Chapter 12

Data Collection and Input

Outline

12-1 Introduction
12-2 Data Collection Techniques
   12-2a Primary Data
   12-2b Secondary Data
12-3 Sampling
   12-3a The Population
   12-3b The Sample
   12-3c The Sampling Unit
   12-3d The Sampling Frame
   12-3e Parameters and Statistics
12-4 Sample Designs
   12-4a Probability Sampling
   12-4b Nonprobability Sample Designs
   12-4c Probability Sample Designs
12-5 Determining the Sample Size
   12-5a Population Size
   12-5b Variability
   12-5c Sampling Error and Confidence Level
12-6 Survey Errors
   12-6a Problems in Survey Execution
   12-6b Manipulation
   12-6c The Wording of Questions
   12-6d Low Response Rates
   12-6e No Response
12-7 Coding the Data
   12-7a The Code Sheet
   12-7b The Code Book
12-8 Data Input
Chapter Summary
Chapter Quiz
Suggested Readings

Key Terms

blank foreign elements
cluster of elements
code book
content analysis
convenience sample
election exit polls
field research
finite population
incomplete frame
infinite population
level of confidence
multi-stage cluster samples
nonprobability sample
parameters
population
primary data
probability sample
purposive sample
quota sample
random-digit dialing
sample
sample frame
sampling error
sampling unit
secondary data
self-selection sample
simple random sample
statistics
stratified samples
systematic sample
variability
Chapter 12

12-1 Introduction

Once you have conceptualized theories, operationalized concepts into variables, formulated your hypotheses, and constructed your measurement instruments, you can begin to collect the data pertinent to your research problem so that you can make generalizations. Collecting and inputting your data to statistical packages constitutes the fifth and sixth stages of the systematic research process.

In this chapter we discuss the various techniques political scientists use to collect data. We also cover the fundamentals of sampling theory and discuss the aims of sampling. Then we discuss the central concepts of sampling: the population, the sampling unit, the sampling frame, and the sample itself. We also discuss the different types of sampling designs to include procedures of probability and non-probability. Next, we cover the considerations involved in determining the sample size and for estimating nonsampling errors. Last, we briefly discuss data codification and entering the data into the SPSSW statistical program.

An understanding of this chapter will enable you to

1. Understand data collection techniques.
2. Understand the purpose of sampling.
3. Distinguish between population, samples, and sampling units.
4. See the importance of sampling frames when drawing a sample.
5. Identify errors associated with sampling frames.
6. Differentiate between probability and nonprobability sample designs.
7. Identify sources or errors associated with sampling.
8. Understand the rudiments of determining an adequate sample size.
9. Draw a random, systematic, stratified, and cluster sample.
10. Understand the basics of random-digit dialing sampling and exit polls.
11. Identify sources of errors associated with surveys.
12. Design and use a code book when entering data into SPSSW.

12-2 Data Collection Techniques

Data collection is one of the more interesting phases of political research. When collecting data, you may find yourself in the classroom, in the campus library tracing public issues through the various media resources, surfing the net or in public depositories examining public records and accounts. You may also find yourself attending legislative sessions, interviewing public officials about political issues, or talking with members of the judiciary about legal procedures and points of law. In short, you may find yourself in the field, or the real world of politics.

Whether you opt to collect your own data (primary data) or rely on data collected by someone else (secondary data), the data collection process will be a unique experience because you will have to sort through several sources of information so that you have the data necessary to test your hypotheses. As a result, your final data set may be a compilation of data from several sources. In any event, you will need to tailor the data to meet the specific needs of your research problem.

12-2a Primary Data

Primary data is data that you personally collect. In this section we discuss several of those methods.

Direct Observation: Field Research

Field research is one of several ways that you can personally collect data. One popular way is to personally observe the behavior of your study group and make notes. For example, you can attend legislative, executive, and judicial proceedings.
One obvious advantage to this method is that you observe the behavior of your study group as it actually happens. As a result, you can control the bias that may be involved with other methods such as the clarification of survey questions by interviewers.

This method also has some inherent disadvantages. For example, some phenomena are not directly observable. How do you observe political alienation or efficacy? This method also requires a considerable amount of skill and training in that it may be difficult to quantify observations. In addition, the direct observation of group behavior can take time. Last, direct observation is generally restricted to a small area because you can be in only one place at a time.

**The Survey**
The survey is an expedient method of data collection. It relies on the verbal report of the respondent. Unlike direct observation, you do not actually observe the subject’s behavior.

There are several advantages associated with this method. You can mail surveys to various areas. This expands the scope of your study. This method is also very efficient in terms of time expenditure.

There are also some disadvantages. One obvious disadvantage is the cost of postage. Another disadvantage is the poor return rate of questionnaires. You can enhance the return rate, however, by using a cover letter that explains the importance of the survey and why the addressee should respond. You should also ensure that the instructions accompanying the survey are clear. Other ways to enhance completion and return include limiting the length of the questionnaire. Only include those questions and items absolutely essential to your investigation. You should also pay attention to the way you word your questions. Take steps to ensure that the wording is concise and written so that questions can be understood by those individuals completing the questionnaire. Also include as many close-ended questions as possible. Not only will this enhance completion, but it will also reduce data input and analysis. Another way to enhance completion is to follow up with the individuals you mailed the questionnaires to. In addition, offer to send participants the results of your data analyses. Last, enclose a postage-paid, self-addressed envelope.

**The Personal Interview**
Many researchers prefer to personally interview their subjects. This method allows you to make face-to-face contact, and, as a result, clarify questions if necessary. This method, however, is also very time consuming. In addition, the clarification of questions can contribute to interviewer bias when trying to explain questions.

**The Telephone Interview**
The telephone interview is less time consuming than the methods just discussed. In addition, you can obtain a random-digits telephone listing from the telephone company to use when interviewing. Telephone interviews also enable you to reach more people and reduce travel costs and mailing costs. It also approximates face-to-face contact.

An obvious disadvantage associated with this method is that the interviewee can terminate the interview by hanging up the phone. In addition, many residences have a call-screening system. Consequently, they may not answer a phone call from an unknown or out-of-area caller. The increase in “mass telemarketing” has made people increasingly wary of responding to these callers.
12-2b Secondary Data

Secondary data is data that is collected by someone else. In this section we discuss several sources of secondary data.

Data Archives

Today many public and private agencies collect, store, and release information at little or no cost. Thus, you should consider making use of available data. Possible general sources include data archives. You need to be familiar with the major social science archives. Organizations that specialize in the distribution of data they collect and store include the Bureau of Applied Social Research at Columbia University. The Opinion Research Center housed at the University of Chicago is another important and often-used archive. The Inter-University Consortium for Political and Social Research located at the University of Michigan, however, is one of the most important data archives at your disposal. Every year since 1952, dozens of scholars interested in American voting behavior have made use of the Consortium’s database on the political attitudes and behavior of the American public. The Consortium also processes and distributes data such as attitudinal data, census records, presidential election results, legislative records, and international information applicable to over 130 countries.

The Almanac of American Politics

*The Almanac of American Politics*, which is organized by state and published annually, is a summary of the important political issues pertinent to each state of the union. *The Almanac* also includes important state census data, such as the size of the population and median levels of education. In addition, you can find information about voter registration, racial concentrations, employment profiles, the demographics of state and congressional public figures, and the ratings of all members of Congress developed by organizations such as the Ripon Society, the American Conservative Union, and Americans for Democratic Action. Thus, *The Almanac* is an exceptional source of information about American politics.

Census Data

The nation’s major data collection agency is, of course, the United States Bureau of the Census. One goal of the agency is to make a variety of census reports available to the public and scholars. As a result, you will find extensive data sets about agriculture, the general population, housing, transportation, construction, wholesale and retail trade, and government.

Most of you, however, will find the census data from *General Population Characteristics, Social and Economic Characteristics*, and *Detailed Characteristics* to be most useful. The first publication contains detailed information on race, gender, and household characteristics for each state, county, city, and town (with populations in excess of 1,000 within the states). The second publication provides for the same jurisdictions as the general census but also includes information about employment status, school enrollment, income levels, and social status. The last publication contains detailed social and economic data for states, cities with populations exceeding 100,000 people, and metropolitan statistical areas.

Statistical Abstract of the United States

Another data source produced by the Bureau of the Census that may be quite useful to you is the *Statistical Abstract of the United States*. The *Abstract* combines data collected from both public and private organizations as well as from unpublished
documents. In addition to population information, the *Abstract* contains data on a wide range of topics such as law enforcement, government finances, and state and local governments. In addition, you will find information about international statistics in the *Abstract*.

