the $\bar{X}$?
10. What is the *Mdn?*
11. What is the *Mo*?

Answers

9. 6.83
10. 4.50
11. 1.00

Calculating Measures of Central Tendency for Data on a Nominal Level of Measurement

One of the best skills in statistics is to know which descriptive statistics can best represent our data sets. Several different factors go into determining which measure or measures of central tendency might be the best to describe a data set. In the first chapter, we learned about the different levels of measurement of data (i.e., nominal, ordinal, interval, and ratio). In this part of the chapter, we learn how to determine which measures of central tendency can be used for data on each level of measurement. Spoiler alert: We won't be able to calculate the mean, median, and mode for data on all levels of measurement. The understanding we gain will have important applications throughout the rest of this book.

To explore which measures of central tendency we can use for data on a nominal level of measurement, let's examine people's favorite colors. To start, what's your favorite color? Is it blue, eggshell, violet, neon green, or gold? When trying to figure out the best measures of central tendency to use in this case, we need to reflect on what nominal data means. The limited amount of information that nominal data gives us severely limits what we can do with it.

Earlier, we learned that data on a nominal level of measurement is where a number is assigned to a category or group. We learned that even though nominal data is represented by a number, there isn't a real numerical value to it. For example, we could assign 1 = Blue, 2 = Black, 3 = Orange, 4 = Green, 5 = Purple, and 6 = Yellow or, alternatively, 1 = Orange, 2 = Blue, 3 = Black, 4 = Purple, 5 = Yellow, and 6 = Green. This is because data on a nominal level of measurement have no numerical value or order to them. If we really wanted chaos, we could have made the number 1,123 = Blue, 666 = Black, 20 = Orange, 777 = Green, 888 = Purple, and 13 = Yellow. One strategy to remember that "nominal" means "category" is to think of nominal as a name. In nominal data, the number represents the name of an item but not a real quantitative value.

Let's use the following pretend data set of my friends' favorite colors to learn more about nominal data and measures of central tendency in Figure 2.2, Favorite Color Data Set.

Figure 2.2 ▪ Favorite Color Data Set

Participant	What is your favorite color? (*X*)
1	5
2	3
3	5
4	3
5	2
6	4
7	4
8	4
9	1
10	5

Note: 1 = Blue, 2 = Black, 3 = Orange, 4 = Green, 5 = Yellow, 6 = Purple.

Let's use Figure 2.2, Favorite Color Data Set, to explore what measures of central tendency we can use with nominal data. We will do this by calculating the mode, median, and mean for this data set and then thinking critically about which can actually apply to this nominal data set.

Let's start exploring which measures of central tendency we can calculate with data on a nominal level of measurement by calculating the mode for Figure 2.2. Since one of the easiest ways to find the mode includes creating a frequency distribution table, let's go ahead and do this for our data set (consult Table 2.5).

Table 2.5 ▪ Favorite Color Ungrouped Frequency Distribution Table

Color	Frequency
Black	1
Blue	1
Green	3
Orange	2
Purple	0
Yellow	3

In the Table 2.5 Favorite Color Frequency Distribution Table, we notice that our data set is bimodal (i.e., our data set contains two modes). Our first mode is green and our second is yellow. In each case, 3 participants claimed it as their favorite. Now that we've found our modes for this data set, let's go back and determine whether the mode is a good representation of central tendency for our nominal data set. What do you think?

If you said that you believe the mode is a good representation of our data set, you are correct! The mode can precisely tell us which colors in our data set our participants endorsed as their favorites—in this case, green and yellow.

Now that we've calculated the mode of the data set, let's move on and investigate whether the median might be a good representation of data on a nominal level of measurement.

Step 1 Put the data in order from least to greatest.

1 2 3 3 4 4 4 5 5 5

Step 2 Determine whether there is an odd or even number of participants.

This data set had an even number of participants.

Step 3 Find the two center values in the data set.

1 2 3 3 **4** **4** 4 5 5 5

Step 4 For a sample with an even number of participants, find the mean of the two center values.

$$Mdn = \frac{4+4}{2} = \frac{8}{2} = 4.00$$

After we follow all the steps for finding the median of a sample with an even number of participants, we find that the median of the data set equals 4. Does 4 (green) represent the sample's favorite color?

