Introduction

Using Statistics to Study the Social World

Why Study Statistics?

We all live in social situations. We observe our surroundings, are socialized into our cultures, navigate social norms, make political judgments and decisions, and participate in social institutions. Social sciences assume that what we can see as individuals is not the whole story of our social world. Political and social institutions and processes exist on a large scale that is difficult to see without systematic research. For most students in a social science statistics class, this basic insight is part of what drove your interest in this field. Maybe you want to understand political processes more thoroughly, understand how inequalities are produced, or understand the operation of the criminal justice system.

Many students reading this book are taking a statistics class because it is required for their major. Some readers are passionate about statistics, but most of you are probably mainly interested in sociology, political science, criminology, anthropology, education, or whatever your specific major is. Whatever your specific interest, statistics can deepen your understanding and build your toolkit for communicating social science insights to diverse audiences. You may think of statistics as a form of math, but, in fact, statistics are more about thinking with numbers than they are about computation. Although we do cover some simple computation in this book, our emphasis is on understanding the logic and application of statistics and interpreting their meaning for concrete topics in the social sciences. There is a good reason that statistics are required for many social science majors: Statistical methods can tell us a lot about the most interesting and important questions that social scientists study. Statistics also can tell you a lot about the questions that motivated your own interest in social sciences.
Statistics and quantitative data are important tools for understanding large-scale social and political processes and institutions as well as how these structures shape individual lives. They help us to comprehend trends and patterns that are too large for us to see in other ways. Statistics do this in three main ways. First, they help us simply to describe large-scale patterns. For example, what is the average income of residents in a given state? Second, statistics help us determine the factors that shape these patterns. This includes simple comparisons, such as how income varies by gender or by age. It also includes more complicated mathematical models that can show how multiple forces shape a given outcome. How do gender, age, race, and education interact to shape income, for example? Third, statistics help us understand how and whether we can generalize from data gathered from only some members of a group to draw conclusions about all members of that group. This aspect of statistics, called inferential statistics, uses ideas about probability to determine what kinds of generalizations we can make. It is what allows researchers to draw meaningful conclusions from data about relatively small numbers of people.

In this book, we emphasize what we can do with statistics, focusing on real social science research and analyzing real data. Readers of this book will develop a strong sense of how quantitative social scientists conduct their research and will get plenty of practice in analyzing social science data. Not all of this book’s readers will pursue careers as researchers, but many of you will have careers that include analyzing and presenting information. And, all of you face the task of making sense of mountains of information, including social science research findings, communicated by various media. This book provides essential tools for doing so.

Recently, some commentators have noted that we have entered a “post-fact,” or “post-truth,” era. People mean different things by this, but one meaning is that the sheer volume of people and agencies producing facts has multiplied to the point that an expert can be found to attest to the accuracy of just about any claim.1 Just think of the amount of information that you are exposed to on a weekly basis from various social media platforms, websites, television, and other forms of media. How do you make sense of it? How do you, for example, decide whether a claim you read online is true or false? Statistics can powerfully influence opinion because they use numerical data, which American culture assumes are objective and legitimate. But not all claims are equally factual, even those that appear to be backed up by statistics. This book will equip you with an understanding of how statistics work so that you can evaluate the meaning and credibility of statistical data for yourself.

When quantitative research is carefully conceived and conducted, the results of statistical analyses can yield valuable information not only about how the social world works but also about how to effectively address social problems. For example, in her 2007 book Marked, sociologist Devah Pager examined how having a criminal record affects men’s employment prospects in blue collar jobs.2 She conducted a study in which she hired paid research assistants, called testers, to submit fake résumés in person to potential employers. The résumés were the same, with the only difference...
being that some of them listed a parole officer as a reference, indicating that the applicant had spent time in prison, while the others did not have a parole officer as a reference. Did résumés without the parole officer reference fare better in the job search process? Yes, they did. On average former offenders were 46% less likely to receive a callback about the job, and the results of the analysis suggested that this difference could be generalized to the overall population of men applying for blue collar jobs, not just the testers in her study.3 Pager also varied the race of the testers applying for jobs—half were white, and half were black. She found that having the mark of a criminal record reduced the chances of a callback by 64% for black testers and 50% for white testers, indicating that the damage of a criminal record is particularly acute for black men.

By varying only whether the applicant had a criminal record, Pager controlled for alternative explanations of the negative effect of a criminal record on the likelihood of receiving a callback for a job. In other words, employers were reacting to the criminal record itself, not factors that might be associated with a criminal record, such as erratic work histories.

Pager’s study contains many of the key elements of statistical analysis that we discuss in this book: assessment of the relationship between two variables (criminal record and employer callbacks); a careful investigation of whether one of the variables (criminal record) has a causal impact on the other (employer callback) and, if so, whether that causal impact varies by another factor (race); and examination of the generalizability of the results.

---

**Research Questions and the Research Process**

Most research starts with a **research question**, which asks how two or more variables are related. A **variable** is any characteristic that has more than one category or value. In the social sciences, we must be able to answer our research questions using data. In many cases, these questions may be fairly general. For example, sociologist Kristen Luker writes about beginning a research project with a question about why women were having abortions despite the availability of birth control.4 A criminologist may begin by wanting to know what kinds of rehabilitation programs reduce recidivism. In other cases, a question may expand on prior research. For example, research has shown that Internet skills vary by class, race, and age.5 Do these factors affect the way Internet users blog or contribute to Wikipedia? Or, if we know that children tend to generally share their parents’ political viewpoints, does this hold true in votes for candidates in primaries?

Some research begins with a **hypothesis**, a specific prediction about how variables are related. For example, a researcher studying political protest might hypothesize that larger protests produce more news media coverage. Other research begins at a more exploratory level. For example, the same researcher might collect data on several possible variables about protests, such as the issue they focus on, the organizations
that sponsor them, whether they include violence, as well as their size, in order to explore what shapes media coverage. Statistical methods can support both approaches to research.

This book focuses on **quantitative analysis**—that is, analyses that use statistical techniques to analyze numerical data. Many social scientists also use qualitative methods. **Qualitative methods** start with data that are not numerical, such as the text of documents, interviews, or field observations. Qualitative data analysis often focuses on meanings, processes, and interactions; like quantitative research, it may test hypotheses or be more exploratory in nature. Qualitative research analysis often uses specialized software programs. Increasingly, many researchers use **mixed methods**, which employ both qualitative and quantitative data and analysis. While this book focuses on quantitative analysis, combining both methods can yield a richer and more accurate understanding of social phenomena than either approach alone.

### Pinning Things Down: Variables and Measurement

Answering any kind of social science research question entails gathering data. Gathering useful data requires formulating the research question as precisely as possible. Quantitative researchers first identify and define the question’s key concepts. **Concepts** are the abstract factors or ideas, not always directly observable, that the researcher wants to study. Many concepts have multiple dimensions. For example, a researcher interested in how people’s social class affects their sense of well-being must define what social class and well-being mean before examining whether they are related. Using existing research and theory, the researcher might define a social class as a segment of the population with similar levels of financial, social, and cultural resources. She might decide that well-being is one’s sense of overall health, satisfaction, and comfort in life. Stating clear definitions of concepts ensures that the researcher and her audience understand what is meant by those concepts in the particular project at hand.

Once researchers specify, or define, their concepts, they must decide how to **measure** these concepts. Deciding how to measure a concept is also referred to as **operationalizing** a concept, or **operationalization**. Operationalization, the process of transforming concepts into variables, determines how the researcher will observe concepts using empirical data. Staying with the example of social class and well-being, how would we place people into different class categories? Using the conceptual definition described above, the researcher might decide to use people’s income, wealth, highest level of education, and occupation to measure their social class. All of these are empirical indicators of financial, social, and cultural resources. To operationalize well-being, the researcher might decide to measure an array of behaviors (e.g., number of times per week that one exercises) and attitudes (e.g., overall sense of satisfaction with one’s life).
This process of conceptualization and measurement, or operationalization, is how concepts become variables in quantitative research. Figure 1.1 offers a visual representation of this process for the concept of well-being.

Figure 1.1 shows how researchers move from defining a key concept to specifying how that concept will be empirically measured and transformed into variables. Starting from the top of the figure and moving down, we can see how the process works. First, the concept of well-being is defined. Next, the dimensions of the concept (physical, mental, and spiritual) are specified. Finally, the researcher establishes empirical measures for each dimension (e.g., frequency of exercise as an indicator of physical well-being). These empirical measures are called variables. The arrow on the right side of Figure 1.1 shows how moving from defining concepts to measuring them shifts from the theoretical or abstract to the empirical realm, where variables can be measured. Studying relationships among variables is the central focus of quantitative social science research.

A variable, remember, is any single factor that has more than one category or value. For example, gender is a variable with multiple categories (e.g., man, woman, gender non-binary, etc.). For some variables, such as body mass index, there is an established standard for determining the value of the variable for different individuals (e.g., body mass index is equal to weight divided by height squared). For variables that lack a clear measurement standard, such as sense of purpose in life, researchers must establish their categories and methods of measurement, usually guided by existing research.

In quantitative social science research, the survey item is among the most common tools used to operationalize concepts. Survey items have either closed- or open-ended response options. Closed-ended survey items provide survey respondents with
predefined response categories. The number of categories can range from as little as two (e.g., yes or no) to very many (e.g., a feeling thermometer that asks respondents to rate their feeling about something on a scale from 0 to 100 degrees). With closed-ended survey items, the researcher decides on the measurement of the concept before administering the survey. Open-ended survey items do not provide response categories. For example, an item might ask respondents to name the issue that is most important to them in casting a vote for a candidate. Open-ended items give respondents more leeway in answering questions. Once the researcher has all responses to an open-ended item, the researcher often devises response categories informed by the responses themselves and then assigns respondents to those categories based on their responses. For example, with an open-ended question about which issues are important to voters, the researcher might combine various responses having to do with jobs or the economy into one category.

Units of Analysis

In the social sciences, researchers are interested in studying the characteristics of individuals but also the characteristics of groups. Who or what is being studied is the unit of analysis. A study of people’s voting patterns and political party affiliation focuses on understanding individuals. But a study of counties that voted for a Republican vs. Democratic candidate focuses on understanding characteristics of a group, in this case counties. In the first case, researchers might seek to understand what explains people’s votes; in the second case, researchers might seek to understand what characteristics are associated with Republican vs. Democratic counties. When the unit of measurement is the group, we sometimes also refer to it as aggregate level. Aggregate-level units that researchers might be interested in include geographic areas, organizations, religious congregations, families, sports teams, musical groups, or businesses. One must be careful about making inferences across different levels of measurement. A county may be Republican, but at the individual level, there are both Democratic and Republican residents of that county. Drawing conclusions about individuals based on the groups to which they belong is an error in logic known as the ecological fallacy.

Measurement Error: Validity and Reliability

Most variables in the social sciences include some amount of error, which means that the values recorded for a variable are to some degree inaccurate. Even many variables that one might suspect would be simple to measure accurately, such as income, contain error. How much money did you receive as income in the last calendar year? Some readers may know the exact figure. But others would have to offer an estimate, maybe because they cannot recall or because they worked multiple jobs and have trouble keeping
track of the income produced by each of them. Still others might purposefully report a number that is higher or lower than their actual income. Researchers never know for sure how much error their variables contain, but we can evaluate and minimize error in measurement by assessing the validity and reliability of our variables.

