1 A BRIEF OVERVIEW OF THE RESEARCH PROCESS

CHAPTER PREVIEW

Steps in the Research Process	Example from Valkenburg and Schouten, "Friend Networking Sites and Their Relationship to Adolescents' Well-Being and Social Self-Esteem"
Step 1: Choose a research area and read the literature	• Impact of social media on self-esteem and well-being among teens
Step 2: Identify the gaps or ways to extend the literature	• Limited research on uses and consequences of social media use among adolescents • Lack of distinction between social and nonsocial Internet use
Step 3: Examine the theory	• Human beings have a desire to protect and enhance their self-esteem. • Self-esteem is strongly related to well-being.
Step 4: Develop your research questions and form hypotheses	• Does the frequency with which teens use networking sites have an impact on their self-esteem and well-being? • Does positive or negative feedback affect self-esteem?
Step 5: Develop your research method	• Online survey among adolescents between 10 and 19 years of age who have a profile on a social networking site
Step 6: Examine the data or other evidence	• Descriptive statistics of frequency of usage and types of feedback received from peers • Regression analysis to determine impact on self-esteem

(*Continued*)

Step 7: Write the research paper	● Introduction ● Literature Review ● Data and Methods ● Results ● Discussion ● Conclusion

Source of example in the second column: Valkenburg, Peter, and Schouten (2006).

1.1 INTRODUCTION

Does the use of ChatGPT to practice homework problems improve scores? Were mask mandates motivated by politics during the COVID-19 pandemic? Do education levels vary by gender identity among those who use online dating applications? Do students from high-income families earn higher SAT scores? These are just some of the examples used in this book to illustrate the endless number of interesting questions that can be examined with **statistics**.

Although the majority of this book is focused on statistics, it is important to understand where and how **data analysis** plays a role in the research process. We begin, therefore, by giving you a brief overview of the research process. This includes choosing a research area, identifying gaps in the literature, examining the theory, developing research questions and hypotheses, identifying your research method, analyzing data, and writing the research paper. Although this is a brief overview, some of these topics are covered in greater detail later in the book. In particular, Chapter 16 on "Writing the Research Paper," offers guidance and examples from published papers for each section of a research paper, including how to structure a literature review, examine the theory, describe your data and methods, report statistical results, discuss your results within the context of the literature and theory, and offer your final conclusions, limitations, and areas for future research.

1.2 WHAT IS RESEARCH

Research is often described as the creation of knowledge. It begins with the construction of an argument that can be supported by evidence. As described by Greenlaw (2009), scholars then create a "conversation" in scholarly journals to discuss the argument. In many cases, scholars will identify gaps in the argument and offer alternate views or evidence. In other cases, scholars may forward or extend the argument by offering new insights or examine the same argument from a different angle. Another equally valid form of research is to replicate what others have done. This can be done by conducting the same research in a different region, in a different time period, over a longer time period, or with a different set of participants. All of these may validate the original argument or disprove it.

1.3 STEPS IN THE RESEARCH PROCESS

1.3.1 Read the Literature and Identify Gaps or Ways to Extend the Literature

When starting a new research project, it is common to begin by choosing a general area, such as poverty, pollution, sports, social media, criminal justice, and so on. Before identifying a research question within the general area, you must begin reading the ***literature***. The literature can be defined as a body of articles and books, written by experts and scholars, that has been *peer reviewed*. A peer review is when two or three scholars are asked to anonymously evaluate a manuscript's suitability for publication and either reject it or accept it, typically with revisions based on their recommendations.[1] Articles in the body of literature will cite other sources and will be written for an audience of fellow scholars. Nonscholarly materials, such as newspapers, trade and professional sources, letters to the editor, and opinion-based articles are not considered part of the literature. They are sometimes used in scholarly papers, but never as a sole source of information.

Most disciplines have their own databases, with articles, book chapters, dissertations, and working papers from their field. Table 1.1 shows a list of the key databases in several fields.

TABLE 1.1 ■ Databases Of Scholarly Literature From Different Fields

Field	Database	Content	Website
Criminal Justice	ProQuest Criminal Justice Database	A comprehensive database of U.S. and international criminal justice journals	www.proquest.com/products-services/pqcriminaljustice.html
	Criminal Justice Abstracts	Titles and abstracts for articles from most significant sources in the field	www.ebsco.com/products/research-databases/criminal-justice-abstracts
Economics	Econ Lit	Over 2,000 journals, plus books, dissertations, working papers, and book reviews	www.aeaweb.org/econlit
Political Science	JSTOR	6,800 political science journals, books, and pamphlets	www.jstor.org/action/showJournals?discipline=43693417
	Academic Search Complete	340 full-text political science reference books and monographs and more than 44,000 full-text conference papers	www.ebscohost.com/academic/subjects/category/political-science
Psychology	PsycINFO	Four million bibliographic records, including more than 2 million digital object identifiers to allow for direct linking to full-text psychology articles and literature. Indexing of more than 2,500 scholarly psychology journals	www.apa.org/psycinfo

(Continued)

TABLE 1.1 ■ Databases of Scholarly Literature From Different Fields (*Continued*)

Field	Database	Content	Website
Public Health	PubMed	Access to 12 million Medline citations dating back to the 1950s	www.ncbi.nlm.nih.gov/pubmed
	PAIS	Political, social, and public policy issues	www.proquest.com
	Nexis Uni	15,000 news, business, and legal sources	www.lexisnexis.com
Sociology	Sociological Abstracts	Abstracts of sociology journal articles and citations to book reviews drawn from more than 1,800 serial publications and abstracts of books, book chapters, dissertations, and conference papers	http://proquest.libguides.com/SocAbs
	JSTOR	8,000 sociology journals, books, and pamphlets	www.jstor.org/action/showJournals?discipline=43693423
	Academic Search Complete	900 full-text sociology journals, abstracts for more than 1,500 "core" coverage journals, data from nearly 420 "priority" coverage journals and more than 2,900 "selective" coverage journals, and indexing for books/monographs, conference papers, and other nonperiodicals	www.ebscohost.com/academic/socindex

