

1

THE WHAT AND THE WHY OF STATISTICS

CHAPTER LEARNING OBJECTIVES

- 1.1 Describe the role of theory, hypotheses, and variables in the research process.
- 1.2 Use independent and dependent variables.
- 1.3 Distinguish between the three levels of measurement and identify their complexities and limitations.
- 1.4 Apply descriptive and inferential statistical procedures to analyze data and evaluate hypotheses.
- 1.5 Explain the importance of statistics in a diverse society.

Are you taking statistics because it is required in your major—not because you find it interesting? If so, you may be feeling intimidated because you associate statistics with numbers, formulas, and abstract notations that seem inaccessible and complicated. Perhaps you feel intimidated not only because you're uncomfortable with math but also because you suspect that numbers and math don't leave room for human judgment or have any relevance to your own personal experience. In fact, you may even question the relevance of statistics to understanding people, social behavior, or society.

In this book, we will show you that statistics can be a lot more interesting and easier to understand than you may have been led to believe. In fact, as we draw on your previous knowledge and experience and relate statistics to interesting and important social issues, you'll begin to see that statistics is not just a course you have to take but a useful tool as well.

There are two reasons why learning statistics may be of value to you. First, you are constantly exposed to statistics every day of your life. Marketing surveys, voting polls, and social research findings appear daily in the news media. By learning statistics, you will become a sharper consumer of statistical material. Second, as a major in the social sciences, you may be expected to read and interpret statistical information related to your occupation or work. Even if conducting research is not a part of your work, you may still be expected to understand and learn from other people's research or to be able to write reports based on statistical analyses.

Just what is statistics, anyway? You may associate the word with numbers that indicate per capita income, COVID-19 vaccination rates, conviction rates, birthrates, divorce rates, and so on. But the word *statistics* also refers to a set of procedures used by social scientists to organize, summarize, and communicate numerical information. Only information represented by numbers can be the subject of statistical analysis. Such information is called **data**; researchers use statistical procedures to analyze data to answer research questions and test hypotheses. It is the latter usage—answering research questions and testing hypotheses—that this textbook explores.

THE RESEARCH PROCESS: THEORY, HYPOTHESES, AND VARIABLES

To give you a better idea of the role of statistics in social research, let's start by looking at the **research process**. We can think of the research process as a set of activities in which social scientists engage so that they can answer questions, examine ideas, or test hypotheses.

As illustrated in Figure 1.1, the research process consists of five stages:

1. Asking the research question
2. Formulating the hypotheses
3. Collecting data
4. Analyzing data
5. Evaluating the hypotheses

Each stage affects the theory and is affected by it as well. Statistics is most closely tied to the data analysis stage of the research process. As we will see in later chapters, statistical analysis of the data helps researchers test the validity and accuracy of their hypotheses.

The starting point for most research is asking a research question. Consider the following research questions taken from several social science journals:

How does racial discrimination affect the health of racial and ethnic minorities?

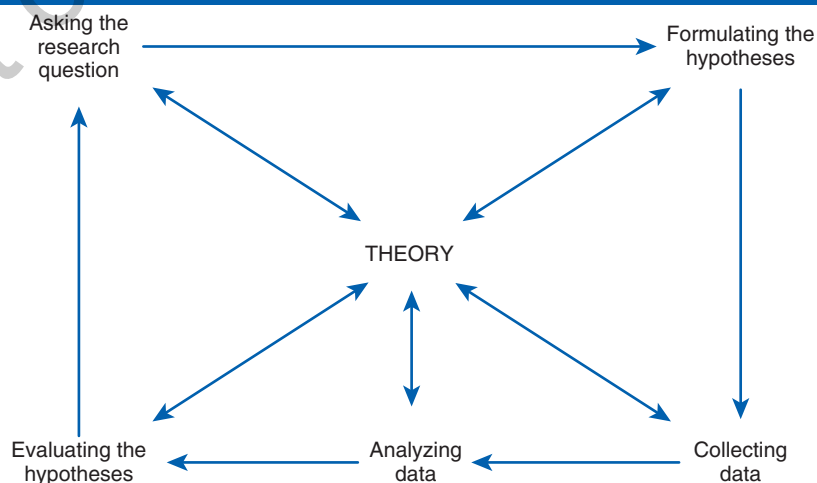
Are the sexual and romantic relationships of students shaped by the colleges and universities they attend?

How does witnessing victimization in prison affect parolees' post-release adjustment?

What factors explain the gender wage gap?

These are all questions that can be answered by conducting **empirical research**—research based on information that can be verified by using our direct experience. To answer research

FIGURE 1.1 ■ The Research Process



questions, we cannot rely on reasoning, speculation, moral judgment, or subjective preference. For example, the questions “Is racial equality good for society?” and “Is an urban lifestyle better than a rural lifestyle?” cannot be answered empirically because the terms *good* and *better* are concerned with values, beliefs, or subjective preference and, therefore, cannot be independently verified. One way to study these questions is by defining good and better in terms that can be verified empirically. For example, we can define *good* in terms of economic growth and *better* in terms of psychological well-being. These questions could then be answered by conducting empirical research.

You may wonder how to come up with a research question. The first step is to pick a question that interests you. If you are not sure, look around! Ideas for research problems are all around you, from media sources to personal experience or your own intuition. Talk to other people, write down your own observations and ideas, or learn what other social scientists have written about.

Take, for instance, the relationship between gender and work. As a college student about to enter the labor force, you may wonder about the similarities and differences between women’s and men’s work experiences and about job opportunities when you graduate. Here are some facts and observations based on research reports: In 2022, women who were employed full-time earned about \$943 (in current dollars) per week on average; men who were employed full-time earned \$1,144 (in current dollars) per week on average.¹ Women’s and men’s work are also very different. Women continue to be the minority in many of the higher-ranking and higher-salaried positions in professional and managerial occupations. For example, in 2021, women made up 19.7% of software developers and 30.02% of chief executives. In comparison, among all those employed as secretaries and administrative assistants, 96% were women. Among all receptionists and information clerks in 2021, 91.7% were women.² These observations may prompt us to ask research questions such as the following: How much change has there been in women’s work over time? Are women paid, on average, less than men for the same type of work? How does the relationship between gender and work vary by race, age, sexuality, and ability?

You may have noticed that each preceding research question was expressed in terms of a relationship. This relationship may be between two or more attributes of individuals or groups, such as gender and income or gender segregation in the workplace and income disparity. The relationship between attributes or characteristics of individuals and groups lies at the heart of social scientific inquiry.

Most of us use the term *theory* quite casually to explain events and experiences in our daily life. You may have a theory about why your roommate has been so nice to you lately or why you didn’t do so well on your last exam. In a somewhat similar manner, social scientists attempt to explain the nature of social reality. Whereas our theories about events in our lives are common-sense explanations based on educated guesses and personal experience, to the social scientist, a theory is a more precise explanation that is frequently tested by conducting research.

A **theory** is a set of assumptions and propositions used by social scientists to explain, predict, and understand the phenomena they study.³ The theory attempts to establish a link between what we observe (the data) and our conceptual understanding of why certain phenomena are related to each other in a particular way.

For instance, suppose we wanted to understand the reasons for the income disparity between men and women; we may wonder whether the types of jobs men and women have and the organizations in which they work have something to do with their wages. One explanation for gender wage inequality is gender segregation in the workplace—the fact that American men and women

are concentrated in different kinds of jobs and occupations. What is the significance of gender segregation in the workplace? In our society, people's occupations and jobs are closely associated with their level of prestige, authority, and income. The jobs in which women and men are segregated are not only different but also unequal. Although the proportion of women in the labor force has markedly increased, women are still concentrated in occupations with low pay, low prestige, and few opportunities for promotion. Thus, gender segregation in the workplace is associated with unequal earnings, authority, and status. In particular, women's segregation into different jobs and occupations from those of men is the most immediate cause of the pay gap. Women receive lower pay than men do even when they have the same level of education, skill, and experience as men in comparable occupations.

