2

Defining and Framing the Population

Chapter 1 introduced the steps of the sampling process: (a) defining the population, (b) obtaining a frame (or list) of the population from which the sample may be drawn, (c) drawing the sample, and (d) executing the research. The extent to which a sample is exposed to coverage bias, selection bias, and nonresponse bias depends on how well these steps are performed.

This chapter addresses the first two steps in the sampling process, defining and framing the population. In this chapter, you will learn the following:

- How to define the population of interest in operational terms
- Possible sources of frames (or lists) of the population
- The various problems that a frame (or list) may have
- How to solve those problems

2.1 DEFINING THE POPULATION

The first step in the sampling process is to define the population. A *sample* is, by definition, a subset of a larger population. The *population,* or *universe,* is the set of elements about which you would like to draw conclusions.

Before selecting the sample, it is necessary to have a clear idea of the population that you wish to study. Failing to think carefully about the population often leads to samples that are convenient but don't match what is needed. For example:

CASE STUDY 2.1

A common error in the field of market research is using information from existing customers to make judgments about the market. Customers are used because they are easy to get, but people who *do* buy a product can't really give you the viewpoint of people who *don't* buy the product.

Consider this example. A chain of nursing homes measured its public image by interviewing every person responsible for registering a patient at one of the chain's nursing homes. These people were asked how they heard of the home and how they rated it on various dimensions. The results were tabulated every month and distributed in an "Image Report." This report showed the chain's image to be stable during a 6-month period in which there was heavy media coverage about poor care in the homes and during which admissions to the homes dropped sharply.

The problem, of course, is that the chain presumably was interested in its image among all *potential* customers, or perhaps the general public, but the population being studied was all *actual* customers. Since people who heard the bad news stayed away from these nursing homes, the research could not capture the company's image problems.

If this example seems extreme, remember that the *Literary Digest* fiasco was, in essence, a case of studying a convenient population (higher income people who appeared on a mailing list) rather than the correct population (all likely voters). Defining the population in advance allows you to avoid such problems.

To define the population for sampling purposes, two questions must be answered:

- What are the population units?
- What are the population boundaries?

2.1.1 Defining Population Units

The first step in defining a population is to define the units. Is it a population of individuals, households, institutions, business establishments, behavioral events, or what?

The definition of population units for any given research project will depend on the nature of the topic and the purposes of the study. For example, if you are doing a study of voting intentions, then you probably will study individuals, because voting is done at the individual level. For a study of home-buying

intentions, you probably will study households, because homes are a household-level purchase. However, if you are studying home *buying* as opposed to *buying intentions,* then your population might consist of purchase transactions.

It is important to recognize that the data source need not be the same as the population unit. Individuals may speak on behalf of households, companies, and/or sales dollars. This use of proxy reporters does not change the definition of the population. Consider the following example.

CASE STUDY 2.2

An entrepreneur whose business was organizing corporate parties in a large city was considering a new hospitality service, as follows. When a business was expecting visitors from out of town, such as potential employees or their spouses, it could call the entrepreneur's company and arrange to have the visitor(s) taken on a tour of the city and provided with other entertainment. The entrepreneur was confident that she could provide this service better than businesses could do it for themselves. However, she didn't know whether businesspeople would be willing to buy the service, as opposed to doing it themselves to save money or show personal interest in the visitor.

To test the idea, the entrepreneur did a mail survey of all companies that had used her party services in the previous 2 years. This population contained 75 companies. The entrepreneur sent two questionnaires to each company, one to the manager of human resources (HR) and one to the chief executive officer (CEO). Overall response rate to the survey was 62%. Of the people who responded, 46% said that their companies would be interested in using the planned service.

If the desired population consisted of individual respondents, then the 46% figure would be meaningful. However, the customers for this service will be *companies,* not individuals. The data from this survey must be interpreted in some way to express interest at the company level. For example, a company might be counted as being interested only if the HR manager and CEO both express interest, on the theory that a purchase will require interest from both.