In all of these databases, you can type in keywords from areas that interest you. You can then peruse article titles and read abstracts to get a sense of the thought-provoking questions and research in your area of interest. Once you have found some key articles that zero in on your research interests, you can review earlier articles that were referenced by the key articles (backward citation searching) and search forward in time to see what other articles have cited your key articles since they were written. For example, if an article was written in 1995, you can find every article written since 1995 that has cited the original article. This can be done through Google Scholar, PubMed, Science Direct, Scopus, and Web of Science. As you find more articles related to your specific topic, you will find that the literature will indicate what has been done in your area of interest, what questions remain, and if there are gaps or contradictions in the literature. All articles will also indicate the flaws in their own research and areas for future research. You can then identify your own research questions based on the contradictions or gaps in the literature or the need for forwarding or extending the argument. As mentioned earlier, you can also replicate what other authors have done by repeating the same study based on a different time period, a different region or country, or a different set of data.

For more information on how to identify gaps in the literature and write a literature review, refer to Chapter 16, "Writing a Research Paper," which offers guidelines on each section of a research paper along with examples from journal articles to illustrate these concepts.

1.3.2 Examine the Theory

A *theory* can be defined as a comprehensive explanation that is supported by a large body of evidence. For example, the theory of comparative advantage used by economists suggests that countries will specialize in producing a good in which they have a lower opportunity cost and trade with each other to benefit from mutual gains. Another example is Darwin's theory of evolution, which is used to explain changes in species over time.

Theories are different from hypotheses and laws. A hypothesis is a testable prediction. For example, you could test the hypothesis that increased exposure to sunlight will lead to higher levels of vitamin D in the body. Unlike theories and hypotheses, which can be updated based on new evidence, a law describes a universal and consistent relationship between two or more variables. For example, the law of demand states that as the price of a good increases, the quantity demanded will decrease, holding all other factors constant.

Theory plays an important role in developing your research questions and hypotheses. In the article used in the chapter preview, for example, Valkenburg et al. (2006) cite the theory that humans have a desire to protect their self-esteem and that self-esteem affects well-being. From this basic theory, they develop their research question related to how social media usage affects self-esteem and thus well-being.

As a second example, the theory of social capital could be used to develop research questions. This theory suggests that individuals benefit from social networks that can offer emotional support, access to resources, and opportunities. Although this theory was first developed within the field of sociology, many fields use the theory of social capital including economics, public health, political science, and education. Using social capital theory as our framework, we could ask how social media usage contributes to the formation of social capital among college students and if this social capital impacts their performance. Our hypothesis could be that social capital leads to better academic performance.

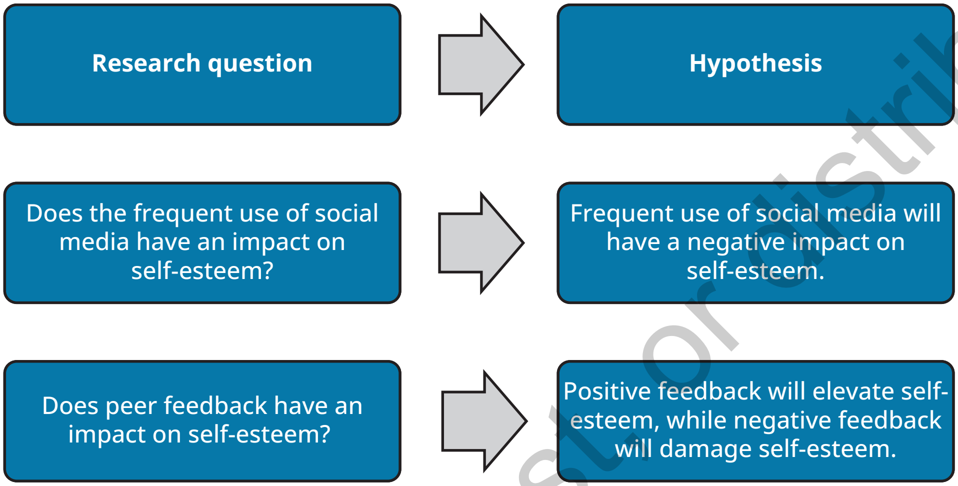
Although theory offers a framework to help develop hypotheses, we also return to the theory when examining the results of our study. In other words, do your results conform to the stated theories? How do they differ? Why might they differ? These concepts are covered in more detail in Chapter 16, "Writing a Research Paper."

1.3.3 Develop your Research Questions and Hypotheses

As described in the previous sections, you begin to form your research questions as you read the literature and examine the theory. Your questions may change in the early stages of the research as you continue to find more articles on the topic or new ways that scholars have examined or answered the questions in your research area.

In the example used in the chapter preview, the authors identify two research questions that are illustrated in Figure 1.1. Each of these questions can then be restated as a *hypothesis*, or an answer to the questions. As you begin your research, you won't know the answer to your research questions, but your hypotheses indicate what you expect to find based on theory. Your research may then find evidence to support or refute your hypothesis, which is a key feature of a hypothesis. It must be testable.

FIGURE 1.1 ■ From Research Question To Hypothesis


Research question
Hypothesis
Does the frequent use of social media have an impact on self-esteem?
Frequent use of social media will have a negative impact on self-esteem.
Does peer feedback have an impact on self-esteem?
Positive feedback will elevate self-esteem, while negative feedback will damage self-esteem.