So far, we have come up with several research questions about the income disparity between men and women in the workplace. We have also discussed a possible explanation—a theory—that helps us make sense of gender inequality in wages. Is that enough? Where do we go from here?

Our next step is to test some of the ideas suggested by the gender segregation theory. But this theory, even if it sounds reasonable and logical to us, is too general and does not contain enough specific information to be tested. Instead, theories suggest specific concrete predictions or **hypotheses** about the way that observable attributes of people or groups are interrelated in real life. Hypotheses are tentative because they can be verified only after they have been tested empirically.⁴ For example, one hypothesis we can derive from the gender segregation theory is that wages in occupations in which the majority of workers are female are lower than the wages in occupations in which the majority of workers are male.

Not all hypotheses are derived directly from theories. We can generate hypotheses in many ways—from theories, directly from observations, or from intuition. Probably, the greatest source of hypotheses is the professional or scholarly literature. A critical review of the scholarly literature will familiarize you with the current state of knowledge and with hypotheses that others have studied.

Let's restate our hypothesis:

Wages in occupations in which the majority of workers are female are lower than the wages in occupations in which the majority of workers are male.

Note that this hypothesis is a statement of a relationship between two characteristics that vary: wages and gender composition of occupations. Such characteristics are called variables. A **variable** is a property of people or objects that takes on two or more values. For example, people can be classified into a number of social class categories, such as upper class, middle class, or working class. Family income is a variable; it can take on values from zero to hundreds of thousands of dollars or more. Similarly, gender composition is a variable. The percentage of females (or any other gender identity) in an occupation can vary from 0 to 100. Wages is a variable, with values from zero to thousands of dollars or more. See Table 1.1 for examples of some variables and their possible values.

Social scientists must also select a **unit of analysis**; that is, they must select the object of their research. We often focus on individual characteristics or behavior, but we could also examine groups of people such as families, formal organizations like elementary schools or corporations, or social artifacts such as children's books or advertisements. For example, we may be interested in the relationship between an individual's educational degree and annual income. In this case, the unit of analysis is the individual. On the other hand, in a study of how corporation profits

are associated with employee benefits, corporations are the unit of analysis. If we examine how often women are featured in prescription drug advertisements, the advertisements are the unit of analysis. Figure 1.2 illustrates different units of analysis frequently employed by social scientists.

TABLE 1.1 ■ Variables and Value Categories

Variable	Categories
Social class	Lower Working Middle Upper
Gender	Male Female Non-Binary
Education	Less than high school High school Some college College graduate

FIGURE 1.2 ■ Examples of Units of Analysis

Individual as unit of analysis:

How old are you?
What are your political views?
What is your occupation?



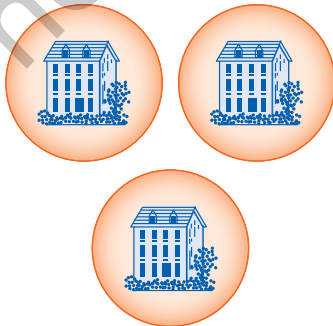
Family as unit of analysis:

How many children are in the family?
Who does the housework?
How many wage earners are there?



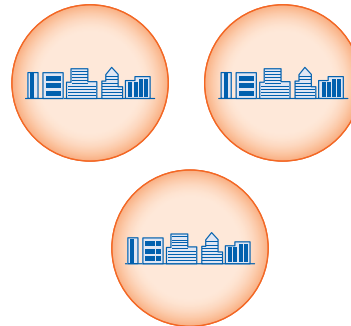
Organization as unit of analysis:

How many employees are there?
What is the gender composition?
Do you have a diversity office?



City as unit of analysis:

What was the crime rate last year?
What is the population density?
What type of government runs things?



LEARNING CHECK 1.1

Identify a social science research question and then formulate a testable hypothesis based on it. Remember that your variables must take on two or more values and you must determine the unit of analysis. What is your unit of analysis?

INDEPENDENT AND DEPENDENT VARIABLES: CAUSALITY AND GUIDELINES

Hypotheses are usually stated in terms of a relationship between two variables with one variable assumed to influence the other. And we often draw on social theories to provide an explanation for social patterns between variables. For example, according to the gender segregation theory, gender segregation in the workplace is the primary explanation (although certainly not the only one) of the male–female earning gap. Why should jobs where the majority of workers are women pay less than jobs that employ mostly men? One explanation is that:

societies undervalue the work women do, regardless of what those tasks are, because women do them. . . . For example, our culture tends to devalue caring or nurturant work at least partly because women do it. This tendency accounts for childcare workers' low rank in the pay hierarchy.⁵

Let's further discuss hypotheses and the role variables play in their construction.

Causality

The distinction between an independent and a dependent variable is important in the language of research. In the language of research, the variable the researcher wants to explain (“the effect”) is called the **dependent variable**. The variable that is expected to “cause” or account for the dependent variable is called the **independent variable**. Therefore, in our example, *gender composition of occupations* is the independent variable, and *wages* is the dependent variable.

Cause-and-effect relationships between variables are not easy to infer in the social sciences. To establish that two variables are causally related, your analysis must meet three conditions: (1) The cause has to precede the effect in time, (2) there has to be an empirical relationship between the cause and the effect, and (3) this relationship cannot be explained by other factors.

Let's consider the decades-old debate about controlling crime through the use of prevention versus punishment. Some people argue that special counseling for youths at the first sign of trouble and strict controls on access to firearms would help reduce crime. Others argue that overhauling federal and state sentencing laws to stop early prison releases is the solution. In the early 1990s, Washington and California adopted “three strikes and you're out” legislation, imposing life prison terms on three-time felony offenders. Such laws are also referred to as habitual or persistent offender laws. Twenty-six other states and the federal government adopted similar measures, all advocating a “get tough” policy on crime; the most recent legislation was in 2012 in the state of Massachusetts. In 2012, California voters supported a revision to the original law, imposing a life sentence only when the new felony conviction is serious or violent. Let's suppose that years after the measure was introduced, the crime rate declined in some of these states (in fact, advocates of the measure have identified declining crime rates as evidence of its success). Does the observation that the incidence of crime

declined mean that the new measure caused this reduction? Not necessarily! Perhaps the rate of crime had been going down for other reasons, such as improvement in the economy, and the new measure had nothing to do with it. To demonstrate a cause-and-effect relationship, we would need to show three things: (1) The reduction of crime actually occurred after the enactment of this measure, (2) the enactment of the “three strikes and you’re out” measure was empirically associated with a decrease in crime, and (3) the relationship between the reduction in crime and the “three strikes and you’re out” policy is not due to the influence of another variable (e.g., the improvement of overall economic conditions).

Guidelines

Because it is difficult to infer cause-and-effect relationships in the social sciences, be cautious about using the terms cause and effect when examining relationships between variables. However, using the terms independent variable and dependent variable is still appropriate even when this relationship is not articulated in terms of direct cause and effect. Here are a few guidelines that may help you identify the independent and dependent variables:

1. The dependent variable is always the property that you are trying to explain; it is always the object of the research.
2. The independent variable usually occurs earlier in time than the dependent variable.
3. The independent variable is often seen as influencing, directly or indirectly, the dependent variable.

The purpose of the research should help determine which is the independent variable and which is the dependent variable. In the real world, variables are neither dependent nor independent; they can be switched around depending on the research problem. A variable defined as independent in one research investigation may be a dependent variable in another.⁶ For instance, *educational attainment* may be an independent variable in a study attempting to explain how education influences political attitudes. However, in an investigation of whether a person’s level of education is influenced by the social status of his or her family of origin, *educational attainment* is the dependent variable. Some variables, such as race, age, and ethnicity, because they are primordial characteristics that cannot be explained by social scientists, are never considered dependent variables in a social science analysis.

LEARNING CHECK 1.2

Identify the independent and dependent variables in the following hypotheses:

- Older Americans are more likely to support stricter immigration laws than younger Americans.
- People who attend church regularly are more likely to oppose abortion than people who do not attend church regularly.
- People who distrust medical research are less likely to be vaccinated against COVID-19 than those who trust medical research.
- Elderly women are more likely to live alone than elderly men.