**Subnational Government Data**

There are numerous data publications available that specialize in subnational governments. For example, the *City/County Data Book* is published every five years to update information on all counties and all cities with populations over 25,000. In this publication you will find population data, employment data, and information dealing with all types of government finances. Crime statistics for states and local governments are also published in the *Data Book*.

Other valuable sources of information on subnational governments include the *Municipal Year Book*, the *County Year Book*, the *Book of the States* published by the Council of State Governments, and the *Census of Governments*. This latter publication provides information for counties, municipalities, townships, school districts, and special districts.

**Voting and Elections**

While there are several data sources dealing with voting and elections, two in particular stand out. The *America Votes* series compiled by Richard Scammon and Alice McGillivray is released every two years following each national election. Included are the results of presidential primaries, state election returns for presidential and senate races by county, state congressional election returns by congressional districts, the results of state party primaries and party run-off elections, and historical state voting profiles since 1946. *America Votes* 23, for example, documents the Republican Party’s unproductive 1998 congressional elections; this was the first time since 1932 that a party not holding the White House actually lost seats. Additionally, the *America Votes* 24 volume records the elections of 2000, in which the presidential winner was not known for weeks after the actual votes were cast. *America Votes* 25 covers the 2002 elections, while *America Votes* 26 will cover the 2004 primary and general elections.

The *Guide to U.S. Elections* is another major source of election information. The *Guide* is a comprehensive compilation of information about America’s electoral process. The latest edition reports the details of the landmark 2000 elections. The *Guide* has been reorganized to include new features such as an expanded discussion of American political parties and their historical development and a chapter about election campaign financing, perhaps the most controversial part of the electoral system. The current *Guide* also includes a chapter that describes the evolution of American elections, with a particular emphasis on the modern period that began with President Franklin Roosevelt’s New Deal political coalition.

**Other Sources**

A very comprehensive source of public opinion data is *The Gallup Opinion Index*, a monthly publication published by the American Institute of Public Opinion. This publication, while publishing information that may be beyond your interests as a political researcher (for example, UFO sightings), also publishes public opinion responses to questions about the president, capital punishment, and other political questions. In addition, the responses are categorized by demographics such as gender, race, party preference, and education. *The Index* also reports trends associated with those questions that were asked on several occasions.
If your area of expertise is international and comparative politics, there are several publications that may meet your needs. For example, the *Statistical Yearbook* is published each year by the United Nations. The *Yearbook* presents demographic, economic, and political data, on a country-by-country basis. Other annual publications include *The International Year Book* and *Statesman's Who's Who* and *The Statistical Abstract of Latin America*. Each publication provides economic, demographic, and political information. *The International Year Book* also provides brief biographical statements about political statesmen. Another publication published by The Network on Urban Research in the European Union, a non-profit-making international association of institutional organizations and individual members founded in Pisa, Italy, in 1989, features extensive data for over 900 cities around the world.

**Content Analysis**

Content analysis is “...any technique for making inferences by systematically and objectively identifying specified characteristics of messages” (Holsti 1968, 601). Many political scientists use content analysis to systematically reduce a text, such as legislative minutes and media products, so they can determine the presence and frequency of some characteristic relevant to their study.

While content analysis is not a method extensively used by political scientists, it can be appropriate in some analyses. An interesting application of the technique is Segal and Cover’s attempt to “...derive independent and reliable measures of the values of all Supreme Court justices from Earl Warren to Anthony Kennedy” (Segal and Cover 1989). They performed a content analysis of the ideological values for the justices from editorials published in several of the nation’s leading newspapers. They coded the statements made within the editorials as liberal, moderate, and conservative. Liberal statements, for example, included those professing support for the rights of defendants in criminal cases, women and racial minorities in equality cases, and the individual against the government in privacy and First Amendment cases. Conservative statements were those with a contrasting view. Each justice was then given a score derived from the codification process. Values ranged from a +1.0 (unanimously liberal) through 0.0 (moderate) to –1.0 (unanimously conservative). Segal and Cover found that these values correlated highly with the votes of the justices for cases dealing with equality and government intrusion into privacy issues (Segal and Cover 1989).

Content analysis has also been used to research the development of international relations (Ithiel de Sola Pool 1981) and to identify the authors of several of the Federalist Papers (Mosteller and Wallace 1964). In summary, content analysis is used to enhance the analysis of nonstatistical information found in documentary, newspaper, and archive records.

**Using the Internet**

An immense amount of information of interest to political and policy researchers is available to users of the Internet. Students and practitioners of political science have increasingly made use of computer networks and the Internet to gather political information. In brief, the Internet links the governors and the governed, links the populace with interest groups, provides a political discussion forum, can be used as a public opinion gauge, and can enhance the democratic process. More specifically, users of the Internet can contribute to political campaign coffers and correspond with their governmental representatives, join interest groups, express their opinions about political issues, and learn about recent U.S. Supreme Court decisions (Davis and Owen 1998, 116–130).
The Department of Defense created the Internet during the late 1960s primarily to serve university and government individuals. Today, access is available to practically all computer users. University students and faculty have ready and easy access to the Internet. Others use relatively inexpensive service providers to access information of interest.

Appendix II (“Political Science World Wide Web Resources Page”) depicts Internet sources that provide information of interest to political researchers, including much of the information we discuss in this chapter. We used the resources in an Internet website page to provide political science students with a beginning point to explore the political content of the World Wide Web. Students can use the page for introductory political science classes as well as upper-level courses in political science (find the page online at http://www.angelo.edu/dept/government/resources.html).

12-3 Sampling

Now that you know something about data sources and ways to collect data, you need to understand the notion of sampling. Often your goal as a researcher is to resolve your research question by observing a group of states, people, or nations. For example, you observe all U.S. congressional members in order to understand legislative behavior, or all nations belonging to the North Atlantic Treaty Organization (NATO) in order to understand international alliances. The entire U.S. congressional body and the entire membership of NATO are known as populations. Each member, or unit, of the population is known as a case. Your goal as a researcher is to generalize about the population from the data you collect.

Typically, however, generalizations are not based on data derived from all the observations, all the respondents, or all the events that are defined by the research problem. Instead, you use a relatively small number of cases (sample) and make inferences about all the cases (population). As an example, pollsters use the responses of a relatively small group of respondents to forecast how the entire population of voters would vote if the election were held at the time the poll was taken. They also use these samples to predict how the population of voters will vote when the actual election is held.

Empirically supported generalizations are usually based on partial information. Why? It may be impossible, impractical, or extremely expensive to collect data from all the potential units of analysis encompassed in the research problem. Yet you can make precise inferences about the population based on a sample when the sample accurately represents the relevant attributes of the population. For example, in marketing research, the preferences of a small sample of households are used to target new products to millions of customers. The Environmental Protection Agency uses a small sample of automobiles of various kinds to obtain data on performance to regulate the performance of all automobiles. In your research efforts you will use a small number of respondents to determine attitudes toward policy, reasons for party identification, and reasons for voter turnout that you can generalize to a larger population.

It is important that you understand what we mean by population, sample, and case. Therefore, we want to spend some more time discussing and presenting examples about these important concepts in empirical research.

12-3a The Population

As we said, the population is the total set of items that you want to analyze; for example, all Third World nations, all individuals who served on the federal judici-
ary, or all Texas cities that have at least 25,000 residents. In other words, the population is the aggregate of all cases that conform to some designated set of specifications. In the first example, by the specifications nations and Third World, we defined a population consisting of all nations of the Third World. For the second example, by the specifications individuals and federal judiciary, we defined a population consisting of all individuals who have served in the federal court system since its inception. And, for the third example, our population is all Texas cities having at least 25,000 residents. You can similarly define populations consisting of all households in a given community, all the registered voters in a particular precinct, or all the books in a public library. Thus, a population is not necessarily a group of people. It can be a group of nations, a group of cities, or even a grouping of events such as cultural or social conflicts.

Populations are the aggregate of all cases that conform to some designated set of specifications. Thus the population has to be defined in terms of content, extent, and time. For example, all registered voters (content), living in Texas (extent), on the date of the gubernatorial election (time).

**Finite and Infinite Populations**

A finite population is a population that contains a countable number of sampling units, or cases; for example, all U.S. Congressional members serving in the 107th Congress, or all registered voters in a particular city in a given year. Sampling designed to produce information about particular characteristics of a finite population is usually termed survey sampling. An infinite population is a population that consists of an endless number of sampling units—an unlimited number of coin tosses, for example.

**12-3b The Sample**

A sample is a subset of all the observations or cases covered by your research question. It is a portion of the population. For example, a portion of all the nations of the Third World, a portion of all the justices that have served on the U.S. Supreme Court, or a portion of Texas cities having a population of at least 25,000 residents.