Developing the research questions is often the most difficult part of the research process and requires a lot of work up front before the **questionnaire** or study design can or should begin.

In addition to identifying the research question, it is also important to begin thinking about your key variables (self-esteem, social media usage, and feedback, in this case) and how they relate to one another. In particular, self-esteem is the ***dependent variable*** because its value depends on the two independent variables: social media usage and feedback received. A dependent variable is defined in general as a variable whose variation is influenced by other variables. This is covered in more detail in later chapters.

1.3.4 Identify your Research Method

Once you have identified your research questions, your next step is to develop your research method. There are many types of research methods, such as qualitative research (narrative research, case studies, ethnographies), quantitative research (surveys and experiments with statistical analysis), and mixed methods that include both qualitative and quantitative approaches. Since this textbook focuses on quantitative analysis of *primary data* (data collected by the researcher) and *secondary data* (data collected by someone else), the remaining chapters in this book will be devoted to sampling, questionnaire design, and data analysis, with a final chapter on writing a research paper. For more complete works on the other types of research methods mentioned, see Leedy and Ormrod (2001) or Creswell and Creswell (2018).

1.3.5 Examine the Data or Other Evidence

As described above, the majority of the remainder of this book covers *data analysis*. This begins in Chapter 6 with **descriptive statistics**, such as the **mean**, **median**, and standard deviation. We then cover testing of hypotheses and exploring relationships through advanced statistical techniques. These include testing a hypothesis about a single mean, two independent means, one-way analysis of variance, chi-squared tests, **linear regression**, and an overview of some more advanced statistical methods. These will be discussed in detail in Chapters 6 through 15.[2]

1.3.6 Write the Research Paper

Once all steps of the research process are completed, you may begin to write your research paper. The typical sections in a research paper are the introduction, the literature review, the method section, the results, a discussion, and the conclusions. Each of these sections is described in Chapter 16 along with examples from published articles. We also review conventional guidelines and style guidelines for reporting statistical results.

1.4 CONCLUSION

This chapter is meant to provide a very brief overview of the research process and where data fits into this process. As mentioned in the "Research Method" section, both primary data and secondary data can be used. Primary data is often used in fields such as sociology, psychology, medicine, and marketing through surveys, experiments, interviews, and observations. Secondary data is common in fields like economics, history, and public policy, where studies employ data from government reports, historical records, or databases. Because both sources of data are important, this book offers examples of each along with chapters that focus on primary data (Chapter 2: Sampling Techniques, Chapter 3: Questionnaire Design, and Chapter 5: Preparing and Transforming Your Data). In Chapter 4, we also offer an example of entering data directly into Stata so that you can better understand the elements of a data set. In other chapters, secondary data sources such as the General Social Survey, the College Scorecard, the National Survey on Drug Use and Health, the OkCupid mobile dating app data, and COVID-19 state-level data are used to generate descriptive statistics and test hypotheses.

In conclusion, this chapter sets the stage for understanding the research process and the role of data analysis, preparing you to effectively utilize both primary and secondary sources in your own research endeavors.

EXERCISES

1. Read the article "Prevalence and Motives for Illicit Use of Prescription Stimulants in an Undergraduate Sample" by Teter, McCabe, Cranford, Boyd, and Guthrie (2005). As you read the article, answer the following questions, which are based on guidelines offered by Greenlaw (2009).
   - **a.** What question or questions are the authors asking?
   - **b.** Describe the theoretical approach that the authors use to develop their research question.
   - **c.** What answers do the authors propose?
   - **d.** In what ways does the current study improve over previous research, according to the authors of the article? In other words, what gaps do the authors identify in the current literature?
   - **e.** What method do the authors use to answer their questions?
   - **f.** What limitations do the authors identify in their study?
   - **g.** What suggestions do the authors have for follow-up research that should be done?

2. Choose a general area of research that interests you. This could be sports, cancer, poverty, social media usage, gaming, and so on. Use the techniques identified in Section 1.2 to narrow your focus as you begin perusing the literature and using forward and backward searching for articles of particular interest to you. Once you have done the initial reading, you should develop a tentative research question and identify five articles that are most closely related to your question. For each of the five articles, answer the following questions:
   - **a.** What question or questions are the authors asking?
   - **b.** Describe the theoretical approach that the authors use to develop their research question.
   - **c.** What is the hypothesis that the authors propose?
   - **d.** What answers do the authors propose?
   - **e.** In what ways does the current study improve over previous research according to the authors of the article? In other words, what gaps do the authors identify in the current literature?
   - **f.** What method do the authors use to answer their questions?
   - **g.** What limitations do the authors identify in their study?
   - **h.** What suggestions do the authors have for follow-up research that should be done?

KEY TERMS

data analysis
dependent variable
descriptive statistics
linear regression
literature