COLLECTING DATA: LEVELS OF MEASUREMENT

Once we have decided on the research question, the hypothesis, and the variables to be included in the study, we proceed to the next stage in the research cycle. This step includes measuring our variables and collecting the data. As researchers, we must decide how to measure the variables of interest to us, how to select the cases for our research, and what kind of data collection techniques we will be using. A wide variety of data collection techniques are available to us, from direct observations to survey research, experiments, or secondary sources. Similarly, we can construct numerous measuring instruments. These instruments can be as simple as a single question included in a questionnaire or as complex as a composite measure constructed through the combination of two or more questionnaire items. The choice of a particular data collection method or instrument to measure our variables depends on the study objective. For instance, suppose we decide to study how one's social class is related to attitudes about women in the labor force. Since attitudes about working women are not directly observable, we need to collect data by asking a group of people questions about their attitudes and opinions. A suitable method of data collection for this project would be a survey that uses a questionnaire or interview guide to elicit verbal reports from respondents. The questionnaire could include numerous questions designed to measure attitudes toward working women, social class, and other variables relevant to the study.

How would we go about collecting data to test the hypothesis relating the gender composition of occupations to wages? We want to gather information on the proportion of men and women in different occupations and the average earnings for these occupations. This kind of information is routinely collected and disseminated by the U.S. Department of Labor, the Bureau of Labor Statistics, and the U.S. Census Bureau. We could use these data to test our hypothesis.

The statistical analysis of data involves many mathematical operations, from simple counting to addition and multiplication. However, not every operation can be used with every variable. The type of statistical operation we employ depends on how our variables are measured. For example, for the variable *first-generation college student*, we can use the number 1 to represent yes and the number 2 to represent no. Similarly, 1 can also be used as a numerical code for the category "one child" in the variable *number of children*. Clearly, in the first example, the number is an arbitrary symbol that does not correspond to the property "first-generation college student," whereas in the second example, the number 1 has a distinct numerical meaning that does correspond to the property "one child." The correspondence between the properties we measure and the numbers representing these properties determines the type of statistical operations we can use. The degree of correspondence also leads to different ways of measuring—that is, to distinct levels of measurement. In this section, we will discuss three levels of measurement: (1) nominal, (2) ordinal, and (3) interval-ratio.

Nominal Level of Measurement

With a **nominal level of measurement**, numbers or other symbols are assigned a set of categories for the purpose of naming, labeling, or classifying the observations. *Gender*, often viewed as synonymous with sex, is an example of a nominal-level variable (Table 1.2). Using the numbers 1, 2, and 3, for instance, we can classify our observations into the categories "female," "male," and "non-binary," with 1 representing females, 2 representing males, and 3 representing non-binary people. We could use any of a variety of symbols to represent the different categories of a nominal variable; however, when numbers are used to represent the different categories, we do not imply anything about the magnitude or quantitative difference between the categories. Nominal categories cannot be rank-ordered. Because the different categories (e.g., males vs. females vs.

TABLE 1.2 ■ Nominal Variables and Value Categories

Variable	Categories
Gender	Male Female Non-Binary
Religion	Christian Hinduism Jewish Muslim
Marital status	Married Single Widowed Other

non-binary people) vary in the quality inherent in each but not in quantity, nominal variables are often called qualitative. Other examples of nominal-level variables are religion, marital status, political party, and race.

Nominal variables should include categories that are both exhaustive and mutually exclusive. Exhaustiveness means that there should be enough categories composing the variables to classify every observation. For example, the classification of the variable *religion* into the categories “Christian,” “Hinduism,” “Jewish,” and “Muslim” violates the requirement of exhaustiveness. As defined, it does not allow us to classify people who are not religious or practice another religion. We can make every variable exhaustive by adding the category “other” to the list of categories. However, this practice is not recommended if it leads to the exclusion of categories that have theoretical significance or a substantial number of observations.

Mutual exclusiveness means that there is only one category suitable for each observation. For example, we need to define *marital status* in such a way that no one would be classified into more than one category. For instance, where would we classify a person who is widowed but has remarried? They could be classified into both categories.

We should always make sure our variables are exhaustive and mutually exclusive. Admittedly the process of doing so is not always straightforward. Take for example the variable *gender* and how it’s often associated with the categories “male” and “female.” This might be a problem for two overlapping reasons as described in the 2022 National Academies of Sciences, Engineering, and Medicine publication *Measuring Sex, Gender Identity, and Sexual Orientation*.⁷ First, *gender* is a “multidimensional construct that links gender identity, gender expression, and social and cultural expectations about status, characteristics, and behavior that are associated with sex traits.”⁸ To reduce gender into a “male” or “female” binary is an oversimplification of a complex phenomenon. It also excludes those who do not identify or fit into either “male” or “female” categories. Second, it treats *gender* as interchangeable with *sex* when one’s gender identity, for example, might not be the same as their internal and/or external sex characteristics. Although it is common for *gender* to be oversimplified in quantitative research, we recommend considering how such an oversimplification might inaccurately represent its intended population and affect the research findings.

Ordinal Level of Measurement

Whenever we assign numbers to rank-ordered categories ranging from low to high or high to low, we have an **ordinal level of measurement**. *Social class* is an example of an ordinal variable. We might

TABLE 1.3 ■ Ordinal Ranking Scale

Rank	Value
1	Strongly agree
2	Agree
3	Neither agree nor disagree
4	Disagree
5	Strongly disagree

classify individuals with respect to their social class status as “upper class,” “middle class,” or “working class.” We can say that a person in the category “upper class” has a higher class position than a person in a “middle-class” category (or that a “middle-class” position is higher than a “working-class” position), but we do not know the magnitude of the differences between the categories—that is, we don’t know how much higher “upper class” is compared with the “middle class.”

Many attitudes that we measure in the social sciences are ordinal-level variables. Take, for instance, the following statement used to measure attitudes toward working women: “Women should return to their traditional role in society.” Respondents are asked to identify the number representing their degree of agreement or disagreement with this statement. One form in which a number might be made to correspond with the answers can be seen in Table 1.3. Although the differences between these numbers represent higher or lower degrees of agreement with the statement, the distance between any two of those numbers does not have a precise numerical meaning.

Like nominal variables, ordinal variables should include categories that are mutually exclusive and exhaustive.

Interval-Ratio Level of Measurement

If the categories (or values) of a variable can be rank-ordered and if the measurements for all the cases are expressed in the same units and equally spaced, then an **interval-ratio level of measurement** has been achieved. Examples of variables measured at the interval-ratio level are *age*, *income*, and *SAT scores*. With all these variables, we can compare values not only in terms of which is larger or smaller but also in terms of how much larger or smaller one is compared with another. In some discussions of levels of measurement, you will see a distinction made between interval-ratio variables that have a natural zero point (where zero means the absence of the property) and those variables that have zero as an arbitrary point. For example, weight and length have a natural zero point, whereas temperature has an arbitrary zero point. Variables with a natural zero point are also called *ratio variables*. In statistical practice, however, ratio variables are subjected to operations that treat them as interval and ignore their ratio properties. Therefore, we make no distinction between these two types in this text.

Cumulative Property of Levels of Measurement

Variables that can be measured at the interval-ratio level of measurement can also be measured at the ordinal and nominal levels. As a rule, properties that can be measured at a higher level (interval-ratio is the highest) can also be measured at lower levels, but not vice versa. Let’s take, for example, *gender composition of occupations*, the independent variable in our research example. Table 1.4 shows the percentage of women in five major occupational groups.

TABLE 1.4 ■ Gender Composition of Five Major Occupational Groups, 2021

Occupational Group	Women in Occupation (%)
Management, professional, and related occupations	52.0
Service occupations	57.7
Production, transportation, and materials occupations	24.3
Sales and office occupations	61.5
Natural resources, construction, and maintenance occupations	5.6

Source: Data from U.S. Department of Labor, 2021, Labor Force Statistics from the Current Population Survey 2021, Table 11.