**Validity** indicates the extent to which variables actually measure what they claim to measure. When measures have a high degree of validity, this means that there is a strong connection between the measurement of a concept and its conceptual definition. In other words, valid measures are accurate indicators of the underlying concept. Imagine a researcher who claims that he has found that happiness declines as people exercise more. How is that researcher measuring happiness? It turns out that he has operationalized happiness through responses to two survey questions: “How much energy do you feel you have?” and “How much do you look forward to participating in family activities?” Do you think answers to these questions are good measures of happiness? They may get at elements of happiness—happier people may have more energy or look forward to participating in activities more. But they are not direct measures of happiness, and we could argue that they measure other things instead (such as how busy people are or their health). What about a researcher who wants to measure the prevalence of food insecurity, in which people do not have consistent access to sufficient food? This could be operationalized in a survey question such as, “How often do you have insufficient food for yourself and your family” or “How often do you go hungry because of inability to get sufficient food for yourself or your family?” It could also be operationalized by the number and size of food pantries per capita or food stamp usage. Which way of operationalizing food insecurity is more accurate? The survey questions have greater validity because both food pantries and food stamp usage are affected by forces other than food insecurity (urban areas may have more food pantries per capita than rural areas, not all people eligible for food stamps use them, and so forth). If the researcher were interested instead in social services to reduce food insecurity, looking at food pantries and food stamps would be a valid measure.

Even if a measure is valid, it may not yield consistent answers. This is the question of **reliability**. Reliable measures are those whose values are unaffected by the measurement process or the measurement instrument itself (e.g., the survey). Imagine asking the same group of college students to rate how often in a typical week they spend time with friends, with the following response choices: “often,” “a few times,” “occasionally,” and “rarely.” These response choices are likely to lead to problems with reliability, because they are not precise. A student who gets together with friends about five times a week might choose “often” or “a few times,” and if you asked her the question again a week later she might choose the other option, even if her underlying estimate of how often she spent time with friends was unchanged. In other words, the same students may give quite different, or inconsistent, responses if asked the question repeatedly.

Measures also tend not to be reliable when they ask questions that respondents may not have detailed understanding or information about. For example, a survey might ask how many minutes a week people spend doing housework, or a survey of Americans...
might ask their opinion of Britain’s foreign policy toward Chile. Because people do not generally precisely track minutes spent doing housework, and Americans are unlikely to know much about British foreign policy, their responses to such questions will be inconsistent.

Reliability and validity do not necessarily coincide. For example, the time shown on a clock may be reliable without being valid. Some households may deliberately set their clocks to be a few minutes fast, ensuring that when the alarm goes off at what the clock says is 6:45, the actual time is 6:30. In this case, the clock consistently—that is, reliably—tells time, but that time is always wrong (or invalid).

Figure 1.2 uses a feeling thermometer, which asks people to rate their feeling about something on a scale from 0 to 100 degrees, to illustrate how reliability and validity can coincide or not. Imagine these are an individual’s responses to the same feeling thermometer item asked five separate times. The true value of the person’s feeling is 42 degrees. In scenario A, the responses have a high degree of validity, or accuracy, because they are all near 42 degrees, the accurate value. There is also a high degree of reliability because the responses are consistent. Researchers strive to attain scenario A by obtaining accurate and consistent measures. In scenario B, there is still a high degree of consistency, and therefore reliability, in the measure. However, validity is low because the responses are far from the true value of 42 degrees. Finally, scenario C reflects both low reliability and low validity. The responses are inconsistent, or scattered across the

![Figure 1.2 Visualizing Reliability and Validity](image-url)
range of the temperature scale, and many fall far from 42 degrees. Notice that there is no scenario D, in which reliability is low and validity is high. This is because the overall accuracy of a measure requires that it be reliably measured.

Levels of Measurement

There is another consideration about how to measure variables—whether they will be measured in a way that will yield data that are numerical. This is very important for statistical analysis because it determines what statistics and graphics can be employed, as we will explain below. Consider a variable measuring employment status. A survey question could ask respondents how many hours they worked in the preceding week. The answers would all be numbers, such as 35 hours, 12 hours, and so forth. Alternatively, a survey question could ask whether respondents are employed full-time, part-time, or not at all. The answers to this question are not numbers, although they can be placed in rank order, since those who are employed full-time are working more than those who are employed part-time. Variables also can be measured in ways that are neither numerical nor rankable. For example, a question about employment might ask what type of job the respondents hold and provide response categories such as “officials and managers,” “professionals,” “technicians,” “sales,” “clerical,” “skilled trades,” and so forth.* These answers are categories, but they do not have any quantitative meaning because none of them can be considered to have a greater value than others.

A variable’s level of measurement refers to whether the “answers,” or possible values of the variable, are numerical; rankable but not numerical; or categorical. Variables with values that are numerical, or quantitative, are called interval or ratio level. For these variables, the distance between each consecutive value of the variable is identical. For example, in the variable number of hours worked, the distance between 20 hours and 21 hours (1 hour) is the same as the distance between 21 hours and 22 hours and between any other adjacent values. Ratio-level variables have a meaningful 0 value that represents a true value of 0 for the variable being measured (such as 0 hours of work or 0 dollars). Interval-level variables do not have a true 0 value. For example, temperature is an interval-level variable because a value of 0 on any temperature scale does not mean the “absence” of temperature. For our purposes, interval- and ratio-level variables are treated in the same way, and we will refer to them as “interval-ratio” variables. Examples of interval-ratio variables include scores on exams, hours or minutes spent on any activity (e.g., hours spent watching television or doing housework), number of times participating in an activity (e.g., number of times per month attending religious services or exercising), number of sexual partners, family members, or children, and many more. Interval-ratio variables can be continuous

---

* All federal agencies in the United States use the Standard Occupational Classification system, which classifies all workers into 867 detailed occupations. A full list of these occupations can be found in the 2018 Standard Occupational Classification Manual at https://www.bls.gov/soc/2018/soc_2018_manual.pdf.
or **discrete**. **Discrete variables** are measured in whole numbers and cannot be broken down further. For example, number of children is a discrete variable because the values of that variable (the number of children) only can be whole numbers. One cannot have 2.5 children. **Continuous variables** have values that can be continually subdivided. Savings measured in dollars, length of employment measured in years, and length of commute measured in miles are all examples of continuous variables. Although we may round these variables (to dollars, days, or half miles), in theory these units can be subdivided further and further.

Variables with values that can be rank-ordered, but which are not numerical and where the distance between each value of the variable is not identical, are **ordinal** level. For example, in the variable employment status, “full-time” represents a greater amount of employment than “part-time,” but the difference between the two categories cannot be expressed in a specific numerical amount. Social science variables that are ordinal level also include questions in which the response categories are not equal in size. For example, when measuring frequency of exercise, a variable could include response categories such as “daily,” “several times a week,” “weekly,” “two or three times a month,” and “monthly or less.” While these categories can clearly be ranked in order of frequency, the difference between exercising daily and exercising several times a week (or between any other two categories) is not numerically precise. Other examples include variables like “How happy are you?” or “How satisfied are you with your job?” that have response categories like “very,” “somewhat,” “little,” or “not at all.”

Finally, variables that are not numerical and cannot be rank-ordered are **nominal** level. The response categories for nominal-level variables are simply categories, without any quantitative meaning. As a result, nominal variables are sometimes also called “categorical” variables. Many variables that social scientists use are nominal level. These are variables such as race, gender, religious affiliation, region of residence, marital status, occupation, or political party affiliation. For example, if the categories of political party affiliation are “Democrat,” “Republican,” “Independent,” and “other,” we cannot rank these categories; they are simply names for the different affiliations.

There is one more important piece of information about levels of measurement. There are many variables in social science research that are scales ranging from “strongly agree” to “strongly disagree.” They are often questions about opinions. These are ordinal variables, since the distance between each pair of categories is not numerically precise. However, in practice, researchers generally treat them as interval-ratio level if they have at least five categories. That means, for example, that a researcher might calculate an average for such a variable, saying, for example, that “On a scale of 1 to 10, average support for measures to reduce climate change was 8.2.”

Why does a variable’s level of measurement matter? It determines what kind of statistical calculations can be performed. Many statistics can be calculated only for

* “Dollars” is technically a discrete variable because its units cannot be subdivided below one cent. However, when dealing with large quantities (e.g., hundreds or thousands), dollars can be treated as a continuous variable.
interval-ratio variables. Consider the mean, or average. You may know that calculating an average requires adding up the values of the variable for all the cases and then dividing by the total number of cases. But you only can add values that are actually numbers, such as hours spent online. You can’t add values for nominal variables. (How would you add “Protestant” + “Catholic,” for example?) You also can’t add values for ordinal variables. (How would you add “Very much” + “Somewhat”?) We will cover this in much more detail in the chapters that follow. For now, remember that determining the level of measurement of a variable is the first important task in statistical analysis.

Causation: Independent and Dependent Variables

A major purpose of statistics in the social sciences is to study relationships among variables. Many social scientists are interested in studying a specific kind of relationship: causal relationships. In a causal relationship, one variable, called the independent variable, causes changes in another variable, called the dependent variable. For example, a criminologist might be interested in studying the effects of rehabilitation programs offered in prison (such as job training) on recidivism, the likelihood of being re-arrested. Does participation in such programs have a causal impact on the likelihood of reoffending?

As we will see in chapter 13, determining whether one variable causes changes in another is no simple task. One might observe, for example, that former offenders who participated in rehabilitation programs have an overall lower rate of recidivism than do those who did not participate in those programs. But to establish that this relationship is causal—that it is the programs themselves that actually deter former offenders from reoffending—the researcher must rule out alternative explanations. For example, it could be that rehabilitative programs are more likely to exist in states that also have higher expenditures on social service programs. The researcher would hold constant or “control” for this third variable—state expenditures on social service programs—to see if the relationship between rehabilitative programs and reoffending were still present. If there were no longer a relationship after holding constant state expenditures on social service programs, this could indicate that lower recidivism rates among those who participate in rehabilitative programs are caused not by the programs but by higher spending on social service programs in general, which also happens to be correlated with the number of rehabilitative programs that states offer.

There are two basic ways of controlling for alternative causal explanations. Researchers using experimental research designs employ experimental control by randomly assigning research participants to treatment and control groups to ensure that participants in one group are not systematically different from those in the other group. Participants in the treatment group receive the “treatment” (e.g., participate in
a rehabilitative program), while those in the control group do not. We would assume that any difference in the outcome (i.e., the dependent variable) between the groups was caused by the treatment because of the random assignment of participants to the two groups. Because experimental designs are often impractical, most social scientists must employ the other method of ruling out alternative explanations: statistical control. Statistical control is employed in a variety of ways in the data analysis process to ensure that a third variable does not account for the relationship between the independent and dependent variables.