mean
median
questionnaire
statistics

2 SAMPLING TECHNIQUES

CHAPTER PREVIEW

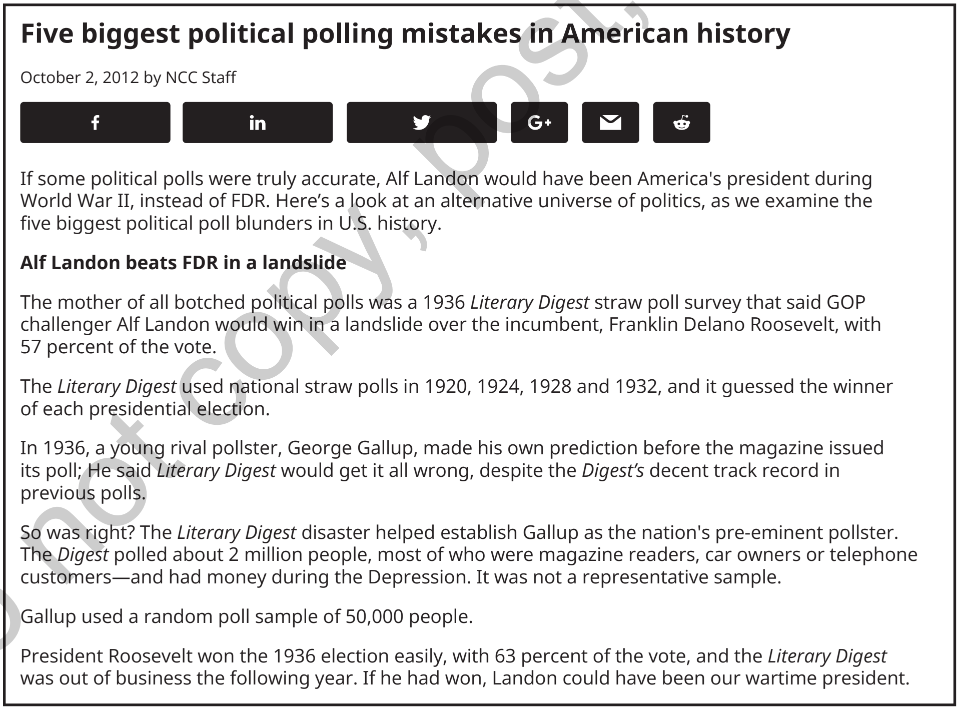
Terms	Definitions
Unit of observation	Type of entity being studied, such as individuals, households, or businesses
Population	The complete set of units that is the topic of a study
Sample	A subset of the population, intended to represent the population, from which data will be collected
Nonprobability sampling	Selection of units based on the discretion of researchers, which means that it is not possible to calculate the probability of selecting each unit
Probability sampling	Selection of units using random numbers, such that it is possible to calculate the probability of selecting each unit
Simple random sample	A sample in which each unit in the population has the same probability of selection
Systematic random sample	A sample in which the selected units are at constant intervals evenly spaced in a list of the units across the population
Multilevel sampling	A sample in which aggregated units (e.g., towns) are selected, followed by the selection of more disaggregated units (e.g., households)
Stratification	Division of the population into different groups, each of which may be sampled differently
Sampling weights	Weights used to calculate population averages in a way that compensates for the effect of the sampling method

2.1 INTRODUCTION

Primary data refer to data collected directly by the researchers. This contrasts with secondary data, which are data collected by another researcher or an organization, such as a government agency. In the social sciences, primary data are often collected through a sample survey, where the researcher interviews (or hires others to interview) a subset of the population on a topic of interest. The quality of the data depends heavily on selecting a good sample and asking the right questions. This was dramatically illustrated by the polling for the 1936 U.S. presidential elections.

As described by the National Constitutional Center in Philadelphia, *The Literary Digest* had run polls in four previous elections, successfully predicting the winner in each. In 1936, they carried out a poll of two million voters and predicted that the Republican candidate Alf Landon would beat Franklin Roosevelt, the Democratic candidate. In fact, Roosevelt won in a landslide, beating Landon in 46 of 48 states. On the other hand, George Gallup used a random sample of just 50,000 voters and correctly predicted that Roosevelt would win (see Figure 2.1).

FIGURE 2.1 ■ Article



Five biggest political polling mistakes in American history

October 2, 2012 by NCC Staff

If some political polls were truly accurate, Alf Landon would have been America's president during World War II, instead of FDR. Here's a look at an alternative universe of politics, as we examine the five biggest political poll blunders in U.S. history.

Alf Landon beats FDR in a landslide

The mother of all botched political polls was a 1936 *Literary Digest* straw poll survey that said GOP challenger Alf Landon would win in a landslide over the incumbent, Franklin Delano Roosevelt, with 57 percent of the vote.

The *Literary Digest* used national straw polls in 1920, 1924, 1928 and 1932, and it guessed the winner of each presidential election.

In 1936, a young rival pollster, George Gallup, made his own prediction before the magazine issued its poll; He said *Literary Digest* would get it all wrong, despite the *Digest's* decent track record in previous polls.

So was right? The *Literary Digest* disaster helped establish Gallup as the nation's pre-eminent pollster. The *Digest* polled about 2 million people, most of who were magazine readers, car owners or telephone customers—and had money during the Depression. It was not a representative sample.

Gallup used a random poll sample of 50,000 people.

President Roosevelt won the 1936 election easily, with 63 percent of the vote, and the *Literary Digest* was out of business the following year. If he had won, Landon could have been our wartime president.

NCC (National Constitution Center). 2012. "The five biggest polling mistakes in U.S. history." National Constitutional Center, Philadelphia. https://news.yahoo.com/news/five-biggest-political-polling-mistakes-u-history-132611721.html

The problem was that *The Literary Digest* relied on lists of "magazine readers, car owners, and telephone subscribers." During the Great Depression, these lists had a disproportionate number of high-income households who opposed Roosevelt and his New Deal policies. In addition, *The Literary Digest* conducted the poll by sending postcards to 10 million voters and relying on respondents to mail back their responses. The response rate was higher among Republicans than Democrats, which also contributed to the incorrect result (Squire, 1988).

The Literary Digest was discredited by this high-profile failure and closed soon after. The success of Gallup's prediction established the national reputation of his firm, which grew to become one of the largest political polling companies. It also catalyzed the development of modern random-sample polling. The lesson for sampling methods is that it is much more important to have a representative sample than to have a large sample. In addition, this experience highlights the fact that a low response rate can distort the results of a survey. Magazine subscriber polls and online polls are not considered scientific or reliable, no matter how many people respond to them.