The variable *gender composition* (measured as the percentage of women in the occupational group) is an interval-ratio variable and, therefore, has the properties of nominal, ordinal, and interval-ratio measures. For example, we can say that the management group differs from the natural resources group (a nominal comparison), that sales and office occupations have more women than the other occupational categories (an ordinal comparison), and that service occupations have 33.4 percentage points more women ($57.7 - 24.3$) than production occupations (an interval-ratio comparison).

The types of comparisons possible at each level of measurement are summarized in Table 1.5 and Figure 1.3. Note that differences can be established at each of the three levels, but only at the interval-ratio level can we establish the magnitude of the difference.

Levels of Measurement of Dichotomous Variables

A variable that has only two values is called a **dichotomous variable**. Several key social factors such as employment status and marital status, are dichotomies—that is, you are either employed or unemployed and married or not married. Such variables may seem to be measured at the nominal level: You fit in either one category or the other. No category is naturally higher or lower than the other, so they can't be ordered.

However, because there are only two possible values for a dichotomy, we can measure it at the ordinal or the interval-ratio level. For example, we can think of “working” as the ordering principle for employment status, so that “employed” is higher and “unemployed” is lower. Using “not working” as the ordering principle, “employed” is lower and “unemployed” is higher. In either case, with only two classes, there is no way to get them out of order; therefore, employment status could be considered at the ordinal level.

Dichotomous variables can also be considered to be interval-ratio level. Why is this? In measuring interval-ratio data, the size of the interval between the categories is meaningful: The distance between 4 and 7, for example, is the same as the distance between 11 and 14. But with a dichotomy, there is only one interval. Therefore, there is really no other distance to which we can compare it. Mathematically, this gives the dichotomy more power than other nominal-level variables (as you will notice later in the text).

For this reason, researchers often dichotomize some of their variables, turning a multicategory nominal variable into a dichotomy. For example, you may see race dichotomized into “white” and “nonwhite.” Though we would lose the ability to examine each unique racial category and we may collapse categories that are not similar, it may be the most logical statistical

TABLE 1.5 ■ Levels of Measurement and Possible Comparisons

Level	Different or Equivalent	Higher or Lower	How Much Higher
Nominal	Yes	No	No
Ordinal	Yes	Yes	No
Interval-ratio	Yes	Yes	Yes

FIGURE 1.3 ■ Levels of Measurement and Possible Comparisons: Education Measured on Nominal, Ordinal, and Interval-Ratio Levels*Possible Comparisons*

Difference or equivalence: These people have different types of education.


Graduated from public high school

Nominal Measurement


Graduated from private high school


Graduated from military academy


Possible Comparisons

Ranking or ordering: One person is higher in education than another.


Holds a high school diploma

Ordinal Measurement



Holds a college diploma


Holds a PhD

Distance Meaningless

Possible Comparisons

How much higher or lower?


Has 8 years of education

Interval-Ratio Measurement


Has 12 years of education


Has 16 years of education

4 years
Distance Meaningful

step to take. When you dichotomize a variable, be sure that the two categories capture a distinction that is important to your research question (e.g., a comparison of the number of white vs. nonwhite U.S. senators).

LEARNING CHECK 1.3

Make sure you understand these levels of measurement. As the course progresses, your instructor is likely to ask you what statistical procedure you would use to describe or analyze a set of data. To make the proper choice, you must know the level of measurement of the data.

Discrete and Continuous Variables

The statistical operations we can perform are also determined by whether the variables are continuous or discrete. Discrete variables have a minimum-size unit of measurement, which cannot be subdivided. The number of children per family is an example of a discrete variable because the minimum unit is one child. A family may have two or three children, but not 2.5 children. The variable *wages* in our research example is a discrete variable because currency has a minimum unit (1 cent), which cannot be subdivided. One can have \$101.21 or \$101.22 but not \$101.21843. Wages cannot differ by less than 1 cent—the minimum-size unit.

Unlike discrete variables, continuous variables do not have a minimum-sized unit of measurement; their range of values can be subdivided into increasingly smaller fractional values. *Length* is an example of a continuous variable because there is no minimum unit of length. A particular object may be 12 in. long, it may be 12.5 in. long, or it may be 12.532011 in. long. Although we cannot always measure all possible length values with absolute accuracy, it is possible for objects to exist at an infinite number of lengths.⁹ In principle, we can speak of a tenth of an inch, a ten thousandth of an inch, or a ten trillionth of an inch. The variable *gender composition of occupations* is a continuous variable because it is measured in proportions or percentages (e.g., the percentage of women civil engineers), which can be subdivided into smaller and smaller fractions.

This attribute of variables—whether they are continuous or discrete—affects subsequent research operations, particularly measurement procedures, data analysis, and methods of inference and generalization. However, keep in mind that, in practice, some discrete variables can be treated as if they were continuous, and vice versa.

Measurement Error

Social scientists attempt to ensure that the research process is as error free as possible, beginning with how we construct our measurements. We pay attention to two characteristics of measurement: (1) reliability and (2) validity.

Reliability means that the measurement yields consistent results each time it is used. For example, asking a sample of individuals, “Do you approve or disapprove of the Centers for Disease Control and Prevention’s response to COVID-19” is more reliable than asking “What do you think of the Centers for Disease Control and Prevention’s response to COVID-19?” Although responses to the second question are meaningful, the answers might be vague and could be subject to different interpretations. Researchers look for the consistency of measurement over time, in relationship with other related measures, or in measurements or observations made by two or more researchers. Reliability is a prerequisite for validity: We cannot measure a phenomenon if the measure we are using gives us inconsistent results.

Validity refers to the extent to which measures indicate what they are intended to measure. While standardized IQ tests are reliable, it is still debated whether such tests measure intelligence or one’s test-taking ability. A measure may not be valid due to individual error (individuals may want to provide socially desirable responses) or method error (questions may be unclear or poorly written). Specific techniques and practices for determining and improving measurement reliability and validity are the subject of research methods courses.

ANALYZING DATA AND EVALUATING THE HYPOTHESES

Following the data collection stage, researchers analyze their data and evaluate the hypotheses of the study. The data consist of codes and numbers used to represent their observations. In our example, two scores would represent each occupational group: (1) the percentage of women and

(2) the average wage. If we had collected information on 100 occupations, we would end up with 200 scores, 2 per occupational group. However, the typical research project includes more variables; therefore, the amount of data the researcher confronts is considerably larger. We now must find a systematic way to organize these data, analyze them, and use some set of procedures to decide what they mean. These last steps make up the statistical analysis stage, which is the main topic of this textbook. It is also at this point in the research cycle that statistical procedures will help us evaluate our research hypothesis and assess the theory from which the hypothesis was derived.

Descriptive and Inferential Statistics

Statistical procedures can be divided into two major categories: (1) descriptive statistics and (2) inferential statistics. Before we can discuss the difference between these two types of statistics, we need to understand the terms *population* and *sample*. A **population** is the total set of individuals, objects, groups, or events in which the researcher is interested. For example, if we were interested in looking at voting behavior in the last presidential election, we would probably define our population as all citizens who voted in the election. If we wanted to understand the employment patterns of Latinas in our state, we would include in our population all Latinas in our state who are in the labor force.

Although we are usually interested in a population, quite often, because of limited time and resources, it is impossible to study the entire population. Imagine interviewing all the citizens of the United States who voted in the last election or even all the Latinas who are in the labor force in our state. Not only would that be very expensive and time-consuming, but we would also probably have a very hard time locating everyone! Fortunately, we can learn a lot about a population if we carefully select a subset from that population. A subset of cases selected from a population is called a **sample**. The process of identifying and selecting this subset is referred to as **sampling**. Researchers usually collect their data from a sample and then generalize their observations to the population. The ultimate goal of sampling is to have a subset that closely resembles the characteristics of the population. Because the sample is intended to represent the population that we are interested in, social scientists take sampling seriously. We'll explore different sampling methods in Chapter 6.