**Getting the Data: Sampling and Generalizing**

During presidential election campaigns, we are inundated with surveys about the candidates’ relative standing. These surveys are meant to give us a sense of who is ahead, who is behind, and by how much. For example, on November 1, 2016, one week before the presidential election, an *ABC News/Washington Post* poll reported that 46% of likely voters expressed support for Donald Trump, compared to 45% for Hillary Clinton. But for obvious reasons, this poll, and every other poll, interviewed a relatively small number of people—it was based on interviews with a sample of 1,128 people. If truth be told, we would not be all that interested in the views of these 1,128 people if they were not representative of the full population of U.S. voters. But they were. Each person in the sample was randomly selected to participate in the survey. This random selection gives us a high degree of confidence that our sample results—Trump 46%, Clinton 45%—are close to what we would have obtained had we somehow managed to interview all 139 million voters.

Inferring from a small sample to a larger population is one of the central goals of statistics. A population includes every individual or case in a category of interest, such as voters. A sample is made up of a small group of individuals or cases drawn from the larger population of interest. If a researcher wishes to generalize from a sample to the population, then that sample must be randomly selected from the population. Most of the time, it is not practical to study all the members of a population directly—unless that population is relatively small and well-defined. For example, we could imagine drawing up a full list of every county in the United States, every country in the world, or every student at your school in order to study them directly. When we are able to study all members of a population, we use a variety of statistical tools to describe variables and their relationships within this population. There is no need to make inferences about the population because we have actual, direct data about the full population. But most of the time, this is not possible. Instead, researchers draw random samples out of populations in order to make inferences about the population based on the characteristics of the sample. Chapters 2–5 focus on descriptive statistics, statistical techniques for describing the patterns found in a set of data, whether those data are based on a full population or a sample. In chapters 6–14, we focus on the idea of “inference” and
the various statistics researchers employ to determine whether and how the results they find in a sample can be generalized. (Chapter 14 also covers some descriptive statistics for examining relationships between variables.) Statistics that examine whether information from a sample can be generalized to a population are called **inferential statistics**.

The ability to infer from a sample to a population is based on the idea of randomness. Randomness is at the core of “probability samples.” In a **probability sample**, every member of the population must have an equal probability of being selected for the sample, and the selection of cases from the population must be made randomly. Most election polls reported by the media employ probability samples. On the other hand, you may have come across Internet polls or call-in polls on the local news. These are **non-probability samples**. In such instances, members of the sample are self-selected, they are not drawn randomly, and most of the time there are biases associated with who chooses to participate and who doesn’t. Although the results of such polls may be interesting, they tell us nothing about a larger population beyond those who responded and are therefore of little to no value.

**Sampling Methods**

There are a variety of methods for drawing a probability sample that allow for inference to a larger population. The most basic method is known as **simple random sampling**. Here, we make a list of all the members of a population and randomly draw our desired number of cases from that population into the sample. We must be able to make a full list of all the members of the population so that we can randomly draw from that list. The list that we draw our sample from is called a **sampling frame**. For example, we could list all 2,600 students enrolled at Smith College, the school where the authors of this book teach, and then randomly draw a sample of 200 of them. Mechanically, these are the steps we might follow to draw this sample:

1. Obtain a list of all 2,600 students at Smith College.
2. Assign every Smith College student a number between 1 and 2,600.
3. Use a random number generator to select 200 numbers between 1 and 2,600.
4. Match each selected number with the student assigned to that number.

We would now have a randomly selected sample of 200 Smith College students.

Because simple random samples require a list of every member of the population, they are practical to use only with fairly small and well-defined populations, such as the students at a small school or all the counties in the state of California. On the other hand, large or constantly changing populations should not be sampled using this method. For example, it would not be possible to list the names of all 139 million voters in the United States.
Stratified random sampling is a variation of simple random sampling. A **stratified random sample** allows the researcher to randomly sample from subgroups in a population to ensure that the sample is representative of population subgroups that are of interest to the researcher, such as students from different class years or residents of rural and urban counties.

Assembling a sampling frame can be harder than it sounds. Sometimes, lists of all members of a population are available through, for example, records of students enrolled at a school, voter registration rolls, telephone directories, or lists of mailing addresses. But these lists are not always publicly available, and the lists themselves can have errors. Sometimes random samples are drawn by randomly dialed telephone numbers (through a computer program that begins with area codes and the three-digit prefix associated with that area code and then randomly selects the final four digits of a phone number). Of course, not everyone has a telephone; cell phone numbers are not listed in directories; and some numbers produced by randomly generated digits will not be working numbers, and others will be assigned to businesses. For paper or face-to-face surveys, researchers can purchase address lists for many areas from the U.S. Postal Service. 

Nevertheless, for large populations, these procedures are cumbersome.

There are methods of probability sampling that do not require a full listing of the target population. The most common is **cluster sampling**, where we randomly sample clusters of cases instead of individuals and then randomly sample individuals from within these clusters. For example, we might not be able to put together a complete list of individuals in a large metropolitan area, but we can assemble a full list of census tracts or city blocks. A cluster sample might start by the researcher putting together a complete list of city blocks, randomly selecting a number of them, assembling a list of households on those city blocks, randomly selecting a number of those households, and then randomly selecting one individual from each household. This method is sometimes called **multistage cluster sampling**. Its main advantage is that it allows the researcher to put together a random sample of individuals from a large population without a complete list of individuals in that population.

Even proper probability sampling techniques can yield a sample that is not representative of a population of interest. This is because of **nonresponse bias**, which occurs when individuals who are invited to take a survey vary systematically in the likelihood that they will complete the survey (or particular survey items). For example, if a survey begins with a question about citizenship status, undocumented immigrants may be less likely to respond to the survey than citizens. Or if a survey is administered during the day, it may be more difficult to reach people who are at work. In these cases, the sample data would not be generalizable to the population because one group of intended respondents was much less likely to answer the survey than others and is, therefore, underrepresented in the sample.

Regardless of the sampling method employed, it is important not to lose sight of our central objectives. We use samples because they shed light on a larger population.
When we study samples, we generate statistics that help us describe characteristics of the sample. We use these statistics to make educated guesses about the value of the unknown population characteristic in which we are interested. For example, we measure the percentage of our sample who state they will support Candidate A because that tells us approximately how much support Candidate A has in the population. We measure the average income in a sample because that tells us approximately what the average population income is. A lot of what we do in the chapters that follow is based on this simple notion: We use statistics to describe a sample and then to infer from that sample to the population.

Sources of Secondary Data: Existing Data Sets, Reports, and “Big Data”

In addition to collecting their own data to address research questions, social scientists often use secondary data, or data that have been collected previously, usually by someone else and often for a purpose that might differ from an individual researcher’s. In these cases, the researcher is usually not involved in the sampling process, but it is still very important that a researcher understand the sampling strategies used to collect any source of secondary data. If the goal of a study is to yield results that can be generalized to a population, only secondary data collected through probability sampling is appropriate.

Fortunately, there are many sources of high-quality secondary data available to social scientists that are collected with generalizability as a primary goal. These data sources are usually the product of large-scale surveys conducted by university researchers with support from various private and public agencies. Most secondary data sets follow a general theme (e.g., political beliefs) yet still ask questions about a wide enough range of topics that researchers can use the data to address a variety of research questions.

Throughout this book, we work with a number of publicly available secondary data sets, all collected using probability sampling. Many of these data sets are available for download on the book’s website, including the following:

1. General Social Survey (GSS)
2. American National Election Study (ANES)
3. World Values Survey (WVS)
4. Police Public Contact Survey (PPCS)
5. The National Longitudinal Survey of Youth (NLSY)\(^8\)

These data sets allow us to address a range of interesting social science topics. The WVS is a cross-national survey with probability samples of nearly 100,000 respondents from sixty countries. The rest of the data sets employ probability samples
of respondents from the United States. The unit of analysis for the GSS, ANES, WVS, and PPCS is the individual. These surveys ask individuals about a range of topics such as their social backgrounds, financial resources, activities, families, opinions, and political beliefs.

Along with the data sets themselves, users can download the codebooks for the data sets. Codebooks are so named because they provide the “code” necessary for interpreting the meaning of each variable. When a data set is created, variables are given names, and numbers are assigned to the categories of the variables. Codebooks contain the following essential information about the variables in a data set:

- the name and description of each variable
- descriptions of each category of every variable
- the numerical value assigned to each category of every variable

Figure 1.3 shows an excerpt from the PPCS codebook, for a variable called V81.

<table>
<thead>
<tr>
<th>Value</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>After 6 a.m. – 12 noon</td>
</tr>
<tr>
<td>02</td>
<td>After 12 noon – 6 p.m.</td>
</tr>
<tr>
<td>03</td>
<td>Don’t know what time of day</td>
</tr>
<tr>
<td>04</td>
<td>After 6 p.m. – 12 midnight</td>
</tr>
<tr>
<td>05</td>
<td>After 12 midnight – 6 a.m.</td>
</tr>
<tr>
<td>06</td>
<td>Don’t know what time of night</td>
</tr>
<tr>
<td>07</td>
<td>Don’t know whether day or night</td>
</tr>
<tr>
<td>98</td>
<td>Refused</td>
</tr>
<tr>
<td>–9 (M)</td>
<td>Out of universe/missing</td>
</tr>
</tbody>
</table>

The codebook tells us that the variable called V81 measures what time of day the respondent’s most recent contact with a police officer occurred. It also tells us that this variable has eight categories: (1) between 6 a.m. and noon, (2) between noon and 6 p.m., (3) don’t know what time of day, (4) between 6 p.m. and midnight, (5) between midnight and 6 a.m. (6) don’t know what time of night, (7) don’t know whether day or night, and (98) refused. The last category listed, –9, represents missing data. Notice that the numbers assigned to each category are only labels for the categories and are not meaningful as numbers. Category 1 does not mean that the respondent had contact with a police officer at 1:00, for example; it means that the contact occurred between 6 a.m. and noon. When researchers use secondary data, they can decide
whether to use the original code for any given variable or recode the variable in some other way. For example, a researcher might use \( V81 \) to create a new variable that measures whether the respondent had contact with the police officer during the day, evening, or night.

**Big Data**

By now, most people have heard the term “big data,” but what does it mean, and how is it related to statistics? There is a key distinction between “big data” and data collected through traditional survey methods. Whereas traditional survey methods collect data for a specific purpose, big data—or organic data—emerge as a by-product of the electronic tracking of people’s behavior online and in the real world. Big data emanate from various sources, such as administrative information (e.g., electronic medical records), social media, and records of online searches. One way of thinking about big data is to imagine individuals’ actions, and especially their online actions, as leaving an invisible residue, or digital trace. This residue constantly adds to the ever-growing store of big data. Big data are collected by corporations (tracking purchasing and search information, for example), by technology companies such as Google and Facebook, and by other entities. Some big data are proprietary, owned and accessible only by those who collect them, but many big data records can be obtained by independent researchers.