Upon first impression, it seems that the median would partially represent the central tendency of this data set because green was one of the two colors most frequently selected by the participants in our sample. However, if we dig deeper, we find that it doesn't work.

The numbers in nominal data don't provide actual numerical value or order to them. Even though numbers represent an assigned color, the order of the numbers doesn't have meaning in explaining central tendency. If we would have assigned different numbers to each of the colors, the median potentially wouldn't represent the favorite color of the sample at all.

Let's examine this using a new pretend data set from different participants by finding the median for it (reference Figure 2.3 for the new data set).

Figure 2.3 ▪ Favorite Color Data Set From a New Sample

Participant	What is your favorite color? (*X*)
1	1
2	1
3	5
4	3
5	2
6	5
7	1
8	4
9	1

Note: 1 = Blue, 2 = Black, 3 = Orange, 4 = Green, 5 = Yellow, 6 = Purple.

Step 1 Put the data in order from least to greatest.

1 1 1 1 2 3 4 5 5

Step 2 Determine whether there is an odd or even number of participants.

This data set had an odd number of participants.

Step 3 For a sample with an odd number of participants, find the center value in the data set.

1 1 1 1 2 3 4 5 5

In this data set, we find that the median equals 2 (black), when it is very clear that 1 (blue) better represents the group's favorite color. Not only is the median not a good representation of our two favorite color data sets that we used as examples here, but we cannot use the median to represent *any* data set that uses a nominal level of measurement.

Now that we've attempted to calculate the mode and median for the data set, let's continue to explore the measures of central tendency for data on a nominal level of measurement by examining the sample mean of the data from Figure 2.2. To calculate the sample mean for this data set, we first need to find the sum of all the *X* values and then divide that number by the sample size.

$$\bar{X} = \frac{\sum X}{n} = \frac{5+3+5+3+2+4+4+4+1+5}{10} = \frac{36}{10} = 3.60$$

Now that we've calculated the mean, let's think about this for a little bit. If we look back at our data set, does 3.60 represent the favorite color of our sample? According to the data set's key, 3 = orange and 4 = green. If this was the best measure of central tendency for nominal data, then our sample's favorite color would be somewhere between orange and green. It really doesn't make a lot of sense to say that the class's favorite color fell between two unrelated colors. This is because when data falls on a nominal level of measurement, there is no numerical order to the groups. Because there is no order connected to nominal data, it doesn't make much sense to put them in order from least to greatest. This example shows us why for nominal data, the best (and only accurate) measure of central tendency will always be the mode. It also makes intuitive sense, because with nominal data, we assign the values to the categories arbitrarily.

Practice Makes Perfect

12. True or False. We can find the median for a data set of favorite dog breeds, where 1 = Plott hound, 2 = Siberian husky, 3 = pit bull, and 4 = golden retriever.
13. Which measures of central tendency can we calculate for a data set of the favorite seasons of the students in your statistics class, where 1 = winter, 2 = summer, 3 = fall, and 4 = spring?

Answers

12. False. We cannot find the median of nominal data.
13. The mode is the only measure of central tendency that we can do for data that falls on a nominal level of measurement like the favorite season of the students in your class.

Calculating Measures of Central Tendency for Data on an Ordinal Level of Measurement

In our last example, we examined which measure of central tendency was the best for nominal data. For this example, we will look at data on an ordinal level of measurement to determine which measures of central tendency we can use. For this example, imagine we need to rate the attractiveness of a group of people on a scale from 1 to 10, with 1 being the least attractive and 10 being the most attractive. We know that these ratings fall on an ordinal level of measurement, because unlike nominal data, there is an order in which 1 = least attractive and 10 = most attractive. However, it would be really hard for us to identify that there were equally spaced intervals between our ratings. We know that a rating of 1 is less attractive than a rating of 2, but we don't know that there is the same amount of attractiveness between the ratings of 1 and 2 as there is between ratings of 2 and 3. Plus different people find different characteristics attractive.