**12-3c The Sampling Unit**

A sampling unit is a single member (case) of a sample; for example, Nigeria, Chief Justice John Marshall, and San Angelo, Texas. As with the population, a sample unit is not necessarily an individual.

**12-3d The Sampling Frame**

After you have defined the population, you need to select a sample frame that adequately represents your population of study. A sample frame is a list of the units in the population. The actual procedures involve a selection of a sample from the list, or sample frame. Your sample frame may be based on congressional records, the census, city directories, or university student records.

In an ideal situation your sample frame will include all sampling units in the population. In reality, however, such a list does not always exist; thus you may have to use a substitute list that contains the same information but may not be as inclusive. Frankfort-Nachmias and Nachmias (2000), for example, wrote about the problems that the U.S. Bureau of the Census encountered when trying to collect data about the nation. Despite Bureau efforts, it is estimated that the census omitted at least five million of the nation’s residents (Frankfort-Nachmias and Nachmias 2000, 165).
Every aspect of the sample design, the population coverage, the stages of sampling, and the actual selection process is influenced by the sample frame (Frankfort-Nachmias and Nachmias 2000, 165). Therefore, before you select your sample, you need to evaluate the sample frame for potential problems.

**Missing Elements**

One problem that you may encounter is an **incomplete frame**. This problem occurs when sample units included in the population are missing from the sample frame list. An outdated frame could contribute to this problem. You will encounter this problem, for example, if you use a list of senators and representatives who served in the 107th Congress to create a sample of the 108th Congress. Thus, you need to ensure that your sample frame is current.

**Blank Foreign Elements**

The **blank foreign elements** problem occurs when the sample units on your sample frame are not a part of the original population. For example, you use a list of the ambassadors who are members of the current United Nations Security Council to select a sample of members who participated in council sessions that met to determine what action should be used against Iraq because of its hostilities toward Kuwait in 1990. These cases should be treated as blanks and omitted from the sample.

**Cluster of Elements**

This problem occurs when the sample units of your frame are listed in clusters, or groups, rather than individually. For example, you decide to use a university registrar’s list to select students to interview. The list, however, is grouped in accordance with the major field of study. Thus, it is possible that you may select too many individuals from a certain major while undersampling another major. The **cluster of elements** problem is another threat to the usefulness of a sample frame.

**Errors in Sampling Frames**

In 1936, the *Literary Digest* conducted a poll to determine if Alf Landon or Franklin Delano Roosevelt would win the 1936 presidential election. The Digest polled almost 2.5 million people. As a result of its poll, the Digest predicted that Landon would defeat Roosevelt in a landslide victory: 57 to 43 percent. If you know your history, Roosevelt turned the tables and won the popular vote by a huge landslide by winning almost 61 percent of the vote compared to a little more than 36 percent of the vote for Landon.

What happened? The Digest sampled more than an adequate number of people to allow prediction. Let’s briefly evaluate their methodology, and their sample frame in particular. The Digest mailed out ten million questionnaires by obtaining addresses from telephone directories and club membership lists. Many of Roosevelt’s supporters, however, were too poor to have phones or belong to private clubs. Thus, the sample frames did not include all possible sample units found in the population. In other words, the frame was incomplete because it excluded many of those sample units most likely to vote for Roosevelt. Consequently, the sample was not representative of the voting public.

12-3e Parameters and Statistics

The population and the sample have measurements you can use to summarize their characteristics. **Parameters** are measures we use to summarize the charac-
teristics of a population based on all items in the population. For example, the average per capita income of all nations of the Third World, the median age of all individuals who served on the federal judiciary, or the median education of citizens residing in all Texas cities that have at least 25,000 residents.

Because of the reasons we discussed, however, we normally use the characteristics of a sample to infer properties about the population from which the sample was drawn. We call these sample characteristics statistics. The major objective of sampling, therefore, is to provide accurate estimates of unknown parameters from the statistics of a subset. In order to accurately estimate unknown parameters from known statistics, however, we need to define the population and collect data from an adequately sized and representative sample.

12-4 Sample Designs

When you draw a sample, you want to ensure that it represents the population from which it is drawn. You have a representative sample when your analyses of the sample units produce results similar to those you would find had you analyzed the entire population. But how do you go about drawing a representative sample? The answer lies in the distinction between probability sampling and nonprobability sampling.

12-4a Probability Sampling

With probability sampling, each element in your population has a known and equal chance to be a part of your sample. The Literary Digest’s 1936 poll would have fulfilled the requirements of probability sampling if the lower income were included in the frame and were given the same chance as the other prospective voters to be included in the sample. A well-designed sample, therefore, assures that, if you were to analyze several samples drawn from a given population, the findings would not differ from the true population figures by more than some specified amount (± 3 percent). A probability sample design makes it possible for you to estimate the extent to which the findings based on one sample are likely to differ from what you would find by studying the entire population. A probability sample design makes it possible for you to estimate the population’s parameters from the sample statistics. Thus, you want to use a probability sample design whenever possible.

On occasion, however, you may need to use nonprobability sample designs. Your study may be limited, for example, because of time constraints or funding. Because nonprobability samples are relatively easy to conduct and explain in comparison to probability samples, we are going to discuss several types of nonprobability designs before we turn our attention to probability sample designs.

12-4b Nonprobability Sample Designs

When you use a nonprobability sample design you cannot specify the probability that each unit of the population has of being included in your sample. In addition, you cannot be assured that every unit has some chance of inclusion. In the Literary Digest’s 1936 sample, for instance, the Digest never determined the voting intentions of lower income voters. There are three major nonprobability designs you can use: haphazard, purposive samples, and quota samples.

Haphazard Sample

When you use a haphazard sample, you select whatever sampling units are at your disposal. A haphazard sample is often called a convenience sample because you
choose people for your sample based on how easy it is to select them within the limits of the situation. Thus, you cannot estimate the representativeness of your sample and the population's parameters. You may, for example, interview the first fifty senators you meet in the halls of Congress who are willing to be interviewed. A common example is the mall survey. When survey takers approach you when you are walking through a shopping mall, you are part of a mall survey.

Another type of haphazard sample is the self-selection sample. With this type of sample, people basically decide for themselves whether they are going to be a part of the sample. For example, when you respond to a media poll by calling a radio station to respond to a question it posed, you are participating in a self-selection sample. These samples are also referred to as selective listener opinion polls, or “SLOP” polls.

There are several problems associated with haphazard polls. First, interviewers often display bias when they approach prospective respondents. Many interviewers, for example, may be reluctant to approach, for whatever reason, certain individuals. Second, there may be time and place restrictions. Possible respondents have to be in the location at the same time the interviewer is at the location. What are the chances, for example, that an interviewer will be able to interview many men if they locate themselves in front of a store that caters strictly to women? Third, call-ins tend to be more educated individuals with higher incomes. In addition, call-ins may be listening to the station because they enjoy that station’s programming. If the program is a political opinion program, the chances are good that respondents may already have a tendency to agree with the program’s host.

Because of these factors, any haphazard sample is unlikely to be representative. Also, information gathered in SLOP polls could be skewed in accordance with the views of the program’s host. So are there any reasons to conduct a haphazard or a SLOP poll? The answer is yes. Time and financial resources from conducting a probability sample may limit you. Also the experience may be fun for you and your participants. The one thing you must remember, though, is that you cannot use the results to infer properties about a population, no matter how many people respond to your poll.

**Purposive Samples**

When you use a purposive sample, you subjectively select your sample units so that your sample appears to represent the population. Purposive samples are often based on some type of trend analysis. Let’s say, for example, that you want to predict the results of a particular political event, such as a gubernatorial election. Your resources, however, are limited. You have neither time nor money to interview possible voters across the entire state. What do you do?

A purposive nonprobability sample may be the answer. To conduct this type of sample, you try to find a particular county or some other geographical division to represent the entire state. That is, you want to find a county, or counties, that consistently vote the way the entire state votes. The underlying assumption that you make is that the subdivisions are representative of their states. Your next step is to interview prospective voters from the subdivisions you determined were representative of the state. The last step is to predict the statewide results based on the interviews you conducted in the subdivisions. While you can see that this is not the ideal way to take a sample, it is not as time prohibitive or costly as a probability sample that would require interviewing voters across an entire state.

With increasing regularity, political consultants have applied this sampling scheme by using focus groups to gain insights into public perceptions of political candidates and controversial issues. The consultants organize a discussion centering on the candidate or on certain political issues. Usually, the group consists of
from ten to fifteen ordinary citizens. Individuals are selected from certain target
groups in the population such as working men, senior citizens, or women voters. The

group may discuss personality traits of the candidate, political advertising,
and other candidate-related issues. Consultants use these groups because they
believe that the sessions capture feelings and perceptions that may not be tapped
by more impersonal telephone surveys. The consultants then can shape campaign
messages to respond to these feelings and perceptions (Bardes et al. 2001, 285).