Descriptive statistics includes procedures that help us organize and describe data collected from either a sample or a population. Occasionally, data are collected on an entire population, as in a census. **Inferential statistics**, in contrast, make predictions or inferences about a population based on observations and analyses of a sample. For instance, the General Social Survey (GSS), from which numerous examples presented in this book are drawn, is conducted every other year by the National Opinion Research Center (NORC) on a representative sample of several thousands of respondents. The survey, which includes several hundred questions (the data collection interview takes approximately 90 minutes), is designed to provide social science researchers with a readily accessible database of socially relevant attitudes, behaviors, and attributes of a cross section of the U.S. adult (18 years of age or older) population. Since 2006, the survey has been administered in English and Spanish. NORC has verified that the composition of the GSS samples closely resembles census data. But because the data are based on a sample rather than on the entire population, the average of the sample does not equal the average of the population as a whole.

Evaluating the Hypotheses

At the completion of these descriptive and inferential procedures, we can move to the next stage of the research process: the assessment and evaluation of our hypotheses and theories in light of the analyzed data. At this next stage, new questions might be raised about unexpected trends

in the data and about other variables that may have to be considered in addition to our original variables. For example, we may have found that the relationship between gender composition of occupations and earnings can be observed with respect to some groups of occupations but not others. Similarly, the relationship between these variables may apply for some racial/ethnic groups but not for others.

These findings provide evidence to help us decide how our data relate to the theoretical framework that guided our research. We may decide to revise our theory and hypothesis to take account of these later findings. Recent studies are modifying what we know about gender segregation in the workplace. These studies suggest that race as well as gender shape the occupational structure in the United States and help explain disparities in income. This reformulation of the theory calls for a modified hypothesis and new research, which starts the circular process of research all over again.

Statistics provides an important link between theory and research. As our example on gender segregation demonstrates, the application of statistical techniques is an indispensable part of the research process. The results of statistical analyses help us evaluate our hypotheses and theories, discover unanticipated patterns and trends, and provide the impetus for shaping and reformulating our theories. Nevertheless, the importance of statistics should not diminish the significance of the preceding phases of the research process. Nor does the use of statistics lessen the importance of our own judgment in the entire process. Statistical analysis is a relatively small part of the research process, and even the most rigorous statistical procedures cannot speak for themselves. If our research questions are poorly conceived or our data are flawed due to errors in our design and measurement procedures, our results will be useless.

USING STATISTICS TO EXAMINE A DIVERSE SOCIETY¹⁰

The increasing diversity of American society is relevant to social science. By the middle of this century, if current trends continue unchanged, the United States will no longer be comprised predominantly of European immigrants and their descendants. Due mostly to renewed immigration and higher birthrates, in time, nearly half the U.S. population will be of African, Asian, Latinx, or Native American ancestry.

Less partial and distorted explanations of social relations tend to result when researchers, research participants, and the research process itself reflect that diversity. A consciousness of social differences shapes the research questions we ask, how we observe and interpret our findings, and the conclusions we draw. Although diversity has been traditionally defined by race, class, and gender, other social characteristics such as sexual identity, physical ability, religion, and age have been identified as important dimensions of diversity. Statistical procedures and quantitative methodologies can be used to describe our diverse society, and we will begin to look at some applications in the next chapter. For now, we will preview some of these statistical procedures.

In Chapter 2, we will learn how to organize information using descriptive statistics and graphic techniques. These statistical tools can also be employed to learn about the characteristics and experiences of groups in our society that have not been as visible as other groups. For example, in a series of special reports published by the U.S. Census Bureau over the past few years, these descriptive statistical techniques have been used to describe the characteristics and experiences of ethnic minorities and those who are foreign born. Using data published by the U.S. Census Bureau, we discuss various graphic devices that can be used to explore the differences and similarities among the many social groups coexisting within the American society. These devices are also used to emphasize the changing age composition of the U.S. population.

Whereas the similarities and commonalities in social experiences can be depicted using measures of central tendency (see Chapter 3), the differences and diversity within social groups can be described using statistical measures of variation (see Chapter 4). In Chapters 3 and 4, we examine a variety of social demographic variables, including the racial/ethnic composition of the 50 U.S. states.

We will learn about inferential statistics and bivariate analyses in Chapters 5 through 12. First, we review the bases of inferential statistics—the normal distribution, sampling and probability, and estimation—in Chapters 5 to 7. In Chapters 8 to 12, we examine the ways in which class, sex, and ethnicity influence various social behaviors and attitudes. Inferential statistics, such as the t test, chi-square, and the F statistic, help us determine the error involved in using our samples to answer questions about the population from which they are drawn. In addition, we review several methods of bivariate analysis, which are especially suited for examining the association between different social behaviors and attitudes and variables such as race, class, ethnicity, gender, and religion. We use these methods of analysis to show not only how each of these variables operates independently in shaping behavior but also how they interlock to shape our experience as individuals in society.¹¹

Whichever model of social research you use—whether you follow a traditional one or integrate your analysis with qualitative data, whether you focus on social differences or any other aspect of social behavior—remember that any application of statistical procedures requires a basic understanding of the statistical concepts and techniques. This introductory text is intended to familiarize you with the range of descriptive and inferential statistics widely applied in the social sciences. Our emphasis on statistical techniques should not diminish the importance of human judgment and your awareness of the person-made quality of statistics. Only with this awareness can statistics become a useful tool for understanding diversity and social life.

A CLOSER LOOK 1.1

A Tale of Simple Arithmetic: How Culture May Influence How We Count

A second-grade schoolteacher posed this problem to the class:

“There are four blackbirds sitting in a tree. You take a slingshot and shoot one of them. How many are left?”

“Three,” answered the seven-year-old European with certainty.

“One subtracted from four leaves three.”

“Zero,” answered the seven-year-old African with equal certainty.

“If you shoot one bird, the others will fly away.”¹²

After years of teaching statistics, we have learned that what underlies many of the difficulties students have in learning statistics is the belief that it involves mainly memorization of meaningless formulas. There is no denying that statistics involves many strange symbols and unfamiliar terms. It is also true that you need to know some math to do statistics. But although the subject involves some mathematical computations, we will not ask you to know more than four basic operations: (1) addition, (2) subtraction, (3) multiplication, and (4) division. For a basic math review, we encourage you to refer to Appendix F.

The language of statistics may appear difficult because these operations (and how they are combined) are written in a code that is unfamiliar to you. These abstract notations are simply part of the language of statistics; much like learning any foreign language, you need to learn the alphabet before you can speak the language. Once you understand the vocabulary and are able to translate the symbols and codes into terms that are familiar to you, you will begin to see how statistical techniques simply provide another source of information with which you can analyze the diverse world around you.

Another strategy for increasing your statistical knowledge is to frame your new learning in a context that is relevant and interesting. Therefore, you will find that we rely on examples from recent sociological literature, pressing social issues, and current events to make real connections to your coursework and your life. A hallmark of our text is the use of real-world examples and data; there are some, but few, cases of fictional data in this book. We emphasize intuition, logic, and common sense over rote memorization and the derivation of formulas. In each chapter, you'll see "Learning Check" boxes where you can apply or test your new knowledge. The chapters also include "A Closer Look" boxes where we provide more detailed or background information about a particular statistical technique or interpretation. Beginning with Chapter 2, we include "Statistics in Practice" and "Reading the Research Literature" features, highlighting the interpretation of data, specific statistical calculations, or published research. We believe being statistically literate involves more than just completing a calculation; it also means learning how to apply and interpret statistical information and being able to say what it means.

What might also help develop confidence in your statistical ability is working with other students. We encourage you to collaborate with your peers as you learn this course material. We have learned that students who are intimidated by statistics do not like to admit it or talk about it. This avoidance mechanism may be an obstacle to overcoming statistics anxiety. Talking about your feelings with other students will help you realize that you are not the only one intimidated by the course. This sharing process is at the heart of the treatment of statistics anxiety—talking to others in a safe group setting will help you take risks and trust your own intuition and judgment. Ultimately, your judgment and intuition lie at the heart of your ability to translate statistical symbols and concepts into a language that makes sense and to interpret data using your newly acquired statistical tools.