Let's meet the participants in this data set. After personality, a lot of people are attracted to people's smiles and eyes, so let's pretend these attractiveness ratings are based on these attributes, listed in Figure 2.4.

Figure 2.4 ▪ Attractiveness Ratings

iStock.com/martin-dm; iStock.com/hsyncoban; iStock.com/Martin Dimitrov; iStock.com/martin-dm; iStock.com/martin-dm

Using this pretend data set, let's calculate the three measures of central tendency that we learned about so far. First up, let's find the mode of our data set by putting the data into an ungrouped frequency distribution table (consult Table 2.6).

Table 2.6 ▪ How Attractive Are These People? Ungrouped Frequency Distribution Table

Attractiveness Rating (*X*)	Frequency
7	2
6	1
5	0
4	0
3	1
2	1

After examining our ungrouped frequency distribution table in Table 2.6, we find that our mode equals 7. This represents our data set because the mode simply tells us what rating was assigned to our participants most frequently. This shows us that we can calculate the mode with data on an ordinal level of measurement.

Now that we've explored calculating the mode, let's move on to calculating the median of this data set.

Step 1 Put the data in order from least to greatest.

2 3 6 7 7

Step 2 Determine whether there is an odd or even number of participants.

This data set had an odd number of participants.

Step 3 For a sample with an odd number of participants, find the center value in the data set.

2 3 6 7 7

After following the steps for finding the median of a sample with an odd number of participants, we find that the median equals 6. Does 6 represent the center of this data set?

It appears that the median does represent this data set that falls on an ordinal level of measurement. Even though the ordinal data doesn't have equally spaced intervals, it does provide us an order to our values, which is essential to finding the median.

Now that we've calculated both the mode and median for our ordinal data set, let's try to calculate the sample mean. We remember that to calculate the sample mean, we first add up all our *X* values together and then divide that sum by the number of participants in our data set.

$$\overline{X} = \frac{\sum X}{n} = \frac{2+6+7+3+7}{5} = \frac{25}{5} = 5.00$$

iStock.com/Liliboas

Now that we've found that the mean of this data set equals 5, let's determine if it is an accurate representation of our data set. To do this, ask yourself if an attractiveness rating of 5 best represents our sample.

Let's start with reviewing what we know about ordinal data. We know that it provides us an order to our data set, but we don't know how much space is between each of the rankings. However, when we add up all the data from our participants and then divide that sum by the sample size, we're assuming that we have equal intervals between our numbers. Because we don't know if we have equal intervals between our numbers in ordinal data, we cannot use the mean to determine central tendency for this data set.

Let's use a metaphor to further explain why we cannot find the mean with data on an ordinal level of measurement. Pretend we have a pie, but we don't know how big each of the pie slices is. Why couldn't we find the mean amount of pie people ate if we only knew how many slices of pie they consumed? We can't find the mean amount of pie people ate because we don't know if the pie slices are the same size. In the same way, to calculate the mean, there must be equally spaced intervals between the numbers.

For this reason, only data that fall on a scale (i.e., on an interval or ratio level of measurement) allow us to calculate the mean. Data on an interval or ratio level of measurement have equally spaced intervals between their numbers. To continue with the metaphor, if we knew that our slices of pie were the same size, then we could calculate the mean amount of pie that people ate. This lesson becomes super important as we move through the rest of the book because we will need to use the mean in most of the remaining formulas. To use formulas that include the mean, we depend on data that is derived from interval or ratio levels of measurement.

In summary, when our data set is based on an ordinal level of measurement, we can calculate both the mode and median for it. However, we cannot calculate the mean for data on an ordinal level of measurement because it doesn't guarantee equally spaced intervals.

Practice Makes Perfect

14. What measures of central tendency can we calculate for people's religious affiliation, where 1 = Agnosticism, 2 = Atheism, 3 = Buddhism, 4 = Christianity, 5 = Hinduism, 6 = Islam, 7 = Judaism, 8 = Wicca, and 9 = Other?

15. What measures of central tendency can we calculate for student grades where 0 = F (0% up to 59%), 1 = D (60% to 69%), 2 = C (70% to 79%), 3 = B (80% to 89%), and 4 = A (90% to 100%)?