Here's an even bigger issue: Assume that these companies have large differences in size and that 7 of the 75 companies account for about 60% of potential revenues. In this case, if the interested companies include the seven key customers, then this service looks promising, regardless of the other companies' opinions. If the interested companies do not include those seven customers, then the service probably is doomed. To reflect this situation, the data should be weighted to reflect the revenue potential for each company. A company that represents 10 times as much revenue as another company should get 10 times the weight.

When you think about it, the desired unit of analysis in this study is *sales dollars,* not people and not even companies. However, dollars (and companies) can't speak. Only people speak, so the data are gathered from people. It is important to remember, though, that these people speak on behalf of the true units of interest, and the results must be treated accordingly.

In some studies—especially large social surveys used for multiple purposes—there may be more than one population unit of interest. For example, the U.S. National Crime Victimization Survey is interested both in the characteristics of *households* that are touched by crime and in the experiences of *individuals* who may have been victimized. Multiple population units can be accommodated in a single study as long as one unit is nested within another, such as individuals within households, workers within companies, expenditure dollars within shoppers, and so on. Such situations are parallel to the case just given. One population unit is selected as the basis for study design—in Case Study 2.2, that initial unit would have been companies, with two individuals selected to report for each company—and weighting is used to express results on the basis of other population units.

2.1.2 Setting Population Boundaries

Once the population units have been defined, the next step is setting the boundaries of the population. *Population boundaries* are the conditions that separate those who are of interest in the research from those who are not. For example, in a study of candidate preferences for a school board election, you might only be interested in people who are likely to vote in the election. Population boundaries may be defined by demographic characteristics (e.g., persons 18 years or older), geography (who reside in the school district), behaviors (who voted in the last election), intentions (who intend to vote in the next election), or any other characteristics that are relevant to the research.

The Need for Operational Specificity in Population Boundaries

The key point in setting population boundaries is to state them in specific operational terms so that everyone can tell who should and shouldn't be measured. "Adults in the Chicago area" is not an adequate definition, because it doesn't tell interviewers whether they should gather data from an 18-year-old in Hammond, Indiana. "Beer drinkers" is not an adequate definition, because it doesn't tell interviewers whether they should interview someone who drank one beer once in his life. The measurement operations that define the boundaries of the population must be clear and specific. Proper definitions of population boundaries take forms such as "people who are at least 18 years of age and have their principal place of residence in Cook County, Illinois" or "people who have drunk beer at least once during the past 3 months," or "people who are at least

18 years of age, reside within the Oak Creek School District, are registered to vote, and say that they 'definitely' or 'probably' will vote in the upcoming election." These boundaries can be translated into unambiguous screening conditions to separate those who are in the population from those who are not.

CASE STUDY 2.3

Sometimes it is easy to define a population in conceptual terms, but difficult to do so in operational terms. For example, in Case Study 2.1, we said that the population of interest was "all potential customers" of a nursing home chain. This definition is easy to understand in conceptual terms but difficult to operationalize. Should we define the population in terms of age? Geography (e.g., proximity to one of the chain's facilities)? Responsibility? Recent purchase? Intention to purchase?

Here is one possible definition: "people who are at least 18 years of age; have their principal place of residence in the 713, 281, 832, or 409 telephone area codes; and who placed a relative in a nursing home facility within the past 12 months."

Here is another definition: "people who are at least 18 years of age; have their principal place of residence in the 713, 281, 832, or 409 telephone area codes; have a relative who is likely to enter a nursing home facility within the next 12 months; and are the person who would have principal responsibility for choosing that facility."

The logic of the first definition is that (1) people who have already been through the decision process are most likely to have formed the opinions of interest in this research, and (2) previous behavior is more solid than intentions. The logic of the second definition is that the group of ultimate interest is the people who will make this decision in the near future. Neither definition includes the people who will actually enter the homes (essentially presuming that the decision will be made for them). Neither definition includes people from outside the area who might have an elderly relative who was (or will be) placed within the area. Both definitions include people from inside the area who might have a relative who was (or will be) placed in a facility elsewhere.