**Quota Samples**

A *quota sample* is another type of nonprobability sample because every unit of
the population does not have the same chance to be included in your sample. The
objective of a quota sample is to select a sample that is similar to the sampling
population. For example, if you know that the population is 52 percent female,
then your sample should be 52 percent female. If the population is 15 percent
Latino, then the sample should be 15 percent Latino.

An obvious limitation to this procedure is that it can be somewhat tedious. You are
doing a lot of mixing and matching. In addition, you need to have a good
idea about those factors that could influence the dependent variable. Do you have
a quota based on religion, occupation, or income level, for example? You also want
to make sure that you don’t omit an important factor. You do not want to repeat
the mistakes made by pollsters who tried to predict the winner of the 1948 presi-
dential election. Basically, the pollsters used a form of quota sampling to predict
the winner. They certainly did not want to repeat the mistake made by the *Liter-
ary Digest* in 1936. So they considered variables such as gender, age, ethnicity,
income, and residence. They did not, however, establish a quota based on party
identification. How could they? This was one of the factors they were trying to
ascertain. The bottom line was that they must have sampled a disproportionate
number of Republicans, because several polls predicted Thomas Dewey would
defeat Harry Truman (Freedman et al. 1978, 302–307).

**12-4c Probability Sample Designs**

In contrast to nonprobability sampling, probability sample designs enable you to
specify the probability that each sampling unit has of being included in the sam-
ple. Further, when you have a probability sample, you can use probability theory to
make estimates of the sample’s accuracy. Provided that you properly select enough
people to interview, a probability sample will produce results that are within a cer-
tain margin of error a certain proportion of the time (Corbett 2001, 100).

There are four major types of probability designs: simple random samples,
systematic samples, stratified samples, and cluster samples. We discuss each one in
turn.

**Simple Random Samples**

The *simple random sample* is the most basic probability design. It is often incor-
porated into the more elaborate designs we discuss later. A simple random sample
is one in which each element of the population has an equal chance of being a part
of the sample. In addition, you know the probability of selection and inclusion in
the sample. For example, when you toss a perfect coin, you know that the proba-
bility that it will turn up heads or tails is 50 percent. The same notion applies to
simple random samples.

To determine the probability applicable to your study, you divide the size of
the sample (n) by the size of the population (N). Thus, if your population is 535
U.S. congressional members and you want to survey 300 of them, the probability of each congressional member being included in the sample is 300/535 or 56.1 percent.

Normally you will use a computer program or a table of random numbers to select units to include in your simple random sample. Let’s consider this example. You want to conduct a survey of ambassadors to the United Nations. You lack, however, the necessary resources to interview each member. So you obtain a current sample frame listing each ambassador. There are 600 names (N) from which a simple random sample of 400 (n) is to be drawn.

To start with, you need to number the list, beginning with 001 for the first ambassador and ending with 600 representing the last ambassador on the list. You should assign a three-digit number because you have three-digit numbers in your population (001–600). Each number will be unique because no two ambassadors will have the same number.

Now refer to Table 12-1. The table is a selection from a random-digits table. A table of random numbers is a table of numbers where there is no pattern to the numbers. The probability of a particular number appearing at a particular point is the same as the probability of any other number being there. Therefore there is no bias in the distribution. Notice that each column has five-digit numbers. You need to drop the last two digits of each number because your sample frame is made up of three-digit numbers. The numbers in the table represent the possible cases that you will include in your sample.

Perusing the first column you will note that the first ten numbers in the column are 100, 375, 084, 990, 128, 660, 310, 091, 17, and 329 (be sure to drop the last two digits!). Of these ten numbers, which ones represent cases that you will select from your frame as sample units? We hope you picked numbers 100, 375, 084, 128, and 310. Why these numbers? Remember, your sample frame consists of 600 ambassadors numbered from 001 to 600. So you will only select numbers from the random table within that range. Numbers greater than 600 are outside your range (990, 660, 310, 852, 635, and 737). If the number is outside the range of numbers assigned to the population, skip it and go to the next number. So far in our example you would interview the 100th, 375th, 84th, 128th, and 310th ambassadors in the sample frame. Carrying this logic through, you need to select 395 more numbers from the table until you have your sample.

An often-asked question is where do you start on the random numbers table? While we started with the first row and the first column, you can select any starting point, such as the first row of the second column. You can also choose to progress

<table>
<thead>
<tr>
<th>Table 12-1 Selection from Random-Digits Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>10097 32533</td>
</tr>
<tr>
<td>37542 04805</td>
</tr>
<tr>
<td>08422 68953</td>
</tr>
<tr>
<td>99019 02529</td>
</tr>
<tr>
<td>12807 99970</td>
</tr>
<tr>
<td>66065 74717</td>
</tr>
<tr>
<td>31060 10805</td>
</tr>
<tr>
<td>85269 77602</td>
</tr>
<tr>
<td>63573 32135</td>
</tr>
<tr>
<td>73796 45753</td>
</tr>
<tr>
<td>76520 13586</td>
</tr>
<tr>
<td>74296 24805</td>
</tr>
<tr>
<td>09303 23209</td>
</tr>
<tr>
<td>19645 09376</td>
</tr>
<tr>
<td>08015 36147</td>
</tr>
<tr>
<td>76850 34072</td>
</tr>
<tr>
<td>82406 45571</td>
</tr>
<tr>
<td>65692 02051</td>
</tr>
<tr>
<td>47048 05325</td>
</tr>
<tr>
<td>64778 03529</td>
</tr>
<tr>
<td>80959 36473</td>
</tr>
<tr>
<td>24037 54876</td>
</tr>
<tr>
<td>15953 22309</td>
</tr>
<tr>
<td>88676 38311</td>
</tr>
<tr>
<td>98951 36653</td>
</tr>
<tr>
<td>65813 36170</td>
</tr>
<tr>
<td>86799 42614</td>
</tr>
<tr>
<td>73053 74818</td>
</tr>
<tr>
<td>82471 68665</td>
</tr>
<tr>
<td>28468 57548</td>
</tr>
<tr>
<td>83491 33606</td>
</tr>
<tr>
<td>07439 09303</td>
</tr>
<tr>
<td>23403 09732</td>
</tr>
<tr>
<td>86238 85790</td>
</tr>
<tr>
<td>25624 73053</td>
</tr>
<tr>
<td>35273 20344</td>
</tr>
<tr>
<td>88435 39292</td>
</tr>
</tbody>
</table>

in any way you want: down the columns, across them, or diagonally. All you need to do is decide ahead of time on your method of selection and be consistent.

**Systematic Samples**

A **systematic sample** is a probability sample in which cases are selected from a sample frame at predetermined intervals. For example, suppose you want to select a sample of 200 county administrators from a list of the 2,000 county administrators serving in Texas (hypothetical). With systematic sampling, you first need to calculate the sampling interval. This is accomplished by dividing the number of cases on the list (2,000) by your desired sample size (200). Or, \( K = \frac{N}{n} \), where \( K \) represents the sampling interval, \( N \) represents the population size, and \( n \) represents the sample size. Thus, for this example the interval is \( K = \frac{2,000}{200} = 10 \).

After you determine the sampling interval, you proceed through the sample frame and select every tenth individual until you have selected 200 names. Basically, you can start with any name on the frame. You can select the first individual, the 1,000th person, or whatever. If you want to be somewhat systematic, however, you might want to use a random start. To do this you randomly select a number from 001 to 2,000 from a random numbers table. Thus, if number 20 is randomly selected, administrators represented by 20, 30, 40, and so on would be included in the sample. Or you could randomly select a number from 1 to 10 and proceed from there.

Systematic sampling is very useful when dealing with a long list of population elements. Consider the following example. A student has repeatedly read about the liberal viewpoints of college professors who teach political science. Therefore, she decides to survey 400 professors teaching political science at major universities across the country and ask them questions that will help her determine their level of liberalism. The student uses the American Political Science Association (APSA) membership list as her sample frame. She is surprised to discover, however, that there are 4,000 professors included on the list. Although a simple random sample may be selected, it would involve a great deal of work. Therefore, the student decides to use a systematic random sample. Her first step is to determine the sampling interval, which she does by dividing the population size (\( N = 4,000 \)) by the size of her desired sample (\( n = 400 \)). This procedure results in a sampling interval of 10. The student then selects the first record at random from the first ten names listed on the APSA membership list (for example, name 7). Next she selects every tenth name until a sample size of 400 is selected (7, 17, 27, 34, and so on). This method is called a 1-in-10 systematic sample.