Answers

14. We can only calculate the mode for this data set, because religious affiliation falls on a nominal level of measurement.

15. Here, we can calculate both the mode and median but not the mean, because this type of grading utilizes an ordinal level of measurement in which there are not necessarily equally spaced intervals between all letter grades.

Calculating Measures of Central Tendency for Data on a Scale With an Outlier in the Data Set

Now that we know we can calculate the mean for data on interval and ratio levels of measurement, let's learn more about how extreme scores influence each of the measures of central tendency. Sometimes, when we collect a data set, we notice that a participant provides data that is drastically different from that of the rest of the participants. We define an **outlier** as an extreme score within a data set. Later in this book, we will learn how to identify outliers in our data set numerically. But for now, let's explore how outliers influence which measures of central tendency we can calculate with scale data. To do this, let's use a pretend data set of the number of times participants reported going skydiving (find the data set in Table 2.7).

iStock.com/Mauricio Graiki

Table 2.7 ▪ Skydiving Data Set

Participant	How many times have you been skydiving? (*X*)
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	10

Let's start by creating an ungrouped frequency distribution table of our data set so that we can easily find the mode of the number of times our participants reported skydiving (consult Table 2.8).

Table 2.8 ▪ Skydiving Ungrouped Frequency Distribution Table

How many times have you been skydiving? (*X*)	Frequency
10	1
9	0
8	0
7	0
6	0
5	0
4	0
3	0
2	0
1	0
0	14

When we examine the ungrouped frequency distribution table, we find that the response that occurred most frequently in our data set was 0. This means that our mode equals 0. Looking back at our data set in Table 2.7, does it appear that the mode accurately represents the data set?

After examining our data set, it does appear that 0 accurately represents the responses from the participants in our sample. Most of our participants selected 0 as an option; in fact, only one participant selected something else. The mode is an appropriate measure of central tendency for scale data with an outlier in it because the mode ignores the outlier. It simply finds the most frequently chosen response.

Now that we've found the mode, let's move on to determine the median for this sample to examine if we can find the median for scale data with an outlier in the data set.

Step 1 Put the data in order from least to greatest.

0 0 0 0 0 0 0 0 0 0 0 0 0 0 10

Step 2 Determine whether there is an odd or even number of participants.

This data set had an odd number of participants.

Step 3 For a sample with an odd number of participants, find the center value in the data set.

0 0 0 0 0 0 0 0 0 0 0 0 0 0 10

Here, we find that the median of the skydiving data set equals 0. Why does the median appear to represent or not represent this data set?

When determining whether we can use the median to represent a data set that's on a scale with an outlier in it, let's think about what information scale data provides us. In scale data, we can put our data set in order because data on interval and ratio levels of measurement have order to them. What happens to the outlier when we order our data from least to greatest? The outlier is placed on the outer edge of the data set, which allows us to focus on what number represents the center of the data set better. This means that we can use the median as a measure of central tendency when representing a data set that is on an interval or ratio level of measurement with an outlier in it.

Now that we've calculated the mode and median for scale data with an outlier in the data set, let's explore whether we can calculate the mean for it too. Let's do this by calculating the sample mean for the skydiving data set.

$$\bar{X} = \frac{\sum X}{n} = \frac{(0+0+0+0+0+0+0+0+0+0+0+0+0+0+10)}{15} = \frac{10}{15} \approx 0.67$$

Here, we find that our mean equals 0.67. Does 0.67 accurately represent the center of the data set? It doesn't look like it represents the middle very well because everyone except one person indicated that they had never been skydiving. Let's explore this even further. What would have happened if we had an outlier in the data set that was even more outrageous (consult Table 2.9)?

Table 2.9 ▪ Skydiving Data Set With an Outrageous Outlier

ID	*X*
1	0
2	67
3	0
4	1
5	0

Let's calculate the mean with the new data set from Table 2.9.

$$\bar{X} = \frac{\sum X}{n} = \frac{(0+0+0+67+1)}{5} = \frac{68}{5} = 13.60$$

Does the mean appear to best represent this new data set? Not really, because 13.60 seems really high to represent the majority of our sample. But why did we get this outrageous number? If we look at the data set, we see that someone went skydiving 67 times! It was not at all typical for people in our sample to go skydiving that frequently, so the outlier of 67 heavily skewed the mean of our data set.