This chapter introduces the basic concepts of sampling, discusses some of the more common sampling methods, and explains the calculation and use of sampling weights. However, it only scratches the surface of a large and complex topic. Readers interested in a more in-depth treatment of sampling methods may wish to consult Rea and Parker (2005), Scheaffer, Mendenhall, Ott, and Gerow (2011), or Daniel (2011).

2.2 SAMPLE DESIGN

As discussed in the previous chapter, any research must begin with careful consideration of the objectives of the study. What are the research questions? What information is needed to answer those questions? What is the ***unit of observation***, defined as the type of entity about which the study will collect information? In social science research, the unit of observation is often individuals, households, businesses, or other social institutions. Table 2.1 gives four examples of units of observation, depending on the research question and information needed.

In statistics, the ***population*** is the complete set of individuals, households, businesses, or other units that is the subject of the study. Table 2.1 gives some examples of populations corresponding to the studies listed. Note that each population is defined in terms of the type of unit of observation, the geographic scope, and the period of time.

The ***sample*** is a subset of the population consisting of units from which data will be collected. *Sampling* is the process of selecting the sample in a way that ensures it will be representative of the population. One option, of course, is to collect data from every unit in the population—that is, to carry out a census. This might be feasible if the population is defined narrowly or if the budget is very large. For example, if the population is defined as all the banks in a given town, it would probably be feasible to carry out a census. Alternatively, the governments of many countries carry out a population census every 10 years.

But for most purposes, it is more cost-effective to conduct a *sample survey*, defined as systematic collection of data from a limited number of units (e.g., households) to learn something about the population. Using the previous four examples, Table 2.1 provides a possible sample for each.

TABLE 2.1 ■ Examples of Research Questions and Surveys

	Example 1	Example 2	Example 3	Example 4
Research question	Which political candidate is favored by voters?	What is the average yield of rice farmers?	Why do students transfer from one university to another?	How do regulations affect small businesses?
Information needed	The opinions of voters regarding each candidate	The rice production and area under rice cultivation among rice farmers	The reasons that students give for wanting to transfer out	The cost of complying with a set of business regulations
Unit of observation	Voters	Rice farmers	Students	Small businesses
Population	All likely voters in the country, defined as those who voted in at least two of the past three elections	All rice farmers in a country, defined as those growing rice in the previous year	All full-time undergraduate students at the university in a year	All businesses in the state that have 10 or fewer full-time workers
Sample	1,500 likely voters	2,000 rice farmers	200 students	5,000 small businesses
Description of survey	A polling firm collects information from 1,500 likely voters about their political views	A statistical agency gathers information from 2,000 rice farmers to estimate the average yield	A university carries out a survey of 200 students to gather information on reasons for transferring	A state agency carries out a survey of 5,000 small businesses in a state

All surveys face a trade-off between the objectives of reducing cost and increasing accuracy. If cost were no object, then one could carry out a census (covering all units), and it would not be necessary to worry about whether the selected units were representative of the whole group. Alternatively, if accuracy were not a concern, one could just sample a handful of units in one location, which would minimize costs. In practice, most surveys are in between these two extremes. A key challenge is to ensure that the sample is selected in a way that accurately reflects the characteristics of the whole group.

2.3 SELECTING A SAMPLE

2.3.1 Probability and Nonprobability Sampling

How does the researcher select a sample for the survey? One intuitive approach is for the researcher to simply choose a set of units based on availability or subjective judgment. This is called *nonprobability sampling* because it is not possible to calculate the probability of selecting each unit. Below is a partial list of some of the various types of nonprobability sampling:

- *Convenience sampling* involves selecting units from available but partial lists or selecting people who are passing by a location, such as a supermarket.
- ***Purposive sampling*** means that the researcher uses knowledge of the field to select units to be studied.
- *Snowball sampling* refers to picking an initial set of units, then a second round of units that are nearby or have links to the first-round selections. There may be additional rounds.

Nonprobability sampling has the advantage of being quick and inexpensive to implement. It is often used with qualitative research focused on in-depth exploration of a topic on a relatively small number of observations. Qualitative research can complement quantitative surveys in several ways. It can be carried out before a random-sample survey to identify key issues, contributing to the design of the questionnaire. Or it can be conducted after a survey to help interpret the results or explain unexpected findings. For an in-depth discussion of qualitative research and mixed methods that combine qualitative and quantitative research, see Creswell and Creswell (2017).

The main disadvantage of nonprobability samples is that they are likely to be biased, meaning that the sampled units do not accurately reflect the characteristics of the population. (The 1936 polling by *The Literary Digest* is an example.) For this reason, it is not possible to infer characteristics of the population from the characteristics of the sample. For example, a nonprobability sample of businesses will probably include mostly large, well-known businesses—those that have more visible locations and those that advertise. Car dealers, supermarkets, and restaurants will probably be overrepresented, while shoe repair shops, cleaning services, and home-based day care providers are likely to be underrepresented or excluded.

For these reasons, almost all larger surveys carried out by researchers and professional polling companies use *probability sampling*, defined as sampling in which the selection is made randomly from a complete list of units. (Indeed, it is also known as **random sampling**.) The researcher defines the population and the selection method but does not have any discretion in deciding which individual units will be included in the sample.

If a random sample is well-designed and large enough, it will be representative of the population. In other words, the characteristics of the sample will be similar to the characteristics of the population. In the example above, the average size of businesses in the sample will be similar to the average size of businesses in the town. In technical terms, the average business size in the sample will be an *unbiased estimate* of the business size in the population. This means that if you took repeated samples using the same method, the average across samples would converge toward the population average as the number of samples increased.