Whereas in survey research, researchers determine the questions and their possible answers by constructing variables and their response categories, big data directly reflect people’s actions without categories imposed by a researcher. As sociologist Amir Goldberg notes, with big data, the approach to data analysis is more open-ended. Big data researchers are less likely to approach their analyses with preformulated hypotheses and more likely to “let the data speak,” opening up possibilities for finding unanticipated patterns in the data. For example, a team of researchers in Wisconsin used linked administrative records from social service agencies in the state to study patterns of disconnection from sources of public assistance for those who are in need of them. One of the key findings is that the traditional notion of what it means for a family to be “disconnected” from public financial assistance—when a family is eligible and in need of financial assistance but no longer receives it—misses a number of other classes of “disconnection” uncovered in the data, such as families who receive food assistance through the Supplemental Nutrition Assistance Program (SNAP) but not financial assistance. If the researchers had relied on a predetermined measure of disconnection, as survey research might have, they would have missed these other ways of thinking about disconnection.

But where big data enthusiasts see possibility, critics argue that its push toward more open-ended approaches to data analysis—letting the data speak—will pull the social sciences away from building theoretically informed explanations for social phenomena and toward simplistic descriptions of social behaviors and attitudes. For example, danah boyd and Kate Crawford point out that cell phone data might show that cell phone users have more social media and text communications with their work
colleagues than with their spouses. Without applying the theoretical tools of the social sciences, we might conclude that coworkers are more important to people than are their spouses. However, it is more likely that text and social media communications reflect what sociologists call “weak ties” but are poor indicators of “strong ties,” or close interpersonal relationships marked by emotional connection. 11

Big data also must grapple with the same considerations about sampling frame, the list of all members of the population, that researchers using probability samples must consider. Namely, is the sampling frame biased? Does it actually contain all members of the population of interest? As many observers have noted, big data from social network sites, such as Twitter and Facebook, represent biased sampling frames because social background and demographic characteristics, such as race and age, are related to whether people use social media sites. 12 Thus, inferences about the general population should not be drawn from big data derived from social media.

One final major concern about big data is ethical and privacy implications. All research involving human subjects must ensure that the safety and privacy of the research participants will not be compromised by participating in the study. Researchers must ensure that all participants give their informed consent to participate in the study. Because big data are made up of the digital traces people leave behind, it is impossible for researchers to obtain the consent of the people whose behaviors left the traces. In addition, for some sources of big data, anonymity cannot always be maintained. For example, using data from credit card transactions for 1.1 million users that did not contain identifiable information (i.e., no names or account numbers), researchers were able to “reidentify” many of the 1.1 million users using limited pieces of information available in the data, such as the price of the transaction. 13

In sum, big data offer new and exciting possibilities for researchers interested in social behavior. There is no question that research using big data will contribute mightily to social science. However, there remains an important place for traditional statistical methods in the social sciences. The findings from research using traditional, theoretically informed statistical methods can provide the context necessary for making sense of the findings yielded by big data.

Growth Mindset and Math Anxiety

“I’m not a math person.” At some point, you likely have heard someone utter this statement, or maybe you have said it yourself. Underneath this statement lies a potentially harmful view of math and one’s relationship to it. In general, this statement communicates a view of one’s mathematical capabilities as fixed and impervious to growth. Saying that one is not a math person also can indicate some level of anxiety about the material itself, perhaps tied to previous difficulties with math. In this section, we discuss how adopting a growth mindset can help all students do better in statistics. For those who have some level of anxiety about studying a subject that does utilize
math, we show how a growth mindset can be a particularly valuable ingredient for success in statistics.

Researcher Carol Dweck has written extensively about the benefits of what she calls a growth mindset approach to learning. As opposed to a fixed mindset, which views intelligence as a fixed and essential characteristic of individuals, a growth mindset views intelligence as something that develops over time through hard work and effort. Research in neuroscience has demonstrated the human brain’s ability to become smarter in response to targeted effort, indicating that the human brain works much more like the vision of the growth mindset than the fixed mindset.

So when we hear that someone is not a math person, we know that neuroscience tells us otherwise. To be sure, individuals differ in their intellectual interests and talents, but most people’s intellectual skills can improve through effort and engagement. In fact, a number of experiments have shown that students who are explicitly taught to adopt the view that intelligence is not fixed, but develops through work and effort, experience greater gains in mathematics learning than control groups. In other words, evidence suggests that adopting a growth mindset when it comes to statistics can go a long way toward actually helping people to do well in statistics. Believing that competence can improve in an area, such as statistics, is just one element of a growth mindset. The other element, equally important, is understanding that this competence is the outcome of applied effort.

Sometimes, adopting a growth mindset when it comes to learning statistics may not be enough to overcome math anxiety, which can be described as “an adverse emotional reaction to math or the prospect of doing math.” With about 17% of the U.S. population having math anxiety, this is no small issue. Fortunately, when it comes to the study of statistics, and particularly the approach taken by this book, there are ways to combat the potentially disruptive effects of math anxiety on learning statistics.

The first way to lessen the effect of math anxiety on your performance in your statistics course is to recognize that, while statistics does depend on basic math skills, most statistics courses taught from a social science perspective draw more upon verbal and inductive reasoning than math skills themselves. The focus of this book is much more on statistical reasoning than the math underlying the statistics. Thus, even students who have some level of anxiety about math can be reassured that this book presents statistics as a tool for understanding social phenomena, requiring students to draw upon only basic math skills.

For students who still have some anxiety about studying statistics stemming from anxiety about their math abilities, research suggests a simple way to counteract that anxiety. A team of psychologists asked college students with high and low levels of math anxiety to complete a math test. They wondered if completing an expressive writing task, in which students were asked to write for 7 minutes “as openly as possible about [their] thoughts and feelings regarding the math problems [they were] about to perform,” would lead to smaller differences in performance on the test between students with high and low levels of math anxiety. In fact, there was a dramatically smaller gap in performance between high- and low-anxiety students in the expressive
writing task group than in the control group in which students were simply given the test. Take a moment to reflect on this: The math performance of math-anxious students improved dramatically when they wrote openly about their math anxieties without any effort to improve their math abilities.

These results suggest that the threat of math anxiety is not primarily a tale of those with high anxiety having worse math skills. As the researchers speculate, it is likely much more a story of how math anxiety distracts one’s cognitive abilities from the task at hand. This study measured the positive effects of expressive writing on performance on a brief math test, but it is plausible to think that there may be positive effects of acknowledging one’s math anxiety on one’s performance in a statistics course. It is worth trying an expressive writing exercise similar to the one in the experiment, in which you openly express your thoughts and feelings about the material in your statistics course.

To recap, our recommendations for counteracting the negative effects of math anxiety on statistics performance include, first, adopting a growth mindset when it comes to mastery of statistics and, second, openly acknowledging one’s math anxiety regularly throughout the course. This advice suggests neither that math anxiety can be easily eradicated nor that it should be completely eradicated. In fact, a frequently replicated empirical finding indicates that both high and low levels of anxiety in a given domain can hurt performance in that domain. The finding has been replicated so many times that the phenomenon has a name: the Yerkes-Dodson Law. Using a sample of students from a university’s Introduction to Statistics course, researchers found that the Yerkes-Dodson Law applied to students’ statistics performance. Students with very high and low levels of statistics anxiety performed worse than students who reported a medium level of anxiety. This research suggests that there is an optimal level of anxiety that motivates students to seek to improve, as a growth mindset would call upon students to do, but does not monopolize students’ cognitive resources in a damaging way.

Using This Book

This book is designed to be used with a growth mindset approach to statistics. This means that we encourage readers to use the book as a tool to help them actively develop and sharpen their understanding of statistics. As with most kinds of knowledge, developing statistical knowledge is not a linear process. Just when you think you understand something, you might find that you’re confused about the concept all over again. This is quite typical with statistics, and you are not alone. Even seasoned researchers can benefit from returning to core statistical concepts to refresh their memories. This means that you should expect to work with and return to various concepts throughout the book many times.

Throughout the book, we offer readers a number of ways to develop and practice their skills and check their understanding of the material. First, each chapter includes
rich examples of how to use statistical tools to answer interesting social science ques-
tions. In addition to reading these examples, we encourage you to work through each of the examples on your own. Second, the chapters include two kinds of boxes separate from the main text. “In Depth” boxes go into more depth about a topic that is covered in the main text, and “Application” boxes walk readers through additional examples employing the relevant statistical method. We know that the temptation is strong to focus on the main text in the name of efficiency, but these boxes provide readers the opportunity to practice and dig deeper into some of the key topics covered in the text. Third, each chapter ends with sections on using two common statistical software programs, Stata and SPSS. (You will use the section for the program that your class is using.) These sections provide an opportunity to apply the tools learned in a given chapter to real social science data using statistical software. After reading the “Using Stata” or “Using SPSS” section, you can complete the exercises on your own or in a lab associated with your statistics class. For an extra challenge, we encourage you to apply the same techniques used in the software sections to variables that are not used in the examples. Fourth, each chapter includes a set of practice problems, designed to help you check your understanding of the material and provide opportunities to challenge yourself. Finally, the companion website to the textbook contains many resources for students to practice their skills. We believe that “statistical intelligence” is something that everyone can obtain through work and effort, and we encourage readers to use the book to challenge themselves to develop this intelligence.

Statistical Software

Statistical software programs can analyze patterns in data sets that include large numbers of cases. Throughout the book, as we explain statistical techniques we often show you how to calculate a result by hand, but these calculations are very time-consuming when data sets are large. Almost all statistical research now relies on computers to do calculations. Statistical software programs ease the computational burden on the user and allow for the analysis of data sets that are too large for the human brain to analyze in a reasonable amount of time.

The first statistical software program was developed in 1957, and since then scientists have developed many more programs. Today, analysts are faced with a dizzying array of these programs, ranging from those designed for general use to those designed for the use of highly specialized statistics.

In this book, we will use Stata and SPSS, two programs that enjoy wide popularity among social scientists. Most students will be using only one of these programs,
Before the technology existed for the electronic storage and analysis of large data sets, data were stored on small “punched cards.” Machines punched small circles or rectangles at specific locations on the card to indicate the value for a specific variable, with the presence or absence of a hole indicating the case’s value for that variable. The U.S. Census Bureau commissioned the inventor Herman Hollerith to develop this “punched card” technology to aid in the collection and analysis of information about the U.S. population. Figure 1.4 shows an image of a census worker punching a card for the 1920 Census.22

Depending on what is available on your campus. You should read only the section of each chapter pertaining to the program you are using in your class. These sections give you the opportunity to use Stata or SPSS to find answers to interesting social science questions using real social science data. At the end of this chapter, we present a general introduction to each program.
This chapter covered the key parts of the process of conducting social science research with quantitative data that precede data analysis. We also discussed available sources of quantitative data and how best to approach learning statistics from a social science perspective. Below, we review key terms.

- The research process proceeds in four major steps:
  1. A social science research question asks how two or more variables are related and must be able to be answered using data.
  2. Defining concepts and their dimensions. Concepts are the abstract factors or ideas that the researcher wants to study. Concepts may have multiple dimensions.
  3. Measurement or operationalization is the process of transforming concepts into observable data, or variables. It includes specifying the dimensions of each concept and establishing the variables that are empirical measures of each dimension. Operationalization determines how the researcher will observe concepts using empirical data.
  4. Sampling is the process of choosing cases from the population to study.