DATA AT WORK

At the end of each chapter, the Data at Work feature will introduce you to people who use quantitative data and research methods in their professional lives. They represent a wide range of career fields—education, clinical psychology, international studies, public policy, publishing, government, and market research. Some may have been led to their current positions because of the explicit integration of quantitative data and research, while others are accidental data analysts—quantitative data became part of their work portfolio. Although "data" or "statistics" are not included in their job titles, these individuals are collecting, disseminating, and/or analyzing data.

We encourage you to review each profile and imagine how you could use quantitative data and methods at work.

MAIN POINTS

- Social scientists use statistics to organize, summarize, and communicate information. Only information represented by numbers can be the subject of statistical analysis.
- The research process is a set of activities in which social scientists engage to answer questions, examine ideas, or test hypotheses. It consists of the following stages: asking the research question, formulating the hypotheses, collecting data, analyzing data, and evaluating the hypotheses.
- A theory is a set of assumptions and propositions used for explanation, prediction, and understanding of social phenomena. Theories offer specific concrete predictions about the way observable attributes of people or groups would be interrelated in real life. These predictions, called hypotheses, are tentative answers to research problems.
- A variable is a property of people or objects that takes on two or more values. The variable that the researcher wants to explain (the “effect”) is called the dependent variable. The variable that is expected to “cause” or account for the dependent variable is called the independent variable.
- Three conditions are required to establish causal relations: (1) The cause has to precede the effect in time, (2) there has to be an empirical relationship between the cause and the effect, and (3) this relationship cannot be explained by other factors.
- At the nominal level of measurement, numbers or other symbols are assigned to a set of categories to name, label, or classify the observations. At the ordinal level of measurement, categories can be rank-ordered from low to high (or vice versa). At the interval-ratio level of measurement, measurements for all cases are expressed in the same unit.
- A population is the total set of individuals, objects, groups, or events in which the researcher is interested. A sample is a relatively small subset selected from a population. Sampling is the process of identifying and selecting the subset.
- Descriptive statistics includes procedures that help us organize and describe data collected from either a sample or a population. Inferential statistics is concerned with making predictions or inferences about a population from observations and analyses of a sample.

KEY TERMS

data	population
dependent variable	reliability
descriptive statistics	research process
dichotomous variable	sample
empirical research	sampling
hypothesis	statistics
independent variable	theory
inferential statistics	unit of analysis
interval-ratio level of measurement	validity
nominal level of measurement	variable
ordinal level of measurement	

INTRODUCTION TO SOFTWARE, DATA SETS, AND VARIABLES

End-of-chapter practice problems have been organized into three sections: SPSS, Excel, and Chapter Exercise calculation and interpretation problems. Chapter Exercises do not require the use of computer software. SPSS Problems are based on the program IBM SPSS Version 28. Excel Problems use Microsoft Excel for MacBook Version 16.64.

Before attempting the SPSS and/or Excel Problems, you will find demonstrations that we strongly encourage you to work through. The demonstrations and related problems are organized by software and labeled accordingly: SPSS Demonstrations are followed by SPSS Problems. The same format follows for Excel. Excel Demonstrations are followed by Excel Problems.

For all SPSS and Excel Problems in this textbook, we will be working with 2021 General Social Survey (GSS) data. The GSS has been conducted biennially since 1972. Conducted by the NORC at the University of Chicago, with principal funding from the National Science Foundation, the GSS is designed to provide social science researchers with a readily accessible database of socially relevant attitudes, behaviors, and attributes of a cross section of the U.S. population. Next to the U.S. Census data, the GSS is the most frequently analyzed source of social science information by educators, legislators, and media outlets. From the GSS, we've created two data sets for use with SPSS and titled them: GSS21SSDS-A and GSS21SSDS-B. They each contain a selection of 50 variables¹³ and 1,500 cases.

We also created one data set for use with Excel and titled it GSS21SSDS-E. It contains 22 variables, 135 cases, and two sheets (Data View and Variable View). We locked both Excel sheets to avoid any changes that might accidentally be introduced as you click around the GSS21SSDS-E data set. All this means is that you cannot change any information in any of the cells. You can easily unlock an Excel sheet by clicking on *Home* → *Format* → *Unprotect Sheet*. If you have not yet installed Excel's Analysis ToolPak (an available add-in), you can do so by selecting *Tools* → *Excel Add-ins* → *Analysis ToolPak* → *OK* in the main toolbar. After adding the Analysis ToolPak, you will find the Data Analysis option off to the right in the Excel Data Tab.

SPSS DEMONSTRATION [GSS21SSDS-A]

To access the data sets for this chapter and view tutorial videos for using SPSS, visit edge.sagepub.com/frankfort10e.

The SPSS appendix found on this text's study site edge.sagepub.com/frankfort10e explains the basic operation and procedures for SPSS for Windows Student Version. We strongly recommend that you refer to this appendix before beginning the SPSS exercises.

When you begin using a data set, you should take the time to review your variables. What are the variables called? What do they measure? What do they mean? There are several ways to do this.

To review your data, you must first open the data file. Files are opened in SPSS by clicking on *File*, then *Open*, and then *Data*. After switching directories and drives to the appropriate location of the files (which we encourage you to save to your desktop or USB storage device), you select one data file and click on *Open*. This routine is the same each time you open a data file. SPSS automatically opens each data file in the SPSS Data Editor window labeled Data View. We'll use GSS21SSDS-A for this demonstration.

One way to review the complete list of variables in a file is to click on the *Utilities* choice from the main menu, then on *Variables* in the list of submenu choices. The SPSS variable labels are listed in the scroll box (refer to Figure 1.4). When a variable label is highlighted, the variable

information for that variable is listed, along with any missing values and, if available, the value labels for each variable category. (As you use this feature, please note that sometimes SPSS mislabels the variable’s measurement level. Always confirm that the reported SPSS measurement level is correct.) SPSS allows you to display data in alphabetical order (based on the variable name) or in the order presented in the file (which may not be alphabetical).

A second way to review all variables is through the Variable View window. Notice on the bottom of your screen that there are two tabs, one for *Data View* and the other for *Variable View*. Click on *Variable View*, and you’ll see all the variables listed in the order in which they appear in the Data View window (as depicted in Figure 1.5). Each column provides specific information about the variables. The columns labeled “Label” and “Values” provide the variable label (a brief label of what it’s measuring) and value labels (for each variable category).