Systematic sampling is more convenient than a simple random sample because you do not need a random digits table other than to possibly pick the first case that you will include in the sample. All you need to do is select every kth unit from the sampling frame. In addition, it is more convenient to use with a large population. Unfortunately, there are also disadvantages you need to consider when using this procedure. There may be a systematic pattern in the data that occurs at every kth unit. This could skew or bias your sample. If you are using a personnel list classified by department and position, for example, it is possible that every kth unit is a department head. This might produce a bias because department heads may have more education, income, and other amenities that are not representative of the entire workforce.

**Stratified Samples**

A **stratified sample** is usually used to ensure that different groups of a population are adequately represented in the sample. This procedure enhances the level of
accuracy of the sample when estimating parameters. The goal, for example, is to break the population down into homogenous strata and select from each stratum. When combined into the total sample, a more heterogeneous sample should occur. Again, let’s use an example to enhance your understanding.

A criminal justice major wants to study the impact of gentrification on the level of crime in an urban neighborhood. To do so, the student decides to examine the attitudes of residents toward crime in the community. The student anticipates that the attitudes of new residents may differ from those of residents who have resided in the neighborhood for an extended period of time. Therefore, as a means of assuring proper representation of both groups, he decides to use a proportional, stratified random sample with two strata: new and long-time residents. The population consists of 2,000 new residents and 3,000 long-time residents. From this population the researcher wants to sample 1,000 residents. The first step for our student is to develop separate lists for each stratum. Then he needs to select the proportional sampling fraction by dividing the population by the size of the sample (5,000/1,000 = 1/5). Next, he should determine the proportion of new residents and long-time residents to be sampled. He does this by multiplying the proportional sampling fraction (1/5) by the number of new and long-time residents. As a result he will survey 400 newcomers (2,000 * 1/5) and 600 long-time residents (3,000 * 1/5). Last, our student can use the simple random sampling methodology we discussed and apply it separately to both lists.

The example and procedures just discussed pertain to proportional stratified sampling. There are times, however, when you might want to use a disproportional stratified sample. Disproportional samples allow you to compare two or more particular strata or a single stratum more intensively. In addition, they are often used if you are concerned that a proportional sample will not give you enough cases to examine from particular strata or groups. This often occurs when examining race as an impact on some political phenomenon. That is, if you use the other types of sampling we discussed, you may not have a sample that has a sufficient number of African Americans or some other racial group in your sample so that you can do more intense analyses. So you would want to use a disproportional sample. Researchers with a major polling group, such as the National Opinion Research Center General Social Survey, will often use this technique.

The steps you take to draw a disproportional sample are similar to those we told you to use with proportional samples. The major exception is that you use several sampling fractions (a different one for each stratum). For example, you believe college students’ majors impact their voting behavior. To test your notion you decide to survey 600 of 3,000 students. In addition, you stratify your sample as 2,100 liberal arts majors, 600 business majors, and 300 agriculture majors. The problem is that you don’t think you will get enough agriculture majors if you use a proportional stratified sample. (How many agriculture majors would you sample?) Therefore, you decide to use a disproportional sample and sample 300 liberal arts majors, 150 business majors, and 150 agricultural majors. If you have followed us to this point, you know that your next step is to calculate the sampling interval for each stratum. Applying our simple formula (N/n), you find you need to survey every seventh person on the liberal arts sampling frame (2,100/300 = 7), every fourth person on the business list (600/150 = 4), and every second person on the agriculture frame (300/150 = 2).

Do you see anything wrong with this procedure? Well, your sample no longer represents the population. After all, you have a disproportionate number of agriculture majors in the sample. Thus, the sample is now biased. So what do you do? Of course, we have an answer. You simply weight the sample.
Perhaps you are concerned about grade point averages (Texans, for example, have all endured the scores of Texas A&M Aggie jokes). As a result of the oversampling of agriculture majors, you are concerned that the sample grade point average is biased downward from the population grade point average. Consider Table 12-2. The table shows that the weighted GPA is 3.0 compared to the sample GPA of 2.83. Our student will use the weighted GPA in his analyses, because he is concerned that the sample GPA may be biased downward because of his disproportional strata sampling technique.

### Multi-Stage Cluster Samples

Multi-stage cluster samples are frequently used because they are an inexpensive way to draw probability samples from a large-scale population. This method requires you to select sample units from several larger groupings through two or more stages until you have a sample of people to interview. You initially select the clusters with a simple random sample or stratified sample. Then you have an option. You can include all the sampling units in each cluster in your sample, or you can select units from the clusters using simple or stratified sampling procedures. In either instance, you will draw from individual population clusters instead of the entire population.

Suppose you are a public administration student who is interested in investigating the relationship between the location of one’s residence and support for an economic development initiative. The initiative, a firefighting school to enhance firefighters’ training, will create several hundred jobs and be a boost to the community’s economy. On the other hand, environmental studies suggest that the smoke from the training exercises could have potential negative effects. You hypothesize that those citizens living closest to the proposed site for the school will be more adverse to its creation.

While designing your study, you discover that there is no single list available that contains the names of every citizen in the community. In addition, it is too expensive to compile such a list. Thus, you decide to use cluster sampling. To use this method, you should take the following steps:

1. Use an up-to-date map to define the area to cover in the study.
2. Define boundaries and exclude those areas that do not include dwelling units. In other words, use only residential areas. These areas make up your primary sampling units, or cluster. For our example, these units could be city census tracts.
3. Number each residential area in a nonpatterned manner. This will compensate for the possibility of an unforeseen bias being built into your sample.

### Table 12-2 Table of Population Proportion and Sample-Weighted GPA

<table>
<thead>
<tr>
<th>Measure/Statistic</th>
<th>Agriculture</th>
<th>Liberal Arts</th>
<th>Business</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of individuals</td>
<td>300</td>
<td>2,100</td>
<td>600</td>
<td>3,000</td>
</tr>
<tr>
<td>Proportion/weight</td>
<td>.1</td>
<td>.7</td>
<td>.2</td>
<td>1.0</td>
</tr>
<tr>
<td>Sample GPA</td>
<td>2.0</td>
<td>3.0</td>
<td>3.5</td>
<td>2.83</td>
</tr>
</tbody>
</table>

Calculation:
- Agriculture: .1 (2.0) = .20.
- Business: .2 (3.5) = .70.
- Liberal arts: .7 (3.0) = 2.1.
- Weighted GPA: .2 + .7 + 2.1 = 3.0.
Select:

Primary cluster of sampling units: [City census tracts, residential areas]  

Second cluster of sampling units: [City blocks]  

Third cluster of sampling units: [Individual households]  

Final cluster of sampling units: [Individuals (registered voters)]

Question: How many sampling frames are needed to use this example?

4. Use a simple random sample or a systematic sample to select the residential areas you will sample from.
5. Divide each residential area into blocks. The blocks will be your second cluster of units.
6. Number each block in a nonpatterned manner.
7. Use a simple random sample or a systematic sample to select the blocks you will sample from.
8. List and number all of the dwelling units in each of the blocks you select. The dwelling units will be your third cluster of units.
9. Number each dwelling unit in a nonsymmetrical manner.
10. Use a simple random sample or a systematic sample to select the dwelling units from which you will interview the residents.
11. Randomly select one person from each household to interview. This particular person should be the only one in the household to be interviewed. If that person is not available, you should return later.

Figure 12-1 depicts the multi-stage clustering process associated with our example.

An inherent problem with this type of sampling design is that sampling error occurs at each stage of the sampling process. To compensate for this problem, you need to select more people in your final sample than if you were using a simple random sample.

Random-Digit Dialing

Random-digit dialing is a way to select a national or statewide sample so that you can conduct a telephone survey. The method uses random numbers for the telephone numbers. Although sets of telephone numbers are available from organizations, there are some inherent problems with using these sets as well as with using a telephone directory. Telephone directories, for example, become outdated quickly. In fact, some are outdated when they are delivered to you through the mail. In addition, some people have unlisted telephone numbers to enhance privacy. These problems, however, can be resolved through the random-digit dialing process.

Let’s pretend that a public policy class wants to conduct a statewide survey in Texas. The class is trying to ascertain the support that Texans have for a proposal to dispose of toxic waste in the vast plains of West Texas. They decide to use random-digit dialing to select their sample. After much research and deliberation, they decide to take the following steps to collect their information.

1. Contact the telephone company to determine how many telephone numbers there are within a particular area code, what the three-digit prefixes are, and how many numbers there are for each prefix.
2. Compute the number of people needed for each area code and prefix.
3. Use a random numbers system for the last four digits of telephone numbers.
4. Dial the telephone number and determine whether it is a residential number. If so, conduct the interview. If not, go on to the next randomly generated telephone number.