- A hypothesis is a specific prediction about how variables are related. Research questions may specify hypotheses or be more exploratory.

- Quantitative analysis uses statistical techniques to analyze numerical data.

- Qualitative methods start with data that are not numerical, such as the text of documents, interviews, or field observations. Qualitative data analysis often focuses on meanings, processes, and interactions; like quantitative research, it may test hypotheses or be more exploratory in nature.

- Mixed methods employ both qualitative and quantitative data and analysis.

- An independent variable is the cause of changes in another variable.

- A dependent variable is affected by another variable.

- Descriptive statistics are statistical techniques for describing the patterns found in a set of data.

- Statistical control controls for alternative causal explanations by using statistical techniques.

- Key terms involving variables and measurement:
  - A variable is any characteristic that has more than one category or value.
  - Level of measurement refers to whether variables are nominal, ordinal, or interval-ratio. It determines what statistical techniques can be applied to variables.
  - Ratio-level variables have numerical values, with identical distances between each value, and a meaningful 0 value that represents a true value of 0 for the variable being measured.
  - Interval-level variables have numerical values, with identical distances between each value, and no true 0 value.
• **Interval-ratio** refers to both interval- and ratio-level variables.
• **Ordinal-level variables** have non-numerical values that can be rank-ordered; the distance between each value of the variable is not identical.
• **Nominal-level variables** are not numerical and cannot be rank-ordered.
• **Scales**, such as those ranging from “strongly agree” to “strongly disagree,” are ordinal-level variables that can be treated as interval-ratio in practice if they have at least five categories.
• **Discrete** variables are measured in whole numbers and cannot be broken down further.
• **Continuous** variables have values that can be continually subdivided.
• **Validity** is the extent to which variables actually measure what they claim to measure; accurate responses.
• **Reliability** is the extent to which responses are consistent and unaffected by the measurement process.
• **Closed-ended** survey items provide survey respondents with predefined response categories.
• **Open-ended** survey items do not provide response categories; respondents supply their own answers.

• Key terms involving sampling and generalizing:
• The **unit of analysis** is the object of study, either individuals or groups.
• When the unit of measurement is the group, we sometimes also refer to it as **aggregate level**.
• The **ecological fallacy** is drawing conclusions about individuals based on groups.
• A **sample** is a small group of cases drawn from a larger population of interest.
• A **population** is every case in the group of interest.
• A **sampling frame** is a complete list of all members of a population.
• A **probability sample** is drawn randomly from a population in which every member has an equal probability of being selected for the sample. Simple random samples, stratified random samples, and cluster samples are all probability samples.
• A **simple random sample** is a random selection of members of a population from a complete list of all members.
• A **stratified random sample** is derived from random samples of subgroups of interest in a population.
• A **cluster sample** is a multistage sample where the researcher randomly samples clusters of cases instead of individuals, followed by randomly sampling individual cases from the sampled clusters. It allows the researcher to put together a random sample of individuals from a large population without having a complete list of individuals in that population.
• A **non-probability sample** is one in which members of the sample are self-selected or they are not selected randomly.
• **Inferential statistics** examine whether information from a sample can be generalized to a population. They can be used with probability samples only.
• Sources of data not collected directly by the researcher:
  - **Secondary data** refers to data that are analyzed but not collected by the researcher.
  - **Big data** are data not collected through traditional sampling methods; the by-product of the electronic tracking of people’s behavior online and in the real world.
• How to use this book:
  - Adopt a **growth mindset**, the view that ability increases over time through hard work and effort and is not a fixed characteristic of individuals.
  - Address math anxiety by openly expressing your thoughts and feelings about doing math in writing.
  - Practice working the problems in the book and on the companion website.

---

**Stata** is one of the most commonly used statistical software programs. With Stata, users have a choice between “point-and-click” and command driven use. With the point-and-click method, the analyst uses Stata’s graphical user interface (GUI), or series of drop-down menus, to select the actions for Stata to perform on the data. With the command approach, analysts use code to write their own commands that tell Stata which analyses to perform. While both methods will yield the same results if done properly, we focus on the command driven approach to Stata in this book. Using the GUI is a perfectly legitimate way to use Stata, but most Stata users prefer to write their own code. This section was developed using Stata for Windows Version 15. If you are using a different version of the software, you may find slight variations in the commands, but, mostly, they will be very similar.

**How Stata Looks**

There are two ways to launch Stata. You can launch the program just as you would launch any program (such as your word processing software) on your computer. You then click on “File” in the upper-left corner, choose “Open,” and navigate to the location of the saved data file. Alternatively, if you have a saved Stata data set on your computer, you can click on that. This will automatically open Stata, and you can begin analyzing this data set. Stata data files always end with the extension “.dta” (e.g., `filename.dta`).
Here, we will use a small example data set called `ExampleStataData.dta`, with information about fictitious college students after their first year of college. Once you open the data set, Figure 1.5 shows how Stata will appear.

![Figure 1.5](image)

The large white space in the middle of the screen is the Results Window, where Stata will display the results of the statistical analyses that you tell it to conduct. The Command Box, circled at the bottom of the screen, is the place where you can type in your syntax commands. The long, narrow Command Window on the left side of the screen is the place where your commands are archived. Each time you enter a command into the Command Box, Stata automatically also places that command in the Command Window on the left. You can copy a command from the Command Window back into the Command Box by clicking on it at any time. (Note that Stata does not save this archive of commands. If you want to save a record of your commands, you will need to enter your code in a separate file, which we discuss below.) On the right side of the screen, circled, is the Variables Window, which lists all of the variables in the data set. We see that in this example data set, there are six variables.

We can view the data by clicking on the Data Editor icon at the top of the screen, circled. (Note that there are two versions of this icon, one that allows the user to edit the data and one that does not allow editing. We are using the icon that allows only for browsing the data, not editing.) You can also open the Data Editor by selecting “Data” in the menu in the top taskbar and then selecting “Data Editor (Browse)”. Once you click on the Data Editor icon, you will see the image shown in Figure 1.6, which displays the data in this data set.

In this data set, the unit of analysis is college student, so each row reports information for a unique college student in the data set. We can see that this data set is small,
including only ten college students. Each column shows the values for a unique variable, in this case six of them. (The actual data sets that we will use in this book have many more cases and variables.) The six variables in this data set include:

- gpa (students’ GPA after the first year of college)
- orgs (the number of student organizations that students have joined)
- sat (students’ satisfaction with their social life on campus, ranked from 1 to 10)
- friends (the number of close friends students have on campus)
- live (whether the student lives on or off campus; 1 = on campus; 0 = off campus)
- commute (the daily commute time for students living off campus, in minutes)

We can see that the first student in the data set had a GPA of 2.8, belonged to five student organizations, rated campus social life as a 9, had twelve close friends on campus, and lived on campus. Why do we see a period in the cell for the commute variable for this student? The period means that there is no value reported for this variable for this student. We refer to missing information on a variable as “missing data.” We could observe missing data for a variable for a number of reasons. For example, a respondent may have refused to respond to that particular question, a researcher may have failed to record the respondent’s value, or that variable may not be applicable to that particular case in the data set. In this case, Student 1 has no value for the commute variable because the student did not live off campus. Only students who lived off campus were asked to report their commute time to campus.

Once you have reviewed the data in the Data Editor, you can close it, and Stata will return you to the initial interface that you saw when you first opened the program. When we walk you through the analyses in the “Using Stata” sections at the end of each chapter, you will type those commands into the Command Box at the bottom of the window. Once you have finished typing a command, press enter, and Stata will display the results of the analysis in the Results Window. The only time that the result
will not be displayed in the Results Window is when you ask Stata to generate a graph or chart, which will always open in a separate window. You will be impressed to find that Stata can produce results, even with massive data sets, in fractions of a second. No matter how quickly you can do computations by hand, Stata will beat you every time!

**Basic Logic of Stata Code**

For most people, the most intimidating part of learning a new statistics software program is learning how the code works. If you have studied any language, you know that every language has its own grammar that lays out its structural rules. As a language, Stata code is no different. All Stata code commands follow the same basic structure:

```
command variable name(s), options
```

Commands can be customized in various ways, as we will see in examples throughout the book, but the basic rule is to state the specific analysis that you want Stata to conduct (the “command”) followed by the variable(s) on which it should conduct this analysis, followed by any special options for that analysis.

**Three Categories of Commands: Create New Variables, Transform Existing Variables, and Analyze Existing Variables**

There are three basic kinds of commands that we will use in this book: (1) those that create new variables, (2) those that transform existing variables, and (3) those that conduct statistical analyses on variables that already exist in the data set. We offer simple versions of each category of command here as a preview. You can try them using the example data set, available on the website.

1. **Create a New Variable:** Here is an example of a command that creates a new variable in Stata:

   ```stata
   generate commute2 = commute/60
   ```

   The “generate” command asks Stata to create a new variable called `commute2`. For each student in the example data set, the value of this new variable is equal to the value of the `commute` variable divided by 60. Because the `commute` variable is measured in minutes, dividing the variable by 60 converts the variable into hours. After running this command, if you click on the Data Editor, you can see that the value for `commute2` is missing if there was also missing information for `commute`. For the three cases that had values for `commute`, the values for `commute2` are now measured in hours. For example, for Student 3 the value of `commute` is 30, and the value of `commute2` is .5.

   We could have chosen any name for the new `commute2` variable as long as it conformed to the rules for Stata variable names. Variable names in Stata can be up to thirty-two characters, including capital or lowercase letters, numbers, and...
underscores. When you create new variables in Stata, we recommend keeping the variable name to around eight characters. Variable names that are longer than twelve characters will be truncated in most Stata output. Variable names are case-sensitive, so be careful about specifying capital or lowercase letters. We recommend sticking to lowercase letters to simplify matters.

2. Transform an Existing Variable: Here is a series of commands that transforms an existing variable, in this case `orgs`, which measures the number of student organizations that students have joined:

   ```
   generate orgs2 = orgs
   replace orgs2 = 1 if orgs>=1
   replace orgs2 = 0 if orgs==0
   ```

   This series of commands follows a convention that we adopt throughout the book: Whenever transforming an existing variable, create a duplicate version of that variable first, and transform only the duplicate version. This preserves the original variable and allows you to check that the transformation proceeded properly by comparing values of the new variable to the original one.

   With these three commands, we create a new variable, called `orgs2`, which indicates whether students joined any organization during their first year of college. In the new variable, all students who joined one or more organizations are combined into one category. The first “generate” command creates the new `orgs2` variable. By writing “= orgs,” we are telling Stata to set the values for the new variable equal to the values for the existing `orgs` variable (e.g., all cases will have the same value for `orgs2` that they have for `orgs`). The second two “replace” commands change, or recode, the values of the `orgs2` variable. The first “replace” command assigns `orgs2` a value of 1 if the student’s value for `orgs` is greater than or equal to 1. The second “replace” command assigns a value of 0 to `orgs2` if the value for `orgs` for that student is equal to 0. (Note that this command is redundant because `orgs2` was already equal to 0 if the student’s value for `orgs` was 0, but we show the command for consistency.) Also notice that we have used two equals signs after “if” in the second “replace” command. This is not an error. It is a rule in Stata syntax that you must use two equals signs after “if” in a command. As we see in the “replace” commands, if an equals sign precedes “if” (or if there is no “if” used in the command), then we use only one equals sign.