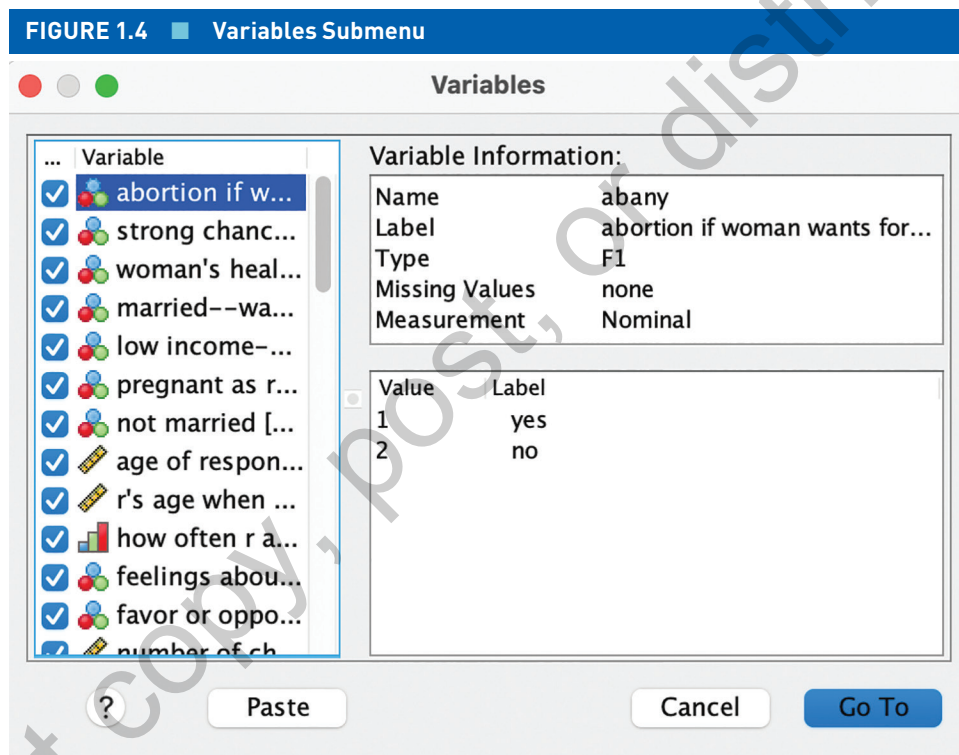


FIGURE 1.5 Variable View Window

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	abany	Numeric	1	0	abortion if wom...	{1, yes}...	None	8	Right	Nominal	Input
2	abdefect	Numeric	1	0	strong chance ...	{1, yes}...	None	8	Right	Nominal	Input
3	abhlth	Numeric	1	0	woman's health...	{1, yes}...	None	8	Right	Nominal	Input
4	abnomore	Numeric	1	0	married--want...	{1, yes}...	None	8	Right	Nominal	Input
5	abpoor	Numeric	1	0	low income--ca...	{1, yes}...	None	8	Right	Nominal	Input

SPSS PROBLEM IGSS21SSDS-A1

- S1. Based on the *Utilities-Variables* option, review the variables from the GSS21SSDS-A. Can you identify three nominal variables, three ordinal variables, and at least one interval-ratio variable? Based on the information in the dialog box or Variable View window, you should be able to identify the variable name, variable label, and category values.

EXCEL DEMONSTRATION [GSS21SSDS-E]

To access the data set for this chapter and view tutorial videos for using Excel, visit edge.sagepub.com/frankfort10e.

Throughout our discussion of the Excel program in this book, we focus on how to create tables and produce summaries of data to enhance your learning of statistics. We do not discuss how to enter data into Excel, for we will be working with an existing data set that we've created especially for this textbook (GSS21SSDS-E). For many students, entering data into Excel is the easy part. It's using the program to summarize the data that most people find challenging. If you would like a greater discussion of the full range of capabilities Excel offers, we suggest you review any number of more exhaustive Excel guides that are available.

Microsoft Office is a relatively affordable software bundle that includes Microsoft Word, Microsoft PowerPoint, Microsoft Excel, and more. If you are taking your statistics class near the end of your undergraduate career, you are likely very proficient with Word and PowerPoint. While we are sure you've heard of Excel, our experience teaching statistics over the years suggests you are probably not very proficient using it to analyze data. Given how readily available Excel is to most computer owners, we believe it is underused in undergraduate statistics courses. Let's change that.

Open up GSS21SSDS-E and examine its contents. The cells shaded in light gray indicate missing data. Also, notice how we've created two Excel sheets: *Data View* and *Variable View*.

FIGURE 1.6 ■ Data View Sheet

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	
1	abany	age	agekbrn	childs	chidldel	class	degree	educ	hapmar	happy	health	hispan	marital	partyid	premarxs	prochoic	qualife	race	sex	sibs	tvhours	xmovie	
2		57	30	2		2 Middl	Bachelo	18		Not Toi	Good	No	Divorced	Independen	Wrong On	Neither A	Good	White	Male	2	3	No	
3	Yes	67	30	2		3 Work	High Sc	12		Pretty	Good	No	Married	Independen		Agree	Good	White	Male	1	1	No	
4		20		0		3 Work	LT High	11		Not Toi	Good	No	Never Married	Independen	Wrong On	Neither A	Good	White	Female	4	2		
5	No		16	2		2 Lower	High Sc	12		Not Toi	Good	No	Divorced		Independen			White	Female	5		No	
6	No	52		0		2 Middl	Bachelo	16		Very Hap	Very Hi	Good	Married	Not Very Str	Not Wronj	Agree	Excellent	White	Male	2	1		
7	Yes	53	18	1		1 Work	Associat	15		Pretty	Fair	No	Divorced		Independen		Agree	Good	White	Female	1		No
8	Yes	25		0		1 Middl	Bachelo	18		Not Toi	Good	No	Never Married	Strong Dem	Not Wronj	Strongly /	Very Goc	White	Male	2	2		
9	Yes	52	40	1		2 Work	High Sc	12		Pretty Hi	Pretty	Excellent	Married	Not Very Str	Not Wronj			White	Female	3	5		
10	No	48		0		1 Work	High Sc	14		Pretty	Good	No	Divorced	Strong Rep.	Not Wronj	Neither A	Good	White	Female	5	3		
11		64	20	2		2 Middl	Graduat	18		Pretty	Good	No	Divorced	Not Very Str	Not Wronj	Agree	Good	White	Female	4	2		
12		49		0		2 Work	High Sc	12		Very Hap	Very Hi	Good	Married	Independen	Not Wronj	Agree	Very Goc	White	Female	4	2	No	
13		37	19	3		1 Work	Bachelo	16		Very Hap	Pretty	Excellent	Married	Not Very Str	Not Wronj	Strongly /	Very Goc	White	Female	1	0	Yes	
14	Yes	47		0		1 Work	Associat	18		Pretty	Good	No	Never Married	Strong Dem	Not Wronj	Agree	Very Goc	White	Male	4	10		
15			27	1		2 Lower	LT High	10		Very Hap	Pretty	Fair	Married	Strong Rep.	Not Wronj	Disagree	Good	White	Male	2	6	No	
16		67	26	1		1 Work	High Sc	15		Very Hap	Very Hi	Excellent	Married		Independen		Agree	Very Goc	White	Female	1		No
17		73		0		2 Work	Associat	14		Pretty	Fair	No	Widowed	Independen	Not Wronj	Strongly /	Very Goc	White	Male	0	6	No	
18		36	27	2		1 Middl	Bachelo	16		Very Hap	Pretty	Good	Married	Strong Dem	Not Wronj	Neither A	Very Goc	White	Female	2	3	No	
19		38	24	2		3 Middl	Bachelo	16		Pretty Hi	Pretty	Good	Married	Strong Dem	Not Wronj	Strongly /	Very Goc	White	Male	2	2		
20		69	17	2		1 Middl	High Sc	11		Very Hap	Pretty	Fair	Married	Not Very Str		Strongly /	Good	White	Male	2		No	
21		29		0		1 Middl	High Sc	12		Not Toi	Good	No	Never Married	Independen		Agree	Very Goc	White	Male	3		Yes	
22		80	22	2		3	High Sc	12		Pretty	Fair	No	Widowed	Strong Dem	Wrong On	Agree	Good	White	Female	4	24		
23	Yes	28		0		1 Upper	Bachelo	16		Not Toi	Excellent	No	Never Married			Strongly /	Excellent	White	Male	2		Yes	
24		85	24	6		4 Middl	High Sc	15		Very Hi	Excellent	No	Widowed	Strong Rep.	Always Wt	Strongly /	Good	White	Male	5	8	No	
25	Yes	73	19	2		2 Middl	High Sc	12		Very Hap	Very Hi	Good	Married	Strong Rep.	Not Wronj	Disagree	Excellent	White	Female	4	3		
26		74	22	4		3 Middl	Graduat	16		Very Hap	Pretty	Good	Married	Independen	Wrong On	Strongly /	Very Goc	White	Male	1	5	No	
27	Yes	52	35	3		1 Middl	Bachelo	17		Very Hap	Pretty	Good	Married	Strong Dem		Neither A	Good	White	Male	3		No	
28	Yes	48	27	3		4 Middl	Graduat	18		Pretty Hi	Pretty	Fair	Married	Strong Dem	Not Wronj	Strongly /	Very Goc	White	Female	2	2		
29		79	23	2		2 Middl	High Sc	13		Pretty	Good	No	Divorced	Not Very Str	Always Wt	Disagree	Very Goc	White	Female	2	5		
30	Yes	64	30	4		8 Upper	Graduat	18		Very Hap	Very Hi	Excellent	Married	Strong Dem	Not Wronj			White	Female	3	1		
31	Yes	67	39	1		2 Middl	Graduat	18		Pretty	Good	No	Divorced	Independen	Not Wronj	Disagree	Very Goc	White	Female	1	2		
32		55	35	3		1 Middl	Bachelo	16		Pretty Hi	Pretty	Excellent	Married	Not Very Str		Strongly /	Very Goc	White	Male	2		Yes	
33	No	72	29	2		1 Work	Associat	15		Very Hap	Very Hi	Excellent	Married		Independen		Neither A	Good	White	Female	3		No

Figure 1.6 shows the Data View sheet. The first row is a list of all of the variables in the data set beginning with ABANY, which stands for “Abortion if Woman Wants for Any Reason.” Beginning with row 2, each row represents an individual respondent. The last respondent in the data set is in row 136 (not pictured). Because our respondents begin on row 2, we know that there are a total of 135 (136 – 1) respondents in our data set.