For our example, pretend that the class determined that twenty-five numbers are to be selected for area code 914, prefix 944. The class then selects a series of four-digit random numbers to complete the telephone numbers. (Of course, they should select more than the twenty-five they need in case they contact nonresidential numbers.) Continuing with our example, the first four-digit random number is 2468. The telephone number (915) 944-2468 is dialed, the caller determines that it is a residential number, and the caller conducts the interview. This process is continued until twenty-five telephone interviews are completed for the area code and prefix.

In summary, random-digit dialing gives every residence with a telephone number about the same chance of being selected for the interview. Consequently, people with new telephone connections and unlisted numbers could be included in the research effort. One problem with this method is the possibility, albeit slight, that our class could reach a household that has more than one telephone number. Thus, two members of the same household could be included in the project. In addition, telephone polls are often overrepresented with female respondents. Another potential problem is that many residences have a call-screening system. Consequently, they may not answer a phone call from an unknown or out-of-area caller.

**Election Exit Polls**

In *election exit polls*, researchers interview voters as they exit the voting booth. The Voter News Service (VNS) was formed to help television networks to quickly report and explain election results during presidential elections. There are several positives associated with exit polls. First, they can produce immediate results. Second, large sample sizes (approximately 10,000 voters) allow for detailed analysis of subgroups and their voting behavior. Third, researchers know for sure that individuals included in the exit poll actually voted. Last, exit polls are based on probability sampling and voting trends. This enhances the prediction and forecasting of results.

For the most part, exit polls have been reliable and useful. The presidential election of 2000, however, raised some concerns about the usefulness of exit polls, and about the VNS in particular. Box 12-1 describes problems associated with VNS Florida exit polls in the 2000 presidential election and forecasting the victor in Florida.

**12-5 Determining the Sample Size**

As discussed, a sample is a subset or portion of the population you have selected to analyze. But how do you determine the size of a sample? This is one of the most often-asked questions in survey research. The answer depends on several factors.

There are numerous suggestions about the necessary size of a sample. For example, the sample size should be a regular proportion (say 5 percent) of the population. Or the sample size should be at least 2,000 cases. No such rule of thumb is adequate. The size of the sample is properly estimated by determining the size and variability of the population you are studying, as well as deciding what level of accuracy is required and how large an error is acceptable. In the fol-
following sections we discuss the components you need to consider when determining the desired sample size. We also present some examples to help you understand our discussion. You should note that the discussion pertains to only a simple random sample.

12-5a Population Size

The first factor you need to consider is the size of the population you are analyzing. Common sense suggests that the larger the population, the greater the needed sample size. A sample of 50 representatives from a legislative body of 55 will probably accurately represent the entire body. On the other hand, a sample of 50 representatives will probably not be representative of a legislative body of 435 people.
Thus, when you increase the sample size, you reduce the probability that you will select an unrepresentative sample. It is important to note that after a point, however, an ever-increasing sample size will have a relatively small impact on the accuracy of your sample. It is possible to achieve a high degree of accuracy for a large population with a relatively small sample. National polling organizations with vast resources, for example, rarely draw a sample size that exceeds 1,500 to 2,000 individuals. In sum, then, population size is really only important when you are dealing with a relatively small population.

12-5b Variability

The more your population varies, the larger your sample size needs to be. For example, if everyone in Dallas, Texas, is Anglo, female, and a Baptist, how large a sample will you need to draw? We hope you answered one because everyone in the population is exactly the same. You know this example is unrealistic because the characteristics of the citizens of Dallas are varied. That is, you will find males, females, African Americans, Latinos, Anglos, and Asian Americans who are of all religious denominations. Dallas is a multicultural city. Thus, you would need to draw a larger sample so that it will be representative of the people of Dallas.

What we are saying is that you need to consider maximum variability, or variance, in the population. When considering attitude toward a dichotomous variable (say, support and opposition) or the distribution of a dichotomous demographic variable (female and male), for example, maximum variance would be a 50 percent to 50 percent distribution. In other words, the population is evenly divided. In fact, as a practical matter, it is difficult to determine the extent of variability in the population. Consequently, it is wise to assume maximum variability and to select a sample size that will accurately represent a population of evenly divided responses (Yes: 50 percent; No: 50 percent) (Cole 1996, 81). Thus, when considering variability, or variance, we always assume maximum variability.

12-5c Sampling Error and Confidence Level

While the size and extent of heterogeneity of the population are important factors when determining the size of your sample, many consider them to be constants in the sample-size formula and that, as such, they will not differ much from situation to situation. Therefore, the more important factors you need to consider when determining the size of your sample are the extent of accuracy you want when predicting from the sample to the population and the degree of confidence you have in your prediction. The first factor is known as sampling error, while the second factor is known as the level of confidence.

Of course, you would like to be 100 percent accurate and confident when making predictions about the population from your sample data. Greater accuracy and confidence, however, require an increased sample size, which can be costly in terms of money and time expenditure. Thus, you will find that you may need to accept some error and a reduced level of confidence. You may be willing, for example, to accept a sample size that will vary 5 percent or less from the population ninety-five times out of one hundred rather than one that will vary 1 percent or less from the population ninety-nine times out of one hundred. What you lose in accuracy, you gain in time and money.

Based on our discussion we can use the following formula to calculate the appropriate sample size needed for a simple random sample.

\[
n = \left( \frac{CL}{SE} \right)^2 \left[ P(1 - P) \right]
\]
Where n = desired sample size. CL = desired level of confidence. SE = sampling error. P = assumed population variance (maximum = 50 percent or .5).

Before we give you an example using this formula, we need to discuss its components. First, the n represents the sample size you will need to meet certain criteria. Second, the CL represents the level of confidence you want to have when predicting from the sample to the population. The CL is normally a Z score (to be discussed in Chapter 14) which, when used with the concept of the normal distribution, represents a certain confidence level. For example, 1.96 represents a confidence level of 95 percent. This means that ninety-five times out of one hundred times your sample will not exceed the error range (SE). Next, the SE is the amount of error you are willing to accept. It usually ranges from 1 percent to 5 percent. Last, the P represents the assumed population variance, or variability. As we discussed, because you do not know the actual population variability, you assume maximum variability (.5).

Now let’s look at the following example. Suppose you want to determine the attitude of a population about ways to reduce government spending. Also assume you are willing to tolerate an error of 5 percent at the 95 percent level of confidence. What size sample will you need to draw? Using the formula, you calculate that you need a sample size of

\[ n = \frac{1.96^2}{0.05^2} \times (1 - 0.5) \]

Table 12-3 summarizes the various sample sizes you will need as determined by various degrees of error tolerance and confidence levels. The table shows you what happens when you want to enhance your predictive capability and the level of confidence. If you want to have a confidence level of 99 percent (CL = 2.58) with 5 percent error, you will need a sample size of 664. If you are willing to tolerate only an error of 3 percent with a 95 percent confidence level, you will need a sample size of 1,068. In sum, our examples show you that to enhance your predictive capability and lessen your tolerable error you will need to increase the size of your sample.

When interpreting and stating the results of a survey, you will have to consider each of the components we discussed throughout the preceding sections: population variability (.5); confidence level (95 percent); and the amount of error (±3) associated with the survey. We suggest that you use the statement in Table 12-4 to guide you when you interpret and state survey results.

Now let’s apply Table 12-4 by interpreting the following hypothetical poll results: A recent national poll showed that 37 percent of the American public said

<table>
<thead>
<tr>
<th>Error Tolerance (%)</th>
<th>Levels of Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95%</td>
</tr>
<tr>
<td>1</td>
<td>9,604</td>
</tr>
<tr>
<td>2</td>
<td>2,401</td>
</tr>
<tr>
<td>3</td>
<td>1,068</td>
</tr>
<tr>
<td>4</td>
<td>601</td>
</tr>
<tr>
<td>5</td>
<td>385</td>
</tr>
</tbody>
</table>

that they could trust the media to report the news fairly most of the time. The margin of error was ± 3 percent with a confidence level of <.05. Using Table 12-4, you state the following: There are ninety-five chances out of one hundred [1] that in the population from which the sample was selected, the percentage of the American public [2] who said that they could trust the media to report the news fairly most of the time [3] was somewhere between 34 percent and 40 percent [4].

12-6 Survey Errors

There are several sources of errors when conducting and examining survey data. In Section 12-3d, for example, we discussed problems associated with sample frames, such as missing elements, blank foreign elements, and clusters of elements. We also discussed sampling error in Section 12-5c. Sampling error occurs because you are using a sample rather than the entire population. It should be obvious that when using a sample it is impossible to be 100 percent accurate when making inferences about the population. Now, let’s examine some other sources of errors associated with surveys.