   It is essential to check your work every time you transform an existing variable. Even seasoned data analysts make coding mistakes, and one small error in a line of code can have profound consequences for the results of a statistical analysis. Remember to check your work, and check it again! As long as you do not violate any of Stata’s syntax, it will run your command, even if the code contains an error for your purposes. While Stata’s brain may be much faster than the human brain, in this way the human brain is smarter. In this case, we
can check our work by making sure that the number of cases with a value of 1 for \textit{orgs2} is equal to the number of cases with value of 1 or higher for \textit{orgs}. (Since we are dealing with a small data set, we can check this manually in the Data Editor.)

3. \textbf{Analyze an Existing Variable:} Here is an example of a command that asks for a simple analysis of an existing variable:

\texttt{summarize friends}

We will see this “summarize” command many times throughout the book, as it is one of the most commonly used commands in Stata. Here, we have asked Stata to summarize the \textit{friends} variable. The output from this command, as shown in Stata’s Results Window, is shown in Figure 1.7.

\begin{verbatim}
. summarize friends

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>friends</td>
<td>10</td>
<td>3.9</td>
<td>3.725289</td>
<td>0</td>
<td>12</td>
</tr>
</tbody>
</table>
\end{verbatim}

\textbf{Figure 1.7}

In response to the “summarize” command, Stata produces a table reporting the following information about the \textit{friends} variable, from left to right: the number of observations for which there are available data (10), the mean (3.9), the standard deviation (3.7), the minimum value (0), and the maximum value (12). (We will cover all of these statistics in detail in chapters 4 and 5.)

As you start to become comfortable with Stata, you might discover that different commands can often be used to generate the same result. If you find yourself developing your own ideas for how you might ask Stata to accomplish a task that vary from the examples that we offer in the book, this is a good sign that you are internalizing the language of Stata.

\textbf{Error Messages in Stata}

It is not uncommon for Stata to return errors instead of output in the Results Window. Whenever you make a mistake that violates one of Stata’s syntax rules, you will see an error message. To show a simple example, if we had misspelled the word “summarize” in the command that we used above as “sumarize,” we would have received the following error message:

\begin{verbatim}
. sumarize friends
command sumarize is unrecognized
r(199);
\end{verbatim}
All error messages give a brief explanation for the error (in red) and provide a link to the error code (in blue). Here, Stata is telling us that the command “sumarize” is not recognized by the program. That is simply because we mistyped the command. If you click on “r(199),” Stata will take you to a brief explanation of the error code, in this case 199. The description of the error code tells us that there is probably a typographical error in our syntax.

This simple error message illustrates the important role that error messages can play in learning Stata syntax. If you read the error message carefully and take the time to click on the error code, you will often identify the source of the problem in your syntax. If you can resist the urge to panic when you see an error message, which can be strong for new Stata users, then error messages can become one of your most valuable Stata teachers.

**Operators in Stata Syntax**

As you become familiar with Stata syntax, it will be important for you to know the basic operators in Stata syntax. For example, as we saw in one of the above commands, if we want to divide the values of a variable by something, we do not write out the word “divide”; we use the “/” sign. The commonly used arithmetic, relational, and logical operators in Stata are shown in Table 1.1.

<table>
<thead>
<tr>
<th>Arithmetic Operators</th>
<th>Relational Operators</th>
<th>Logical Operators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addition: +</td>
<td>Greater than: &gt;</td>
<td>And: &amp;</td>
</tr>
<tr>
<td>Subtraction: –</td>
<td>Less than: &lt;</td>
<td>Or:</td>
</tr>
<tr>
<td>Multiplication: *</td>
<td>Greater than or equal to: &gt;=</td>
<td>Not: ! (or) ~</td>
</tr>
<tr>
<td>Division: /</td>
<td>Less than or equal to: &lt;=</td>
<td>Through: /</td>
</tr>
<tr>
<td>Raise to a power: ^</td>
<td>Equal to: = =</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Not equal to: ! (or) =~</td>
<td></td>
</tr>
</tbody>
</table>

**Data Files, Do-Files, and Saving Your Work**

As we work with Stata throughout the book, we will use a number of data files, which contain the actual data that you will ask Stata to analyze. As we noted above, the common feature of all Stata data files is that they end in the extension “.dta.”

Do-files are another kind of Stata file, and they end in the extension “.do”. Do-files are where you can write and save commands for future use. While you need not use a do-file in order to follow along with any of the Stata exercises in this book (all commands can be typed directly into the Command Box), we recommend starting a do-file if you are working on projects outside of the exercises in this book. To open a do-file, simply click on the do-file icon in Stata, circled in Figure 1.8.
Instead of typing single commands into the Command Box, you can collect all of your code in one savable do-file. To use the do-file, simply write your commands directly in the file, with each command on a separate line. (You can also copy and paste...
commands that you have run from the Command Window into a do-file.) You can either run one command at a time, as you would in the Command Window, or select chunks of syntax to run all at once. The easiest way to do this is to highlight the command(s) that you want to run, select the “Tools” menu, and then “Execute selection.” Once you have run the commands, you can toggle to the Results Window to review the output. You will see that the commands you just ran now also appear in the Command Window. If you want to make comments to yourself in the do-file, say, to leave comments about what a set of commands is meant to accomplish, simply precede those notes with one or more asterisks (*), which signals to Stata that the subsequent text is a comment and not commands.

Saving: It is important to recognize that data files and do-files are separate files, saved independently of one another. Saving a data file does not mean that you have saved the do-file that you have been using, and vice versa.

Sources of Stata Help

The Stata User’s Guide, Stata Base Reference Manual, and Stata help files provide a useful source of extra help and support as you learn Stata. The most recent versions of the Stata User’s Guide and Base Reference Manual can be downloaded for free at http://www.stata.com/bookstore/users-guide/ and http://www.stata.com/bookstore/base-reference-manual/, respectively. The Stata User’s Guide gives an overview of how Stata works, and the Base Reference Manual offers detailed descriptions and examples of all Stata commands. Chapter 3 of the User’s Guide lists a number of sources of information about Stata, including the Stata website, the Stata YouTube channel, and the Stata blog. Typing “help” followed by any Stata command (e.g., “summarize”) into the Command Window will open the help file associated with that command, which describes how to use the commands and provides links to the place in the User’s Guide where that command is described.

SPSS is a statistical software program that enjoys wide popularity among social scientists. SPSS can be used in two modes: “point-and-click” and command driven use. With the point-and-click method, the analyst uses SPSS’s graphical user interface (GUI), or series of drop-down menus, to select the statistical analyses for SPSS to perform on the data. With the command approach, analysts use a special language called syntax to write their own commands that tell SPSS which analyses to perform.24 While both methods will yield the same results if done properly, in this book, we focus on the point-and-click approach to SPSS, which is the more widely used method. The SPSS exercises that follow (here and in subsequent chapters) were developed using SPSS Version 25 for the Mac. If you are using a different version of SPSS, there may be very slight differences in the appearance of commands and procedures.

How SPSS Looks

There are two ways to launch SPSS. You can launch the program just as you would launch any program (such as your word processing software) on your computer. You
then navigate to “File” ➔ “Open Data” and select the SPSS data set you wish to open. Alternatively, if you have a saved SPSS data set on your computer, you can click on that. This will automatically open SPSS, and you can begin analyzing this data set. SPSS data files always end with the extension “.sav” (e.g., filename.sav).

Regardless of how you launch the software, SPSS operates with two windows, and you will navigate back and forth between them. If you launch the program by clicking on the SPSS icon, you will see the SPSS “Data Editor” window, but it will be empty (since you have not yet indicated what data set you are using). If you launch SPSS by clicking on the saved SPSS data file, then it will open the “Data Editor” window, which will be populated with information about the variables in the data set and the actual data. When you launch SPSS, it will also open an “output” window at the same time it opens the “Data Editor” window. The output window is where the results of your statistical analyses will appear, but it will be empty when you first open the program.

For illustration purposes, we will use a small SPSS data set, called ExampleSPSSData.sav, with information about fictitious college students after their first year of college. You can try the procedures we show here, using the example data set, available on the website. We launch SPSS by clicking on this file. This opens the SPSS “Data Editor” window, shown in Figure 1.10.
The “Data Editor” window has two options, “Data View” and “Variable View,” circled at the bottom of Figure 1.10. When the “Data View” window is active, as shown in Figure 1.10, we see the data. Each row corresponds to a case in our data set. In this data set, the unit of analysis is the student, so each row reports information for a unique college student in the data set. We can see that the data set is small, with data for only ten college students. Each column shows the values of a different variable. The names of the variables are shown at the top of each column. The six variables in this data set are:

- **gpa** (students’ GPA after the first year of college)
- **orgs** (the number of student organizations that students have joined)
- **sat** (students’ satisfaction with their social life on campus, ranked from 1 to 10)
- **friends** (the number of close friends students have on campus)
- **live** (whether the student lives on or off campus; 1 = on campus; 0 = off campus)
- **commute** (the daily commute time for students living off campus)

In the illustration above, we can see that Respondent 1 has a GPA of 2.80, belonged to five student organizations, rated campus social life as a 9, had twelve close friends on campus, and lived on campus. If you look at the commute variable on the right, you will note that we have data for only three of the ten respondents; for seven respondents, the data are missing. SPSS puts in a period to represent data that are missing. We could observe missing data for a variable for a number of reasons. For example, a respondent may have refused to respond to that particular question, a researcher may have failed to record the respondent’s value, or that variable may not be applicable to that particular case in the data set. In this case, Respondent 1 has no value for the commute variable because the student did not live off campus. Only students who lived off campus were asked to report their commute time to campus.

There is an alternative way of examining the information in this data set. Instead of “Data View,” we can select “Variable View” by clicking on “Variable View” at the bottom of the window (circled in Figure 1.10). Instead of seeing the raw data for each respondent, we see summary information about the variables. The “Variable View” window is shown in Figure 1.11.

SPSS now shows us the variable name, the type of variable (numeric or text), the width of the variable as it is recorded in the data file, the number of decimal places, the variable’s label, and the labels associated with each value of the variable. We can also use this window to add or edit certain information about the variables. We will return to this point shortly.

Before you begin to analyze your data, we always recommend getting a sense of your data set by looking at both the “Data View” and the “Variable View” windows. Once you begin analyzing data, SPSS will deliver the output in a different window, called the “Output” window. One can use the SPSS drop-down menus when either window—“Data Editor” or “Output”—is active. Any time you ask SPSS to run a procedure, it will append the output to the bottom of the “Output” window. You will
be impressed to find that SPSS can produce results, even with massive data sets, in fractions of a second. No matter how quickly you can do computations in your head, SPSS will beat you every time!