Another advantage of a random sample is that we can estimate the *sampling error* of our sample-based averages—that is, the error associated with selecting a sample rather than collecting data from every unit in the population. As described in more detail in Chapter 8, the sampling error of a variable is based on (a) the size of the sample, (b) how it was selected, and (c) the variability of the variable in question. If the sample is large or the variability is low, the sample error is likely to be small. One way to describe the sampling error is the *95%* **confidence interval**, defined such that there is a 95% probability that the true average lies between the two numbers. If a political poll reveals that 45% of voters approve of a state governor with a **margin of error** of 3 percentage points, this means that the 95% confidence interval is 45% ± 3 percentage points or 42% to 48%. In other words, there is a 95% probability that this confidence interval contains the true level of approval (if you polled every voter in the state). This topic is discussed in more detail in Chapter 7.

Note that a sample does not have to represent a large percentage of the population to be precise. In national political polls, a sample of 800 to 1,200 is usually sufficient to reduce the margin of error to less than 5 percentage points, in spite of the fact that the sample is roughly 0.001% (or 1 in 100,000) of the total voting population in the United States. It is also useful to note that these calculations count only sampling error. They do not include other sources of error, such as respondents who give false answers or pollsters misidentifying who will decide to vote.

In a large majority of surveys, it is worth the additional effort to select the units randomly. The remainder of this section describes the methods used for different types of random sampling.

2.3.2 Identifying a Sampling Frame

To select a random sample, a researcher needs a *sampling frame*—that is, a list of sampling units in the population from the sample is selected. Ideally, the sampling frame would be a complete list of the units in the population, but this is not always possible. Sometimes an available list is smaller than the target population. For example, a researcher may wish to define the population as all rice farmers in a region, but the available list may include only members of a cooperative of rice farmers, thus excluding rice farmers who are not members. It is important to either complement the list with additional sampling to capture information on nonmembers or recognize this gap in describing and interpreting the results.

Other times, an available list may include more units than the target population. For example, suppose you want to survey likely voters, but the only information available is a list of registered voters, including some who rarely vote. In this case, one option is to contact all voters, ask each respondent if they voted in two of the past three elections, and proceed with the interview only if the answer is yes. Alternatively, the researcher could collect voting patterns and opinions from all registered voters and then examine the patterns for different definitions of *likely voter* in the analysis.

In some situations, no sampling frame is available. This is particularly common when the sampling unit is a specific type of household or business. For example, if a researcher wants to conduct a survey of bicycle repair shops, fortune tellers, or beekeepers in a place where these businesses are not registered, it may not be possible to obtain a complete list to serve as a sampling frame, even at the local level. In such a situation, the researcher must create a sampling frame.

One approach is to use area sampling. The researcher obtains a set of maps of local areas, such as counties or urban neighborhoods. Using maps of each area, the researcher divides it into smaller units of similar size. One common approach is to use a grid to divide the map into equal-sized squares. Another option (relevant for urban surveys) is to use city blocks as the smaller unit. In either case, the researcher selects a sample of the smaller units and then collects information from all the sampling units within the selected unit. For example, to implement a survey of small-scale food shops, the city is divided into 80 neighborhoods using a map, and 20 neighborhoods are selected. Each selected neighborhood is divided into blocks using a street map. The survey team then visits a randomly selected set of eight blocks in each neighborhood. Within each block, every small-scale food retailer is interviewed.

In the absence of maps and a sampling frame, it may be necessary to carry out a listing exercise, in which the survey team first prepares a list of the sampling units within a given area. The sampling units are then numbered, and a random selection is made for follow-up interviews. This can be a time-consuming process, so it is useful to define the area as small as possible given the information available.

2.3.3 Determining the Sample Size

How large should a survey sample be? Not surprisingly, it depends. To explain the factors that determine the minimum sample size, it is helpful to use an example. Suppose we are designing a survey to test whether there is a gender difference in the salaries of recent graduates from a college. Would it be enough to interview 70 graduates, or do we need a sample of 700? To answer this question, we need five pieces of information:

1. How small a difference in salaries do we want to be able to measure? In our example, if we want to detect a male–female salary difference as small as 3%, the sample size will have to be relatively large. If, on the other hand, we are satisfied with only being able to detect salary differences that are 20% or more, a smaller sample will suffice.

2. How much variation is there in salaries? If all the graduates have similar salaries, then we can estimate the mean (average) salary of men and women more precisely, so a small sample would be sufficient. If, on the other hand, there is a wide variation in salaries, then we would need a larger sample to achieve the same level of precision in the estimate.

3. How small do we want to make the probability of incorrectly concluding that there *is* a difference between the salaries of men and women? The larger the sample size, the smaller the risk of making this type of error.

4. How small do we want to make the probability of incorrectly concluding that there is *no* difference between the salaries of men and women? Again, the larger the sample, the lower the risk.

5. How was the sample selected? The sample design influences the size of sample needed to reach a given level of precision.

If we have information (or at least educated assumptions) about these five factors, we can estimate the number of graduates that need to be interviewed in the survey. We will not describe the methods here because they make use of concepts taught in later chapters. However, a brief survey of the methods can be found in Appendix 9.

2.3.4 Sample Selection Methods

This section describes four types of sampling methods: (1) simple random sampling, (2) systematic random sampling, (3) multistage (or cluster) sampling, and (4) stratified random sampling. The Stata code to implement each of these methods is shown in Appendix 7, though it requires a solid understanding of Stata. We recommend studying Chapters 4 to 7 before reading Appendix 7.

2.3.4.1 Simple Random Sampling

Once we have the sampling frame, how do we select the sample? One approach is to select a ***simple random sample***, in which the entire sample is based on a draw from the sampling frame, where each sampling unit has an equal probability of being selected. The probability of selecting each unit is n/N, where n is the number of units to be selected and N is the total number of units in the sampling frame. One disadvantage of a simple random sample is that the selected units may be "clumped" together in the sample frame, resulting in a sample that is less representative than desired. To address this problem, researchers are more likely to use a systematic random sample, as discussed next.