12-6a Problems in Survey Execution

When you conduct a survey, you can make mistakes. Your mistakes can cause problems when interpreting the results. Random errors are not a major concern. Systematic errors, however, are more serious (Corbett 2001, 107). Examples include the way you select your sample and the use of an outdated sample frame. Additionally, systematic errors can occur due to the manipulation and dishonesty of those who perform surveys.

12-6b Manipulation

There are many ways that individuals and organizations can manipulate surveys. Consider the following examples reported in the July, 1995, edition of the Reader’s Digest (Barnes 1995, 49–54).

**Example 1**

American trial lawyers claimed that Americans rejected liability reforms that were proposed in the Republican Party’s Contract with America. To support their claim, they used the results of a poll that asked whether “manufacturers should be held less accountable for placing dangerous products on the market.” Seventy-seven percent of the respondents disagreed with the statement.
Business groups, on the other hand, claimed that Americans supported liability reform. Their evidence was another poll that asked whether “volunteer organizations should be exposed to excessive punitive damage awards.” Seventy-seven percent agreed that it was a bad idea. In summary, trial lawyers and business groups used the results, respectively, to support their goals.

**Example 2**

Government agencies also manipulate surveys. Most nonmilitary foreign aid is administered by the Agency for International Development (AID). Despite the fact that AID is currently providing urgent assistance to the people of Afghanistan affected by civil conflict, a debilitating three-year drought, and American military efforts to get the Taliban regime to turn over Osama Bin Laden, economists and policy analysts have criticized the agency for years as an example of an unnecessary federal agency.

In addition, the majority of Americans do not have much knowledge about the activities and responsibilities of the agency, and almost 74 percent of Americans polled since 1972 have responded that federal spending to assist other nations was too high (National Opinion Research Center [NORC] General Social Survey [GSS], 1972–1998). Thus, there is not much political or public support to continue the agency.

AID, however, used the results of a 1995 poll funded and conducted by the Center for International and Security Studies at the University of Maryland to demonstrate the support the American people had for foreign aid and to perpetuate the continuance of their programs and agency. Specifically, AID used the following statements as proof that there is strong support for maintaining foreign aid at current spending levels or higher.

- “Should the United States share at least a small portion of its wealth with those in the world who are in great need?” More than 80 percent responded in the affirmative.
- “Should the United States cut foreign aid by a little, somewhat, a lot, or eliminate it entirely?” Only 8 percent of the respondents selected the last option.

AID did not report conflicting evidence in the same poll. For example, 75 percent of the respondents echoed the results reported in the NORC polls. That is, federal spending to assist other nations was too high. Nor did AID report findings that showed that 83 percent of the respondents believed that very little aid actually reached the needy because of corrupt administrators and governments. In short, the agency reported only results that presented its efforts in a favorable light.

**Example 3**

Advocacy organizations are also skillful at manipulating surveys and poll results. Only a little over 22 percent of Americans polled since 1972 have responded that federal spending through the National Endowment for the Arts (NEA) to assist the arts was too high (National Opinion Research Center [NORC] General Social Survey [GSS], 1972–1998). Conservative groups and politicians, however, have been concerned that the NEA was financing art that most Americans considered pornographic. A poll commissioned by conservative groups revealed that 67 percent of the respondents opposed government support of individuals to promote the arts. Consequently, People for the American Way Action Fund, a liberal lobbying group, commissioned their own survey. The poll preceded a list of questions with flattering statements about the NEA, such as “The goal of the NEA is to foster...”
professional excellence in the arts in America, to nurture and sustain them, and to help create a climate in which they may be enjoyed by the widest possible public.” The follow-up question, and the one used by the organization to counter negative poll results about the NEA, was “Do you think that the NEA wastes taxpayers’ money when it fosters professional excellence in the arts in America?” Predictably, 69 percent of the respondents responded no.

12-6c The Wording of Questions
In addition to the use of manipulative questions similar to the ones just presented, poorly phrased survey questions can also lead to erroneous conclusions. Michael Corbett identifies the following as types of questions to avoid (Corbett 2001, 108–109).

**Leading Questions**
Leading questions are worded so that respondents will respond as intended by the person conducting the survey. Often such questions use adjectives that suggest there is a “proper” response to the question. Consider the following example: “Shouldn’t the government regulate the Internet to protect your children from smut and filth peddlers?” Well, if you put it that way, I guess so.

**Social Desirability**
These questions play upon the conscience of the respondent. In a sense, they are similar to leading questions. Social desirability questions strongly suggest that there is a socially accepted response and respondents select that response to avoid looking bad. For example, although barely half of qualified American voters turn out for elections, polls show that almost 80 percent of Americans believe that voting is a very important obligation. Additionally, polls reflect voting participation in excess of 65 percent (National Opinion Research Center [NORC] General Social Survey [GSS], 1972–1998).

**Vagueness**
Respondents answer vague questions different ways. This, of course, can lead to error. Vague questions also diminish the reliability of a survey because they contribute to inconsistent results. Consider how you would answer the following survey question: “What action should authorities take when a child frequently skips school and the parents don’t do anything about it?” Before you respond, we have another question for you: What does the word *frequently* mean? The word is vague, and people will interpret it differently from one another when responding to this question.

**Double-Barreled Questions**
Double-barreled questions require respondents to choose between two options that are not mutually exclusive. For example: “Do you favor protecting the environment or spurring economic growth?” Many Americans favor both and do not see them in conflict with each other.

12-6d Low Response Rates
Research has shown that people having less education, income, and political interest also have less willingness to participate in a survey. Their refusal to participate in a survey can influence the final results because you may not end up with a representative sample. Thus, it behooves researchers to maximize their response rate.
12-6e No Response

Often, people will not answer particular questions. People may not answer a question because they do not have an opinion about a particular issue, or they do not know the answer to a question. These types of nonresponse are legitimate. We do not want someone responding to a question just to answer the question. To make sure, though, that they have a legitimate reason for not responding, you should consider including “Don’t know” and “No opinion” as possible responses.

Survey error can occur, however, when respondents do not give their opinion because they do not want to respond to a question. People may not give their opinion because they believe the survey taker may perceive their opinion negatively. Avoiding question wording problems can help alleviate this problem. Additionally, skilled interviewers can make the respondent feel more comfortable and less threatened.

12-7 Coding the Data

Now that you are familiar with sources of data, we need to spend some time talking about the assembly, storing, and processing of the data you collect. An important step is to assemble your data in a way that will enhance input and analysis. In some instances, you may be able to analyze data directly from the source; for example, the archive or questionnaire. For the most part, however, you will need to input your data to one of the many statistical packages loaded in mainframe systems or personal computers.

Therefore, before you proceed to the actual analysis stage, you will want to assemble your data in a form amenable to computer processing. As a first step, you will need to code your data. Coding is the process of assigning numbers to all possible responses to all questions or items that will constitute your database. Based on this coding process, you will be able to input and store your data. Before we look at how to construct a code sheet, we need to share some rules of coding presented by Frankfort-Nachmias and Nachmias (2000).

First, code numbers should make intuitive sense for variables that can be rank ordered. For example, higher scores should be assigned higher code numbers (1 = Low; 2 = Medium; 3 = High).

Second, the coding categories must be mutually exclusive. That is, each unit of analysis should fit into one and only one category. The following coding scheme dealing with individual income for the 2003 tax year is not mutually exclusive:

1 = $0 – $10,000; 2 = $10,000 – $30,000; 3 = $30,000 – $50,000.

If you made exactly $30,000 during the tax year, which category do you belong to? Do you belong to category 2 or to category 3?

Third, the coding scheme must be exhaustive. In other words, every response must fit into a category, with few responses being classified as “Other.” Our example about individual income does not include categories for individuals who earned in excess of $50,000 during the 2003 tax year.

Last, your categories must be specific enough to capture differences while using the smallest possible number of categories. You do not want to use a coding scheme dealing with individual income that has categories such as 1 = $0 – $2,500; 2 = $2,501 – $5,000; 3 = $5,001 to $7,500; and so on.

12-7a The Code Sheet

At one time, many political researchers used punch cards to transfer their data to more convenient and permanent modes of storage such as magnetic tape and disk. Today, with the advent of personal computers and the like, many use code sheets.
If you prepare your survey instrument correctly, you can probably input your data into the computer directly from your survey instrument. This will save you much time and effort. Secondary data, however, usually requires the need for a code sheet.

A code sheet is nothing more than a sheet of paper having several columns and several rows. The columns contain the information, or variables, you collected to describe the cases in your sample. Variables such as age, income, and education, for example, will make up the columns. Each row corresponds to each case or sampling unit. You will, for example, have 50 rows if the cases in your study are the states of the union. You will have 535 rows if the cases in your study are congressional members.