If you click on the “Variable View” tab at the bottom of the file, you will move to the second sheet that contains a list of 22 variables in GSS21SSDS-E. The first row is for information purposes only and is not a variable. See Figure 1.7. This sheet offers more information about each variable, including each variable’s label.

FIGURE 1.7 Variable View Sheet

GSS21SSDS-E

	A	B	C	D	E	F	G	H	I
1	NAME	LABEL	VALUES	MISSING	MEASURE				
2	ABANY	Abortion if woman wants for any reason	1=Yes 2=No		Nominal				
3	AGE	Age of respondent	89=89 or older		Scale				
4	AGEKDBRN	R's age when 1st child born			Scale				
5	CHILDS	Number of children	8=8 or more		Ordinal				
6	CHLIDDEL	Ideal number of children	7=Seven+	8	Ordinal				
7	CLASS	Subjective class identification	1=Lower class 2=Working class 3=Middle class 4=Upper class	5	Ordinal				
8	DEGREE	R's highest degree	0=LT High Sch. 1=HS 2= Associate/Junior College		Ordinal				
9	EDUC	Highest year of school completed	3=Bachelors 4=Graduate		Scale				
10	HAPMAR	Happiness of marriage	1=Very happy 2=Pretty happy 3=Not too happy		Ordinal				
11	HAPPY	General happiness	1=Very happy 2=Pretty happy 3=Not too happy		Ordinal				
12	HEALTH	Condition of health	1=Excellent 2=Good 3=Fair 4=Poor		Ordinal				
13	HISPANIC	Hispanic	1=No 2=Yes						
14	MARITAL	Marital status	1=Married 2=Widowed 3=Divorced 4=Separated 5=Never Married		Nominal				
15	PARTYID	Political party affiliation	0=Strong democrat 1=Not very str dem 2=Ind, close to dem 3=Independent 4=Ind, close to repub 5=Not very str rep 6=Strong republican	7	Ordinal				
16	PREMARSX	Sex before marriage	1=Always wrong 2=Almost always wrong 3=Wrong only sometimes 4=Not wrong at all	5	Ordinal				
17	PROCHOIC	I consider myself pro-choice	1=Strongly agree 2=Agree 3=Neither agree nor disagree 4=Disagree 5=Strongly disagree		Ordinal				
18	QUALLIFE	R's quality of life	1=Excellent 2=Very good 3=Good 4=Fair 5=Poor		Nominal				
19	RACE	R's race, three categories	1=White 2=Black 3=Other						
20	SEX	Respondents sex	1=Male 2=Female		Nominal				
21	SIBS	Number of brothers and sisters			Scale				
22	TVHOURS	Hours per day watching tv			Scale				
23	XMOVIE	Seen x-rated movie in last year	1=Yes 2=No		Nominal				
24									
25									
26									
27									
28									
29									
30									
31									
32									

Ready Accessibility: Good to go

EXCEL PROBLEM [GSS21SSDS-E]

- E1.** Closely examine the Variable View tab of GSS21SSDS-E and identify two nominal variables, two ordinal variables, and two interval-ratio variables.

CHAPTER EXERCISES

- C1.** In your own words, explain the relationship of data (collecting and analyzing) to the research process. (Refer to Figure 1.1.)
- C2.** Construct potential hypotheses or research questions to relate the variables in each of the following examples. Also, write a brief statement explaining why you believe there is a relationship between the variables as specified in your hypotheses.
- Political party and support for a mask mandate on airplanes to reduce the transmission of COVID-19
 - Income and race/ethnicity
 - The crime rate and the number of police in a city
 - Life satisfaction and marital status
 - Age and support for student loan forgiveness
 - Care of elderly parents and ethnicity
- C3.** Determine the level of measurement for each of the following variables:
- The number of people in your statistics class
 - The percentage of students who are first-generation college students at your school
 - The name of each academic major offered in your college
 - The rating of the overall quality of a smartphone, on a scale from “excellent” to “poor”
 - The type of transportation a person takes to school (e.g., bus, walk, car)
 - The number of hours you study for a statistics exam
 - The rating of the overall quality of your campus coffee shop, on a scale from “excellent” to “poor”
- C4.** For each of the variables in Exercise 3 that you classified as interval-ratio, identify whether it is discrete or continuous.
- C5.** Why do you think men and women, on average, do not earn the same amount of money? Develop your own theory to explain the difference. Use three independent variables in your theory, with annual income as your dependent variable. Construct hypotheses to link each independent variable with your dependent variable.
- C6.** For each of the following examples, indicate whether it involves the use of descriptive or inferential statistics. Justify your answer.
- The number of unemployed people in the United States
 - Determining students’ opinion about the food options at the student union based on a sample of 100 students
 - The national incidence of breast cancer among Asian women
 - Conducting a study to determine a community’s understanding of the Black Lives Matter movement, gathered from 1,000 residents
 - The average GPA of various majors (e.g., sociology, psychology, English) at your university
 - The change in the number of immigrants coming to the United States from Southeast Asian countries between 2015 and 2020

- C7.** Construct measures of political participation at the nominal, ordinal, and interval-ratio levels. (*Hint:* You can use behaviors such as voting frequency or political party membership.) Discuss the advantages and disadvantages of each.
- C8.** Using an original database that they constructed, Jessica L. Schachle and Jonathan S. Coley (2022)¹⁴ examined the number of Asian, Black, Latinx, and Native American students groups across 1,910 four-year, not-for-profit, U.S. colleges and universities. Their analysis involves a number of variables for each school including the following: presence of a student of color organization broken down by the following four racial/ethnic groups (Asian, Black, Latinx, or Native American); the number of such student organizations; the percentage of students of color; whether or not the school offered an “ethnic studies” major or something more specific including “Black or African American Studies, Chicano or Hispanic/Latinx Studies,” etc.; and finally a variable noting if the school had a Diversity, Equity, and Inclusion (DEI) center.
- What is the unit of analysis in Schachle and Coley’s study?
 - Which two variables did the authors likely use as dependent variables? And what are their levels of measurement?
 - Identify two independent variables in their research. And identify the level of measurement for each.
 - Schachle and Coley conclude that “schools that have higher percentages of students of color, offer ethnic studies majors, and maintain centers devoted to issues of racial diversity, equity, and inclusion” are more likely than those that don’t to have at least one student of color organization. From a sociological perspective, what might explain this finding?
- C9.** Variables can be measured according to more than one level of measurement. For the following variables, identify at least two levels of measurement. Is one level of measurement better than another? Explain.
- Individual age
 - Annual income
 - Religiosity
 - Student performance
 - Social class
 - Number of children
- C10.** Statistics are more than just batting averages. They can be used to make sense of the diverse society we live in. For example, we could use statistics to understand how vaccination against COVID-19 varies by class, race, and age. Identify three questions that you would be interested in using statistics to explore and explain how each is sociologically relevant.