2.3.4.2 Systematic Random Sampling

A ***systematic random sample*** is one in which there is a fixed interval between selected units. First, a unit is randomly selected from among the first N/n units in the sampling frame. Subsequently, units are selected every N/n units. For example, a systematic random sample of 20 households from a list of 200 households starts with a randomly selected unit from the first $N/n = 10$ units. Suppose the random selection picks unit 4. After that, we select every $N/n = 10$ units, that is 14, 24, 34, and so on up to 194. The main advantage is that it spreads out the selected units evenly across the sampling frame. If the sampling frame does not follow any order, this will not make a difference. But typically, the sampling frame is sorted by some characteristic, such as location or size. In this case, a systematic random sample will ensure that the selected units are balanced in terms of that characteristic. For example, if the sampling frame is sorted by location from north to south, then a simple random sample might include a disproportionate number of units in the north. However, a systematic random sample spreads out the sample so that the number of selected units in the north and south will be proportional to the actual number of units in the north and south.

2.3.4.3 Multistage Sampling

Multistage sampling refers to a selection process in which the selection occurs in two or more steps. (This is also called cluster sampling.) For example, suppose we are carrying out a national survey. The researcher may randomly select 10 of the 50 states, 5 counties in each state, and 100 households in each county, for a total sample of 5,000 households. This represents a three-stage random sample, corresponding to the three levels of selection: states, counties, and households.

There are several possible motivations for multistage sampling:

- First, it may be used to overcome limitations on the availability of a full sampling frame. Often, it is not possible to use single-stage sampling because there is no sampling frame that covers the entire population of interest. In the case above, suppose the household lists are available only from county officials. It would be very expensive and time-consuming to gather lists from every county in the country to prepare a national sampling frame for a simple random sample. In contrast, it would be much easier to randomly select a subset of counties in the first and second stages and then get the list for each selected county for third-stage selection of households.
- Second, it may be used to ensure that the sample is well distributed across certain categories. In the example above, the design ensures that the sample includes 10 states and 5 counties within each state.
- Third, multistage sampling may be used to reduce the cost of data collection. Even if a national sampling frame is available, visiting 5,000 randomly selected households would be much more costly than visiting households in 50 counties.

2.3.4.4 Stratified Random Sampling

Stratification refers to dividing the population into categories (or **strata**) and specifying the sample size for each one rather than allowing the distribution to be determined by chance. The strata must not overlap each other, and they must cover the entire population. For example, national household surveys are often stratified into rural and urban areas, with a separate selection of households in each area. National surveys may be stratified by region as well. Surveys of enterprises are often stratified by size, specifying the number of small, medium, and large firms that will be included.

There are three reasons to design a stratified sample. First, stratification may be used to ensure that the sample for each stratum is large enough to allow reliable estimates at the stratum level. For example, suppose a country has six administrative regions, but one of them only has 2% of the national population. In an unstratified random sample of 1,200 households, roughly 2% of the sample (24 households) would be selected from the small region. If the sample is stratified by region, the researcher can ensure that each region has 200 households, which may be enough to generate reliable results for each region. In this case, stratification would be used to oversample the small region, meaning that the percentage of households sampled in the small region is larger than its share in the overall population. The other five regions would be undersampled in this process.

Second, stratification can be used to ensure that each stratum is proportionally represented in the sample. In this sense, stratification fulfills the same function as systematic sampling where the sampling frame is organized by stratum. If the strata are internally more homogeneous than the population, stratification will improve the precision of estimates when compared with a simple random sample.

A third reason for stratification is to adapt to differences in the variability of key indicators across strata. As discussed earlier and as we will discuss in Chapter 8, the precision of survey-based estimates in measuring a variable of interest is partly determined by the variability of the variable of interest. (In the extreme, if there were no variability and all units were the same, a sample of one would be sufficient!) For example, suppose a survey is designed to estimate national income. In general, the variability of income is greater in urban areas than in rural areas. Because of this, it is useful to oversample urban households, meaning that we select a larger share of urban households than rural households. Well-designed stratification can reduce the confidence interval in survey-based estimates without increasing the overall size of the sample.

2.4 SAMPLING WEIGHTS

Sampling weights are numbers used to estimate population **parameters** (e.g., means and percentages) from sample statistics, compensating for "distortions" that may be introduced by sampling. For example, suppose 90% of the population lives in rural areas, but the sample is stratified so that it is 50% urban and 50% rural. In this case, the average income in the sample will be disproportionately affected by urban households. If urban incomes are higher, the average income for the sample will be higher than the average income of the population. In other words, the average income from the sample is biased upward because it has a disproportionately large number of urban households. Using sampling weights, however, we can calculate the weighted average, which will give greater weight to each rural household and lesser weight to each urban household, providing an unbiased estimate of the average income among the population.

2.4.1 Calculating Sampling Weights

Sampling weights are calculated as the inverse of the probability of selection. They can also be interpreted as the number of units in the population that each unit in the sample represents.

In the case of simple random sampling or one-stage systematic random sampling, the probability of selecting any one unit is *n*/*N*, where *n* is the size of the sample and *N* is the size of the population. Thus, the sampling weight (*w*) is calculated as the inverse:

$$w = \frac{N}{n} \tag{2.1}$$

Note that the sampling weight is the same for all units. Such a sample is considered *self-weighted* because the sample average is equal to the weighted average and represents an unbiased estimate of the population average. In this case, the main use of sampling weights is to extrapolate from sample totals to population totals. For example, suppose a survey of seniors at a university collects information on 100 out of 2,000 seniors. The weight is 2,000/100 = 20, so each senior in the sample represents 20 in the senior class. The *average* spending on books in the sample is an unbiased estimate of the average spending in the population. But if you wanted to estimate the *total* spending on books by the senior class, you would just multiply the total for the sample by 20.