How much did that outlier actually impact the mean of our data set? Let's find out what the mean would have been if we didn't have that outlier in the data set. Let's calculate the sample mean if the data set would have just been 0, 0, 0, and 1 with no outlier of 67 in it.

$$\bar{X} = \frac{\sum X}{n} = \frac{(0+0+0+1)}{4} = \frac{1}{4} = 0.25$$

iStock.com/OGULCAN AKSOY

If we compare the mean of the data set that had the outlier in it ($\overline{X} = 13.40$, Table 2.9) to the mean when we removed the outlier from the same data set ($\overline{X} = 0.25$), we discover a considerable difference between them. We found that the outlier pulled the mean toward it. This shows us that the mean of data on an interval or ratio level of measurement fails to accurately represent the data set if there is an outlier in it. The mean is overly sensitive to outliers in a data set.

For this very reason, experts typically use the median to report annual household income even though it falls on a ratio level of measurement. This is because there are outliers such as the annual incomes of billionaires like Oprah Winfrey, Jeff Bezos, and Taylor Swift in comparison to the rest of the earners in the world. If we calculated the mean annual household income, the earnings of the billionaires of the world would disproportionately skew the mean. If we report the median annual household income instead of the mean, outliers, such as Bill Gates, don't overly influence the measure of central tendency we use to represent that data.

Practice Makes Perfect

Read the examples of data and identify which measure(s) of central tendency would be the best fit for that data set and explain why.

16. Favorite temperature in degrees Fahrenheit: 72 68 75 78
17. Number of tattoos: 36 1 2 0 5
18. Favorite city in the world: 1 1 5 4 2 3 3 1
Favorite city in the world, where 1 = Kansas City, 2 = Tokyo, 3 = Venice, 4 = Paris, 5 = Rio de Janeiro

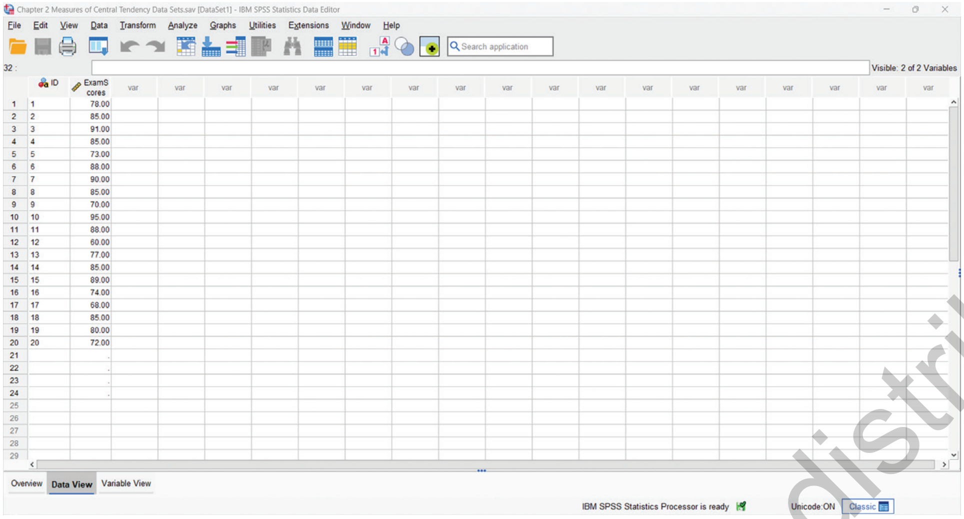
Answers

16. We could use the mean, median, or mode to represent this data set because it is on a scale and free of outliers.
17. We could use either the median or the mode but would not be able to use the mean because of the outlier of 36.
18. The only appropriate measure of central tendency for this data set is the mode because the data is on a nominal level of measurement.