Three Categories of Commands: Create New Variables, Transform Existing Variables, and Analyze Existing Variables

There are three basic kinds of commands that we will use in this book: (1) those that create new variables, (2) those that transform existing variables, and (3) those that conduct statistical analyses on variables that already exist in the data set. We offer simple versions of each category of command here as a preview.

1. Creating a New Variable: One of the variables in the `ExampleSPSSDATA.sav` data set is called `commute`. It represents the daily commute time for students living off campus, measured in minutes. We are going to tell SPSS to create a new variable called `commute2` by taking each respondent’s score on the original `commute` variable and dividing it by 60. Thus, the new variable, `commute2`, will represent respondents’ commute time, measured in hours.

   To create a new variable, we use the SPSS drop-down menu called “Transform”; we click on it and then click on “Compute Variable,” as shown in Figure 1.12.

   **Transform → Compute Variable**
Figure 1.12

This sequence opens up the “Compute Variable” dialog box, as shown in Figure 1.13.

Figure 1.13
We will type the new variable’s name, commute2, into the “Target Variable” box, and we will define the construction of the new variable by putting the appropriate information (in this case, “commute/60”) into the “Numeric Expression” box. Once completed, the “Compute Variable” dialog box will look as shown in Figure 1.14.

![Compute Variable dialog box](image)

**Figure 1.14**

After we click “OK,” the new variable, commute2, will appear in the “Data Editor” window under “Variable View.” SPSS always appends new variables to the bottom of the list, as shown in Figure 1.15.

You can see that the value for commute2 is also missing if there was missing information for commute. For the three cases that had values for commute, the values for commute2 are now measured in hours. For example, for Student 3 the value of commute is 30, and the value of commute2 is .5.

2. **Transform an Existing Variable**: The data set has a variable called orgs, which represents the number of student organizations that students have joined. If you go to the “Data Editor” window and switch to “Data View,” you will see that the values range from 0 to 5. We will “recode” this variable so that there are only two categories: 0 (for students who do not belong to any organizations)
Using SPSS

and 1 (for students who belong to at least one organization). SPSS offers us two methods of recoding: “Recode into Same Variables” and “Recode into Different Variables.” If we choose the first method, then SPSS will overwrite the existing orgs variable, and we will lose the original information. As a result, it is always a good idea to choose the second method: “Recode into Different Variables.” When we use this method, SPSS maintains the original variable and creates a new variable according to whatever specifications we provide. This way we do not lose any information.

We are going to tell SPSS to recode orgs into orgs2 (while maintaining the original variable), using the specifications shown in Table 1.2.

Table 1.2

<table>
<thead>
<tr>
<th>Values on Original Variable, orgs</th>
<th>Values on New Variable, orgs2</th>
<th>Value Labels for New Variable, orgs2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>No organizations</td>
</tr>
<tr>
<td>1, 2, 3, 4, 5</td>
<td>1</td>
<td>At least one organization</td>
</tr>
<tr>
<td>All other values</td>
<td>System-Missing</td>
<td></td>
</tr>
</tbody>
</table>
To recode a variable, we use the following SPSS sequence of menu commands:

**Transform → Recode into Different Variables**

This opens up the “Recode into Different Variables” dialog box, shown in Figure 1.16.

![Figure 1.16](image)

- We move `orgs` into the “Input Variable” box by clicking on `orgs` and then clicking on the arrow to the right of the variable list.
- We type the new variable name, `orgs2`, into the “Output Variable: Name” box. We then click on “Change.”
- We click on “Old and New Values.” This opens another dialog box, shown in Figure 1.17, where we tell SPSS how to recode the variable, using the specifications presented in the table above.

We enter these recode commands one at a time. After we put the appropriate information into the “Old Value” and “New Value” boxes, we must click on “Add” so that SPSS adds them to the “Old → New” window. Note as well that SPSS offers us a series of shortcuts. For example, instead of entering the following:

-old value 0 = new value 0,
-old value 1 = new value 1,
-old value 2 = new value 1,
-old value 3 = new value 1,
-old value 4 = new value 1,
-old value 5 = new value 1,
SPSS allows us to perform this recode using a single entry with a range of values (e.g., 1 through 5 = 1) by selecting “Range” under the “Old Value” options.

If we have a large data set and are not sure of the lowest and highest values, SPSS gives us lowest and highest value shortcuts as well.

We have told SPSS that:

- “Old Value” 0 (on the orgs variable) will be 0 on the new variable (orgs2).
- Any respondent with a score between 1 and 5 on the original orgs variable will be a “1” on the new orgs2 variable.
- Just in case we missed anything, “All other values” on the original variable will be “System-missing” on the new variable.

Once we have finished entering the recode specifications, we click on “Continue,” then “OK,” and SPSS will create the new, recoded variable orgs2. Once again, to be certain, look at the “Data Editor” window and make sure “Variable View” is active, and you will see orgs2 appended to the bottom of the list.

When we create a new variable using the “Recode” or “Compute” procedure, we have the option of adding information about the newly created variable. For example, say we want to tell SPSS that when it produces output involving the new orgs2 variable, it should include the labels that are associated with each of the two categories. We need to tell SPSS that on the new orgs2 variable, 0 is equal to no organizations, and 1 is equal to at least one organization.
can easily add this information when the “Variable View” is active in the “Data Editor” window.

To do that, we click on the cell (circled in Figure 1.18) where the orgs2 variable meets the column labeled “Values.” This highlights the cell.

![Image](image.png)

**Figure 1.18**

With the cell highlighted, we can click on the small box on the right side of it. This opens the “Value Labels” dialog box, where we tell SPSS what labels should be attached to values 0 and 1 (shown in Figure 1.19). We do this one value at a time, and we click “Add” after we have entered each value and label into the appropriate boxes. When this is complete, we click “OK,” and SPSS will store these value labels.

We can also use the “Variable View” window to add a full label to the variable name itself (for example, we might call it “Recoded Organizations”). To do that, we click on the cell where the orgs2 variable meets the column headed by “Label.” Similarly, we can tell SPSS to treat certain values as “missing” for any variable by clicking on the appropriate cell in the “Missing” column and telling SPSS which values should be considered “missing.”
3. **Analyze an Existing Variable**: Here is an example of a sequence of menu commands that asks for a simple analysis of an existing variable.

**Analyze → Descriptive Statistics → Descriptives**

![Figure 1.19](image1)

**Figure 1.19**

This sequence will open the “Descriptives” dialog box, shown in Figure 1.20. We are going to ask SPSS to calculate some basic statistics for a variable called *friends* (number of close friends students have on campus). To do that, we move *friends* into the “Variable(s)” box and click “OK,” as shown in Figure 1.21.

![Figure 1.20](image2)
We will see this “Descriptives” procedure many times throughout the book, as it generates some commonly used statistics. Here is the output from this procedure, which you will find in the SPSS “Output” window, as shown in Figure 1.22.

<table>
<thead>
<tr>
<th>Variable(s):</th>
<th>Options...</th>
</tr>
</thead>
<tbody>
<tr>
<td>friends</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.21

<table>
<thead>
<tr>
<th>Variable(s):</th>
<th>Options...</th>
</tr>
</thead>
<tbody>
<tr>
<td>friends</td>
<td></td>
</tr>
</tbody>
</table>

SPSS has produced a table reporting the following information about the friends variable, from left to right: the number of cases for which there are available data (10), the minimum value (0), the maximum value (12), the mean (3.9), and the standard deviation (3.73). (We will cover all of these statistics in detail in chapters 4 and 5.)

Figure 1.22

<table>
<thead>
<tr>
<th>friends</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>0</td>
<td>12</td>
<td>3.90</td>
<td>3.725</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Saving Your Work

When you finish a session in SPSS, you can save the data file, save the output file, or both. In either case, you simply click on the save icon or navigate to “File” ➔ “Save.” You might save a data file if you have created new variables that you want to be able to use later. You might save an output file if you want to be able to revisit the output or use it in a report later.
1. A team of researchers is studying the effect of inequality in metropolitan areas on aggregate levels of trust. The researchers define inequality as differences in access to resources between groups of people in the area. This variable is measured by an index of inequality for each metropolitan area, ranging from 0 (no difference in access to resources between groups) to 100 (some groups have access, and others have none). The other variable is average level of trust in metropolitan areas.

a. What is the independent variable in this study?

b. What is the dependent variable?

c. What is the unit of analysis?

d. What is the sampling strategy?

2. You encounter this item on a survey about how people use social media:

   *In a typical week, how many times do you post content on a social media site?*
   *Please choose the one response that fits best.*

   • Never
   • Rarely
   • Often
   • All the time

a. Is this a closed- or open-ended survey item? Explain why.

b. What is the variable being measured by this survey item?

c. What are the categories of this variable?

d. What level of measurement is this variable? Explain how you know.

3. Before administering the social media survey mentioned in Problem 2 to a sample of respondents, a researcher assesses the survey item shown in Problem 2 for reliability and validity. He pilots the survey on a small group of respondents, asking them to answer the question every Monday for ten weeks. He notices variation in many of the individuals’ responses. One respondent gave the following responses over the ten weeks (displayed in Table 1.3).

<table>
<thead>
<tr>
<th>Table 1.3</th>
<th>One Individual’s Reported Frequency of Social Media Postings, Over Ten Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency of Responses</td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>2</td>
</tr>
<tr>
<td>Rarely</td>
<td>2</td>
</tr>
<tr>
<td>Often</td>
<td>4</td>
</tr>
<tr>
<td>All the time</td>
<td>2</td>
</tr>
</tbody>
</table>
a. Based on this information, how would you characterize the reliability of this measure?

b. What changes could be made to the response categories to increase the reliability of the measure?

4. In response to the survey item in Problem 2, another respondent chooses the “often” category three times and the “all the time” category seven times in each of the ten weeks of the pilot study. However, compared to the rest of the respondents in the pilot group, this respondent actually posts quite infrequently on social media.

a. For this pilot respondent, is reliability high or low for this survey item?

b. What about validity? Explain your answer.

5. A researcher wants to study how people cope with the stress of being homeless. Her goal is to make inferences about the total homeless population in the city from one probability sample. She decides to devise her sampling frame by asking all of the homeless shelters in the city for lists of the people who stayed at their shelter for at least one night during the last thirty days. She plans to draw a random sample from the list of residents. It is the beginning of the summer, and the researcher wants to finish collecting data by the end of the summer.

a. What is the researcher’s sampling strategy?

b. Do you think that this sampling frame will yield a sample that is representative of the city’s entire homeless population? Explain your answer.

c. Does collecting data over the summer months present any threats to the generalizability of the data? Explain your answer.

6. After deciding that it would take too many resources to get a complete list of people who had stayed at every homeless shelter in the city, the researcher from Problem 5 decides to draw a random sample of homeless shelters across three cities. Then she will randomly sample people who have stayed at each of the sampled shelters. What kind of sampling design is the researcher using now?