In the case of a single-stage stratified sample, we carry out the calculation for each stratum. The weight for stratum i (w_i) is calculated as follows:

$$w_i = \frac{N_i}{n_i} \tag{2.2}$$

where N_i is the population of stratum i and n_i is the sample size for stratum i. Taking the example of urban–rural stratification, suppose there are 900,000 rural households and 100,000 urban households in the population, and the sample contains 4,000 households divided equally between urban and rural areas. The weight for rural households would be 900,000/2,000 = 450, and the weight for urban households would be 100,000/2,000 = 50. In other words, each rural household in the sample represents 450 households in the rural population, while each urban household in the sample stands for just 50 in the urban population. Calculating weighted averages would give more weight to rural households in the sample, thus compensating for the fact that they were undersampled in the survey.

For multistage sampling designs, the calculation of the sampling weights is a little more complicated, but it follows the same general rule: the sampling weight at each stage is the inverse of the probability of selection. There is a separate ratio for each stage in the sampling. Consider the example of a three-stage random sample:

- In the first stage, we select 10 of the 50 states.
- In the second stage, we select 5 counties in each of the 10 selected states.
- In the third stage, we select 100 households in each selected county.

The sampling weight for each county (w_c) is the product of three ratios, each representing the inverse of the probability of selection in that stage of selection:

$$w_c = \frac{50}{10}\frac{C_s}{5}\frac{H_c}{100} \tag{2.3}$$

where 50 is the total number of states, 10 is the number of states selected, C_s is the total number of counties in state s, 5 is the number of counties selected in each state, H_c is the total number of households in county c, and 100 is the number of households selected in each county.

This equation can be adapted to other multistage sample designs, keeping in mind the fact that the number of terms should be equal to the number of stages in the sampling. A simple way to double-check the calculation of the sample weights is to sum the sample weights over the units in the sample. The total should be roughly equal to the number of units in the population.

Up to this point, we have been discussing a type of weight called inverse probability sampling weights (IPSW). The other type of weight is relative sampling weights, defined as the IPSW for each unit divided by the average IPSW. As such, the average value of relative weights is always 1.0. For estimating weighted means and percentages of the population, relative weights and IPSW give the same results. However, relative weights cannot be used to estimate population totals, while IPSW can be used for this purpose.

2.4.2 Using Sampling Weights

How are the sampling weights used? Suppose our variable of interest in a national survey is household income. We can estimate national income as a weighted sum of household income across the sample using the following equation:

$$X = \sum_{i=1}^{n} x_i w_i \tag{2.4}$$

where X is the estimate of the total for the population (e.g., national income), x_i is the value of the variable for household i (e.g., household income), and w_i is the IPSW for household i. As a reminder, $\sum$ is the summation sign, so the right side of the equation means that we should take the sum of $x_i w_i$, as i goes from 1 to n. In other words, $X = x_1w_1 + x_2w_2 + x_3w_3 + \ldots + x_nw_n$.

Estimates of population means can be calculated as the weighted average:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} \tag{2.5}$$

The numerator is an estimate of the sum of x across the population, as shown in Equation 2.4. The denominator is the sum of the weights across the sample, which is an estimate of the total size of the population. Thus, the overall expression is an estimate of the average value of x across the population.

If x_i is a **binary variable** taking values of 0 or 1, then this equation gives an estimate of the proportion of the population for which $x_i = 1$. In the case of categorical variables, such as region or marital status, the average has no meaning, but the variable can be broken up into a set of binary variables, one for each category. Equation 2.5 can be used to estimate the proportion of the population in each category.

However, statistical packages, such as Stata, will do these calculations for us. In Chapter 6, we show how sampling weights can be used to adjust the calculations of totals, means, and percentages in Stata.

EXERCISES

1. Suppose you have a sampling frame of 1,200 hardware stores in a state, numbered from 1 to 1,200. Describe how you would select a systematic random sample of 100 stores for a survey. Give an example of what the sample might look like, showing the store numbers of the first five selected stores.
2. Give three possible reasons why one might want to use a multistage random sample rather than a single-stage random sample.
3. Describe the general circumstances under which it would be useful to apply area sampling to select units to interview. Give an example of a situation in which area sampling would be useful.
4. You have been hired to design a survey of political opinions in 10 swing states, but you need to have a large enough sample (say, 800 respondents per state) to generate reliable results for each state. What type of sampling method do you need to use?
5. Assuming you have a list of all households in each state and can use simple random sampling in each, how would you calculate the sampling weight for each household in the survey?
6. There are 20,000 people in the country of Wakanda. Most of the population (i.e., 17,500) live in urban areas and the rest live in rural areas. If you drew a stratified sample of 250 people from urban areas and 250 people from rural areas, what would be the sampling weights for urban and rural areas?
7. Suppose that we develop a multistage sampling design and choose five states (out of 50), three counties within each state, and 300 households in each county. In the state of Pennsylvania, where there are 67 counties, we randomly select the following three counties (see Table 2.2):

TABLE 2.2 ■ Population of Three Counties in Pennsylvania

County	Population
Montgomery	819,000
Bucks	630,000
Allegheny	1,200,000

Assuming there are three people per household, what is the sampling weight for the selected households in each of these three counties?

KEY TERMS

binary variable
confidence interval
error
estimate
margin of error
parameters
population
purposive sampling
random sampling
sample
simple random sample
strata
stratification
systematic random sample
unit of observation