SPSS Bonus Content Unlocked: Finding Measures of Central Tendency

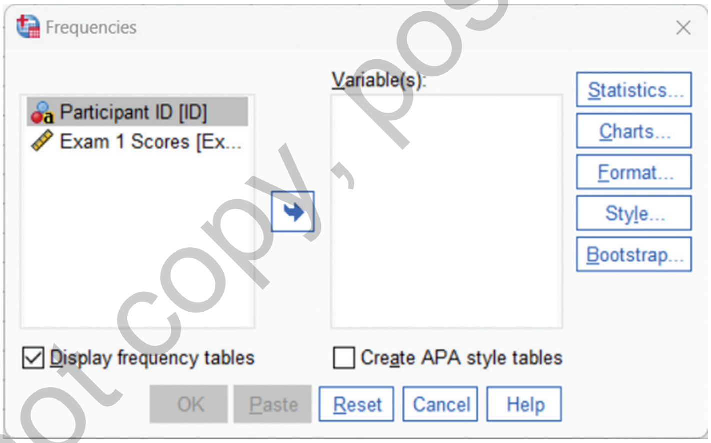
Now that we know when to use which measures of central tendency, let's use a practice data set to find the mean, median, and mode using SPSS. When we open up the "Chapter 2 Measures of Central Tendency Data Set," SPSS takes us directly to the "Data View" that we find in Figure 2.5.

Figure 2.5 ▪ Ch 2 SPSS Data View



To calculate these three measures of central tendency, we click "Analyze" >> "Descriptive Statistics" >> "Frequencies" and it'll bring us to the "Frequencies" dialog box we find in Figure 2.6.

Figure 2.6 ▪ Frequencies Dialog Box



After that, we click "Exam 1 Scores" to highlight it and click the blue right arrow to move the "Exam 1 Scores" into the "Variable(s):" box like we find in Figure 2.7.

Then we need to click "Statistics…" and endorse the "Mean," "Median," and "Mode" boxes underneath "Central Tendency" like we find in Figure 2.8 and then click "Continue."

After that we click "OK" in the "Frequencies" dialog box (refer to Figure 2.9).

Once we click "OK," SPSS will provide us an output displaying the mean, median, and mode (refer to Figure 2.10). The results show that the $\overline{X} = 80.90$, $Mdn = 85.00$, and $Mo = 85.00$.

Figure 2.7 ▪ Moving Variable Frequencies Dialog Box

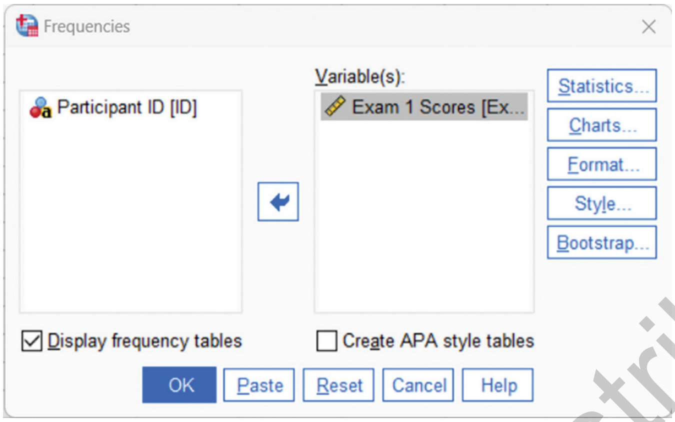


Figure 2.8 ▪ Measures of Central Tendency

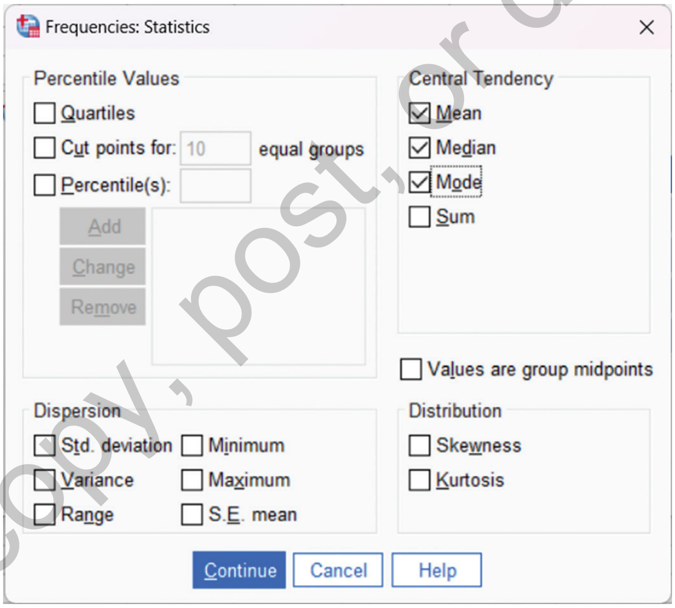


Figure 2.9 ▪ Click "OK"