The amount of column space assigned to any variable is equal to the largest value for that variable. Data should also be entered as “right-justified” entries. After the code sheet is finished, it is necessary to develop a code book and then input the data to the applicable computer program.

12-7b The Code Book

The code book provides information about the data file you will create in statistical packages such as the Statistical Package for the Social Sciences. It will also tell you what the numbers in the columns represent. Therefore, you need to take some time when putting your code book together.

There are several items that you should include in your code book. First, you want to assign a name for each variable. Most variable names will be self-explanatory (Age = Age). On other occasions you may need to abbreviate the name of the variable because many programs restrict the length of the name to eight to ten characters. For example, you may opt to use “PCI” to represent per capita income.

Second, you want to include a variable label. The label describes the concept measured by the variable. For example, a variable label for “PPID” may be “political party identification.”

Third, you want to depict whether the variable is numeric or an alpha variable, the number of spaces the variable will occupy (the width), and the number of decimal places for the variable in case the variable is measured as a percentage.

If the data you plan to enter is primary data, that is, survey data, you may want to depict the text of the original question as an entry in your code book. You may also want to show the response possibilities (value labels) and the numbers assigned for each possibility. For example: “How do you rate President George W. Bush’s performance in addressing the budget deficit?” 1 = poor; 2 = fair; 3 = neutral; 4 = good; 5 = excellent.

Table 12-5 is part of a code book used by several scholars who were studying the impact of early voting legislation on the county party chairs in Texas counties.

In summary, it is essential that you take time when preparing your code book. You will find that it is a valuable tool when you are entering and analyzing data at a later time.

12-8 Data Input

In the past, political scientists would enter and analyze data through one of two ways: the batch mode or the interactive mode. The first method, for the most part,
required the use of data input cards. The second method allowed the researcher to
directly input data through a computer terminal. Each method, however, was asso-
ciated with a mainframe computer or minicomputer housed in a computer cen-
ter. While efficient and cost effective, these methods had some problems. First,
they required you to wait your turn for data update and printouts because you
were sharing the system with other users across campus. Second, you had to wait
several hours or even overnight to receive your printouts of your input and analy-
sis efforts. Third, many mainframe systems were not “user friendly” because you
had to be versed in the operations of the system as well as the statistical program.

While the mainframe was an efficient and effective system for the university,
it was slow and cumbersome for the individual. As a result, students and
researchers now use personal computers (PCs) and any of a number of statistical
packages to handle their data input and analysis efforts. The major difference
between personal computers and the central site computers just discussed is that
the PC is self-contained. This means that there is no need to have a computer line
linkage between the PC and other units. Although technology may soon render
the PC obsolete, for the time being, it is the most popular type of computer used
by researchers and students.

With the advancement of PCs in terms of cost and effectiveness, many statis-
tical software packages have been developed that can be used with PCs. This has
also contributed to the advancement and popularity of the PC. With the PC and
appropriate statistical software, you can, within a matter of minutes, analyze the
results of your data input and analysis efforts. No more time is wasted in sharing
or waiting in line for output products. In addition, you can make corrections
much more quickly with the PC than with the other types of computers. Last,
many of the major software packages are user friendly because they provide online
direction and assistance. If you have ever worked with central mainframes or
minicomputers, you will appreciate this feature.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Variable Label</th>
<th>Width</th>
<th>Decimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>COUNTY (Alpha)</td>
<td>Name of county</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>PARTY</td>
<td>Political party</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(1 = Democratic; 2 = Republican; 3 = Other)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PARTYORG</td>
<td>Impact on party organization</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(1 = None; 2 = Some; 3 = Great)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHONEBANK</td>
<td>Impact on phone banks</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(1 = None; 2 = Some; 3 = Great)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RALLIES</td>
<td>Impact on rally efforts</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(1 = None; 2 = Some; 3 = Great)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Adapted from Laurence F. Jones, Delbert Taebel, and Nirmal Goswami. “The Politics of Early Voting in Texas: A Research
Chapter Summary

In this chapter we discussed several techniques political scientists use to collect data. Specifically we discussed the steps applicable to direct observation techniques such as field research; mail, personal and telephone interviewing; and survey research. We also discussed secondary data analysis and content analysis. For each technique we discussed the advantages and disadvantages. For example, while the mail questionnaire has advantages such as low cost and anonymity, its disadvantages include a low response rate. Thus, some prefer to use personal interviews or telephone interviews. The personal interview, however, is extremely time consuming and costly. The telephone interview does not ensure that everyone in a population has the opportunity to be interviewed.

Secondary data analysis is used with data collected by others such as researchers or government agencies such as the U.S. Census Bureau. Secondary analysis is useful because it can be used for comparative purposes and to determine change over a period of time (longitudinal analysis). In addition, the collection of secondary data is somewhat cheaper than the collection of primary data.

We covered the fundamentals of sampling theory and the aims of sampling. We said that the central concepts of sampling were the population, the sampling unity, the sampling frame, and the sample itself. We discussed the different types of sampling designs to include procedures of probability and nonprobability.

We also spent considerable time telling you about ways to determine the sample size of a simple random sample. We said that the size of the sample depended on several factors, including the size of the population, the variability of the population, the amount of error you are willing to accept when predicting from the sample to the population, and the level of confidence you want to have when making those predictions. To enhance your comprehension we gave you a simple formula and a quick reference table you could use to determine the size of your simple random sample.

We also discussed sample error and sources of survey errors. For example, we told you that all samples contain sample error because we are not working with a population. We also discussed the problems associated with inappropriate sample frames, the manipulation of surveys by individuals and organizations, and problems associated with the ways questions are worded.

Last, we briefly discussed data codification and ways to enter the data into statistical programs while emphasizing the importance of code sheets and code books. We also discussed why the PC is a very attractive option to researchers and students. Individuals, for example, can see the results of their input and analysis efforts in a relatively short period of time. The statistical software packages used with PCs are also more user friendly than those used with central processing units.

Chapter Quiz

1. A parameter is to a statistic as
   a. collectively exhaustive is to mutually exclusive.
   b. discrete is to continuous.
   c. a sample is to a population.
   d. a population is to a sample.
   e. descriptive statistics is to inferential statistics.

2. In deciding how many people to include in a probability sample of a large population, which of the following factors does not need to be taken into consideration?
   a. Approximately how many people are in the population?
   b. How much accuracy do you need in the results?
   c. How much confidence do you want that your results are actually within the specified range of accuracy?
   d. How much variability is there in the variables?

3. Suppose that a national survey was conducted to determine whether people favored or opposed stricter gun laws. The survey results indicated that 70 percent favored stricter gun laws, 23 percent opposed, and 7 percent were undecided. The accuracy level (or margin of error) for the survey was ± 3 percent and the confidence level was .05. Considering the 70 percent figure for those who favored stricter gun laws, which of the following is the best interpretation of this result?
   a. There are ninety-five chances out of one hundred that in the population from which the sample was selected the percentage who favor stricter gun laws is somewhere between 67 percent and 73 percent.
   b. There are three chances out of one hundred that in the population from which the sample was selected the percentage who favor stricter gun laws is somewhere between 65 percent and 75 percent.
c. There are five chances out of one hundred that more than 3 percent of the respondents gave wrong answers to the question.
d. There are five chances out of one hundred that this survey result is within 3 percent of the results that would be attained if the survey were done again.

4. In the distribution of a dichotomous demographic variable (gender, for example), maximum variance would be a ___________ distribution.
   a. 50 percent to 50 percent
   b. 0 percent to 100 percent
   c. 100 percent to 0 percent
   d. None of choices a through c represents maximum variance in the distribution of a dichotomous variable.

5. __________________ samples are usually used to ensure that different groups of a population are adequately represented in the sample.
   a. Systematic
   b. Stratified
   c. Haphazard
   d. Simple random

6. A(n) ____________________ is a sample in which each person in the population has an equal chance of being selected throughout the selection process.
   a. haphazard sample
   b. simple random sample
   c. multi-stage cluster sample
   d. purposive sample

7. A(n) __________________________ is a probability sample in which cases are selected from a sample frame at predetermined intervals.
   a. haphazard sample
   b. simple random sample
   c. multi-stage cluster sample
   d. purposive sample
   e. systematic sample

8. With a(n) ________________________ sample, people basically decide for themselves whether they are going to be a part of the sample.
   a. SLOP sample
   b. simple random sample
   c. multi-stage cluster sample
   d. purposive sample

9. A(n) ___________________________ is a measure we use to summarize the characteristics of a population based on all items in the population.
   a. statistic
   b. parameter
   c. variable
   d. concept

10. _______________________________ are usually based on either convenience or self-selection.
   a. Haphazard samples
   b. Simple random samples
   c. Multi-stage cluster samples
   d. Purposive samples

---

Suggested Readings