7. A popular website called Over the Top is known for publishing popular Top Ten lists. One of its recent lists, “The Best New Ice Cream Flavors of the Summer,” was shared widely on social media. The data for this list come from a survey of thirty ice cream shops around the country. Managers of each shop were asked to check whether they would be selling the following ten flavors during the coming summer:

- chocolate mocha
- apricot
- green tea
- black raspberry
- burnt sugar
- chocolate swirl
- ginger lemongrass
- chocolate
Practice Problems

47

- vanilla peanut swirl
- tamarind coconut pineapple

**a.** What is the variable being measured, and what is its level of measurement?

**b.** For its list, Over the Top ranked each flavor according to how many shops would be serving that flavor. Do you agree with Over the Top’s decision to call the list “The Best New Ice Cream Flavors of the Summer?” Explain your answer.

**c.** Propose a revised title for the list and explain why it works better than Over the Top’s title.

8. Employees of a humanitarian organization are concerned about the practice of separating migrant children from their parents when families try to cross the U.S. border illegally. They want to know how they can use their organization’s resources to decrease the number of children separated from their families. One group thinks that investing its resources in as many social media campaigns as possible will be more effective at reducing the numbers, while another group thinks that focusing on staging as many protest events as possible will work better.

**a.** Turn the two proposed strategies into two separate social science research questions. Be sure to pose each as a question.

**b.** Which variables are the independent variables, and which are the dependent variables for each question?

**c.** What level of measurement is each of the variables from each research question?

9. A famous study found in a probability sample of children that children’s average score on a self-discipline scale, ranging from 0 to 100, was 50. A researcher conducting a follow-up study decides to measure the self-discipline of children in her small sample relative to the average self-discipline score from the previous study, 50. To calculate each child’s relative self-discipline, she will subtract the average score from previous research (50) from each child’s score in her sample. Figure 1.23 shows the scores for the first five children in her sample.

**a.** The last column in Figure 1.23, circled, shows the relative measure of self-discipline for each child (the child’s raw score minus the average self-discipline score).
score from the previous study). Is the variable relative self-discipline an interval- or ratio-level variable? Explain how you know.

b. The researcher drew a larger sample of children and decided to transform the relative self-discipline variable into a new variable. This new variable has three categories: “low” (scores below the average score of 50), “medium” (scores equal to the average score of 50), and “high” (scores above the average score of 50). Assign each child in Figure 1.23 to the proper category of this new variable.

c. What level of measurement is the new variable? Explain your answer.

10. An open-ended survey item asked residents of a neighborhood to explain what they liked least about their neighborhood. Here are six of the responses:

- None of my neighbors talk to each other.
- When I walk around, I don’t feel like I can trust people I pass on the street.
- I have to drive too far to get to a grocery store.
- If I want to walk anywhere, I have to leave my house an hour early.
- I wish we still had block parties, but people stopped organizing them a few years ago.
- There isn’t enough public transportation.

a. Assign each of the six responses to one of the two categories of a new variable: (1) Transportation or (2) Social Relationships.

b. Propose a name for this new variable.

c. What level of measurement is this new variable? Explain how you know.

11. Figure 1.24 shows an excerpt from the GSS codebook for one of its survey items. It shows the distribution of responses to the question over time, since the GSS first asked it in 1988.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Exclusively male</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>601</td>
<td>4147</td>
<td>1021</td>
<td>968</td>
<td>1022</td>
<td>1059</td>
<td>759</td>
<td>800</td>
<td>758</td>
</tr>
<tr>
<td>Both male and female</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>118</td>
<td>43</td>
<td>34</td>
<td>14</td>
<td>40</td>
<td>25</td>
<td>25</td>
<td>36</td>
</tr>
<tr>
<td>Exclusively female</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>455</td>
<td>3475</td>
<td>882</td>
<td>879</td>
<td>882</td>
<td>893</td>
<td>703</td>
<td>707</td>
<td>654</td>
</tr>
<tr>
<td>Don’t know</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No answer</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>45</td>
<td>435</td>
<td>239</td>
<td>78</td>
<td>54</td>
<td>71</td>
<td>31</td>
<td>45</td>
<td>43</td>
</tr>
</tbody>
</table>

**Figure 1.24** Excerpt of GSS Codebook for One Variable

a. What is this variable measuring?

b. How many categories does this variable have, and what are the labels for each of them? What are the numbers assigned to each category?

c. Between 1988 and 2016, how many respondents said that their sex partners had been both male and female?

d. What level of measurement is this variable? Explain how you know.

e. Which categories should be defined as “missing data”? Explain why.
12. Implicit Association Tests (IATs) are tests that measure how strongly different concepts are linked in the brain. A user’s implicit bias is revealed when an IAT shows that a group of people is associated with a negative concept in the user’s mind (e.g., “gay” with “immoral”). For example, as shown in Figure 1.25, in an IAT for LGBTQ bias, test takers see “gay” and “straight” images (Screens 1 and 2, respectively). They must press a key that associates each image with opposing terms: “good” or “bad.” In this case, the IAT measures the amount of time, in seconds, it takes for the user to choose whether the “gay” and “straight” images match with “good” or “bad.” In this section of the test, if the user matches “straight” images with “good” more quickly than “gay” images with “good” (and “gay” images more quickly with “bad” than “straight” images with “bad), this may indicate implicit bias because it shows that the person may make automatic connections between “straight” and “good,” and “gay” and “bad.”

A researcher wants to use the IAT to test whether sexual identity is related to implicit LGBTQ bias. She hypothesizes that LGBTQ people will have less LGBTQ bias than straight people.

a. What is the independent variable? How is it measured? What is its level of measurement?

b. What is the dependent variable? How is it measured? What is its level of measurement?

13. State the level of measurement for each variable listed below:

a. Party identification, where 1 = Democrat; 2 = Republican; 3 = Independent; 4 = Other
b. Party identification, where 1 = Democrat; 2 = Independent; 3 = Republican
c. Income, measured in dollars
d. Income, where 1 = low; 2 = middle; 3 = high
e. Feeling thermometer rating of the military, 0 to 100 degrees
f. Feeling thermometer rating of the military, where 0–25 = very cold; 26–49 = cold; 50 = neutral; 51–74 = warm; 75–100 = very warm
g. How many days exercised in a week, 1–7
h. Gender, where 1 = girl and 2 = boy
i. Girl, where 0 = is not a girl and 1 = is a girl

14. Table 1.4 describes the research question and data collection strategy for five studies. For each study:

a. Identify the unit of analysis.

b. Identify whether original data are being collected or if secondary data are being used. If original data are being collected, identify the sampling strategy: non-probability sampling, simple random sampling, stratified sampling, or cluster sampling.

Table 1.4 Research Questions and Sampling Design for Five Studies

<table>
<thead>
<tr>
<th>Research Project</th>
<th>Unit of Analysis</th>
<th>Sampling Method</th>
</tr>
</thead>
</table>
| 1 Research Question: Are college Republicans or Democrats more likely to engage in campus activism?  
*Data Collection:* Make a list of all members of the Republican and Democrat student organizations at a large university. Use a random number generator to draw a sample from the list. | | |
| 2 Research Question: Is gender related to the number of close friends that high school students have?  
*Data Collection:* Make a list of all high schools in Texas and draw a random sample of schools. Draw random samples of students from each of the sampled schools. | | |
| 3 Research Question: Are companies’ parental leave policies related to the gender composition of their upper-level staff?  
*Data Collection:* Find ten companies that will allow the researcher to collect information about their organizational policies and staff. | | |
| 4 Research Question: Is the frequency of a city’s misdemeanor arrests related to the frequency of felonies committed in the city?  
*Data Collection:* Use publicly available data from the Bureau of Justice Statistics. | | |
| 5 Research Question: Who has the highest frequency of contact with the police: African Americans, whites, Asians, or Hispanics?  
*Data Collection:* Create separate sampling frames of all African American, Asian, white, and Hispanic residents of the same geographic area. Draw random samples from each of the four sampling frames. | | |
15. Confirmation bias is our tendency to disregard and discredit information that does not conform to what we already believe while accepting information that does support our beliefs.
   a. Write a few sentences regarding your beliefs about your ability to do well in this statistics course.
   b. What are some examples of information that confirms that belief?
   c. What are some examples of information that challenges that belief?
   d. If you believe that you have the capacity do well in this course, what can you do, according to the growth mindset approach to intelligence, to make your belief a reality?
   e. If you believe that you do not have the capacity to do well in this course, what can you do, according to the growth mindset approach to intelligence, to create a different reality?

Stata Problems

Open the GSS2016.dta. Here, we will focus on the variable agekdbrn, which measures the age of respondents at the time their first children were born.

1. Click on the Data Browser to visually scan the data. You will notice that the values for some variables are presented as “value labels,” while others show the numerical values assigned to the variable’s categories. Find the variable called agekdbrn. Are the categories of the variable presented as numbers or labels?

2. Use the “summarize” command to generate summary information about agekdbrn. How many people answered this question? What is the youngest age and oldest age at which first children were born to this sample of respondents? What is the mean age at which first children were born?

3. Use the “generate” command to create a new version of agekdbrn and name it agekdbrn2.

4. Use the “replace” command to transform agekdbrn2 into a variable with three categories: “Under 30,” “30 to 40,” and “Over 40.” Assign 1 to “Under 30,” 2 to “30 to 40,” and 3 to “Over 40.”

5. Click on the Data Browser again and scroll to the last column in the spreadsheet. Scroll down the agekdbrn2 column. How can you tell that the recoding occurred?

SPSS Problems

Open GSS2016.sav. Here, we will focus on the variable agekdbrn, which measures the age of respondents at the time their first children were born.

1. Navigate to “Data View” in the “Data Editor” window to visually scan the data. You will notice that the values for some variables are presented as “value labels,” while others show the numerical values assigned to the variable’s categories.
Find the variable called *agekdbrn*. Are the categories of the variable presented as numbers or labels?

2. Use the “Descriptives” dialog box to generate summary information about *agekdbrn*. How many people answered this question? What is the youngest age and oldest age at which first children were born to this sample of respondents? What is the mean age at which first children were born?

3. Use the “Recode into Different Variables” dialog box to create a new version of *agekdbrn* called *agekdbrn2*, with three categories: “Under 30,” “30 to 40,” and “Over 40.” Assign 1 to “Under 30,” 2 to “30 to 40,” and 3 to “Over 40.”

4. Navigate to “Data View” in the “Data Editor” window and scroll to the last column in the spreadsheet. Scroll down the *agekdbrn2* column. How can you tell that the recoding occurred?

---


8 We use several waves of the GSS data, each of which can be downloaded here: http://gss.norc.org/Get-The-Data. We use several waves of the ANES data, which can be downloaded here: http://www.electionstudies.org/studypages/download/datacenter_all_NoData.php. We use Wave 6 of the WVS data, which can be downloaded here: http://www.worldvaluessurvey.org/WVSDocumentation/W6.jsp. We use the 2011 wave of the PPICS data, which can be downloaded here: http://www.icpsr.umich.edu/icpsrweb/NACJD/studies/34276/version/1. We use the 2009–2010 wave of the SSS data, which can be downloaded here: https://nces.ed.gov/surveys/ssocs/data_products.asp. We use the 1997 wave of the NLSY data, which can be downloaded here: https://www.bls.gov/nls/nlsy97.htm.


23 See chapter 13 in the Stata User’s Manual for more on operators.