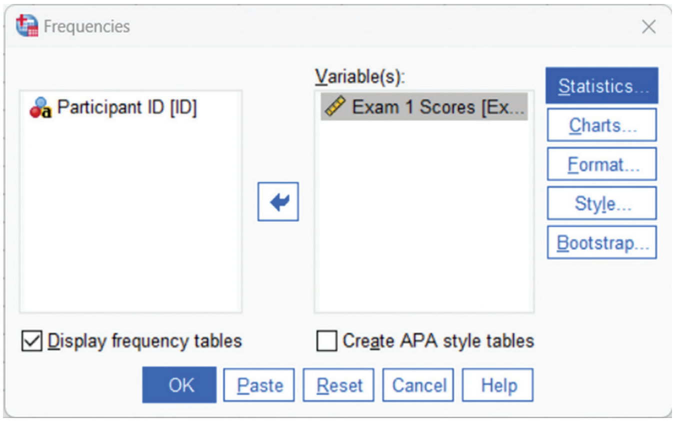
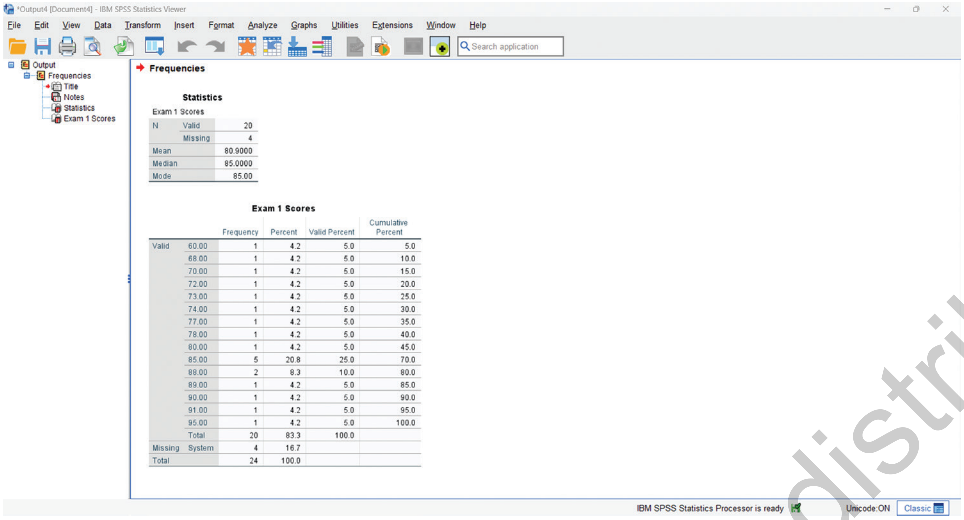


Figure 2.10 ▪ Measures of Central Tendency Output



Frequencies

Statistics

Exam 1 Scores

N	Valid	20
	Missing	4
Mean		80.9000
Median		85.0000
Mode		85.00

Exam 1 Scores

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	60.00	1	4.2	5.0	5.0
	68.00	1	4.2	5.0	10.0
	70.00	1	4.2	5.0	15.0
	72.00	1	4.2	5.0	20.0
	73.00	1	4.2	5.0	25.0
	74.00	1	4.2	5.0	30.0
	77.00	1	4.2	5.0	35.0
	78.00	1	4.2	5.0	40.0
	80.00	1	4.2	5.0	45.0
	85.00	5	20.8	25.0	70.0
	88.00	2	8.3	10.0	80.0
	89.00	1	4.2	5.0	85.0
	90.00	1	4.2	5.0	90.0
	91.00	1	4.2	5.0	95.0
	95.00	1	4.2	5.0	100.0
	Total	20	83.3	100.0	
Missing	System	4	16.7		
Total		24	100.0		

Summary

This chapter introduced descriptive statistics and its three major branches including measures of central tendency, variability, and distribution. We explored the key measures of central tendency, the mean, median, and mode, and how to think critically about which of these measures of central tendency can be best utilized based on the characteristics of a data set.

Symbol Guide

$\bar{X}$ sample mean (X-Bar)

Σ Summation symbol tells us to add everything up after it.

μ population mean (lowercase mu)

M mean (used for either sample or population mean)

Mo mode

Mdn median

Formulas

Formula 2.1 Sample Mean

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

Formula 2.2 Population Mean

$$\mu = \frac{\Sigma X}{N}$$

Terms to Know

Arithmetic mean
Bimodal
Frequency distribution table
Mean
Measures of central tendency
Measures of distribution
Measures of variability
Median
Mode
Multimodal
Outlier
Parameter
Statistic

Putting in the Work

1. Match the type of descriptive statistic with the correct explanation of its function.

• measures of distribution	These types of descriptive statistics try to describe the middle of the data set.
• measures of central tendency	These types of descriptive statistics provide information about how frequently every value occurred in a data set.
• measures of variability	These types of descriptive statistics try to describe how the individual data points differ from one another in the data set.

2. True or False. It would be considered a parameter if we found the mean number of times a sample of 10 movie theater customers went to the movies last year.
3. Which of the following is *not* a measure of central tendency?
   a. mean absolute deviation
   b. mode
   c. arithmetic mean
   d. median
4. Calculate the mean for the following data set: 56, 80, 60, 72, 43, 11, 22, 30
5. Calculate the mean for the following data set: 23, 11, 20, 19, 33, 20
6. Calculate the mode for the following data set: 15, 20, 12, 35, 12, 15, 20, 10, 12, 11, 10
7. Calculate the median for the following data set: 100, 999, 20, 300, 425, 427, 555, 638, 465
8. Calculate the median for the following data set: 5,234, 8,888, 2,345, 5,001, 6,550, 9,990
9. Calculate the mode for the following data set: 33, 35, 31, 32, 33, 31, 25, 33, 31, 34
10. If we asked 15 people what their favorite temperature was in degrees Celsius, what would be the most appropriate measure(s) of central tendency for that data set if there were no outliers?
    a. mean
    b. median
    c. median and mode
    d. mean, median, and mode

11. If we asked a group of 20 people what their favorite genre of music was, where 1 = pop, 2 = R&B, 3 = rock, and 4 = country, what would be the most appropriate measure(s) of central tendency for that data set if there were no outliers?
    - **a.** mean
    - **b.** mode
    - **c.** median and mode
    - **d.** mean, median, and mode

12. If we asked 17 people to tell us how many hours they watched Netflix last week, what would be the most appropriate measure(s) of central tendency for that data set if there was one outlier in it?
    - **a.** median
    - **b.** mode
    - **c.** median and mode
    - **d.** mean, median, and mode

13. If we asked 30 people to rate their pain level from 0 to 10, where 0 indicated no pain and 10 indicated the highest level of pain, what would be the most appropriate measure(s) of central tendency for that data set if there were no outliers?
    - **a.** mean
    - **b.** median
    - **c.** mode
    - **d.** median and mode

14. Which measure(s) of central tendency are the most influenced by outliers?
    - **a.** mean
    - **b.** median
    - **c.** mode
    - **d.** median and mode

Use the ratio data set in Table 2.10 to answer questions 15 through 17.

Table 2.10 ▪ Ratio Data Set

Participant	*X*
1	3
2	2
3	5
4	1
5	8
6	3
7	5
8	3

15. What is the *Mo* of the data set in Table 2.10?

16. What is the *Mdn* of the data set in Table 2.10?

17. What is the $\bar{X}$ of the data set in Table 2.10?

18. Match the symbol to the correct meaning.

Mdn	a. population mean
Mo	b. summation symbol
μ	c. sample mean
M	d. median
Σ	e. mean
X	f. mode
$\bar{X}$	g. participant data