Is one of these definitions better than the other? Would a third definition be better? Before we can answer these questions, we have to resolve issues such as the following: Who makes the decision regarding selection of facility? Does the answer vary, such that some residents place themselves and others are placed by family members? How long before entry do they make that decision? How long before entry do they begin to gather information? Are "out-of-area" placements a small enough segment to ignore? Once we resolve such issues, the results can be translated into operational criteria, but the task will not be simple.

Other Issues in Setting Population Boundaries

In addition to the need for specificity, population boundaries are often set with an eye toward the cost-effectiveness of the research. This might be relevant for our nursing home example: For

example, we might restrict the research to certain telephone area codes that provide a large majority of the chain's admissions, even though it is recognized that some customers from outside the area are missed by this definition. In doing so, we would be making an implicit assumption that the potential coverage bias caused by excluding certain population members is not serious enough to warrant the extra cost of obtaining these observations.

Population boundaries also will implicitly reflect any method limitations. For example, if you do a landline telephone survey, your operational population is limited to people with landlines, whether or not you state that fact in your population definition. If you don't have foreign language interviewers, your operational population in the United States is limited to people who speak English. If you do a mail survey, your operational population is limited to people who can read and write. Also, many studies are operationally limited to adult participants who live in households.[1,2]

In cases where the classification of population members depends on their own reports (as opposed to information from records), the operational boundaries of a population also may be limited by population members' willingness or ability to report qualifying conditions or behaviors. Willingness can be an issue if the qualifying condition is socially sensitive (e.g., being gay or lesbian, having used illegal substances, having been a crime victim, being human immunodeficiency virus [HIV] positive, or even conditions such as having received a traffic ticket, living with an unmarried partner, or having a certain level of income). Ability can be an issue if the condition is unobserved or vague. For example, people who are HIV positive may not know that they are HIV positive. Somebody who has been in a fistfight may not know that it was, technically, a criminal assault.

As a result of these various factors, there is almost always some mismatch between the conceptual population of interest and the actual, operational

1. Many studies are limited to adults to avoid legal or procedural complications associated with gathering data from minors. Also, while it has been our experience that children as young as 6 years can be reliable respondents regarding their own behaviors, a restriction to adults may be based in concerns that younger respondents will not be able to provide reliable information. For example, in research designed to measure the psychological response of New York City schoolchildren to the World Trade Center attack of September 11, 2001, researchers limited the study to students in the 4th through 12th grades "mostly to save time and money" but also because the researchers feared that "it would be more difficult to assess the effects on such young children" with the methods employed (Goodnough, 2002, p. A1).

2. The restriction to households excludes people who reside in group quarters. Group quarters are nonhousehold places where people live or stay in a group living arrangement, including places such as college residence halls, fraternities or sororities, military quarters, prisons, nursing homes, rooming houses, convents, and homeless shelters. About 3.5% of U.S. adults live in group quarters, but the rate is around 10% for people ages 18 to 24 (many at college) and 15% for people age 85 and older (many in nursing homes) (U.S. Census Bureau, 2012, Table 73).

population. The usual direction of error is toward undercoverage of the true population. This is one factor to consider in evaluating the quality of a sample.

2.2 FRAMING THE POPULATION

After defining the population, you must obtain a frame of the population before sampling can begin. A *frame* is a list or system that identifies members of the population so the sample can be drawn. A single study may use more than one frame: for example, in the United States, we might use census data to select places where we will do a study and local directories to select people within those places.

Lists generally are the preferred type of frame. With a computerized file or printed directory that lists the population, it is easy to draw selected members. For example, at most universities, a sample of students may easily be obtained by drawing names from the student directory. However, lists are not always available; for example, if you want to sample visitors to a particular website, to learn something about them and get their opinions on the site, you generally will not have a list from which to sample. In these situations, some sort of counting system must be used to keep track of the population members and identify the selections (e.g., every fourth visitor).

We will begin our discussion of sampling frames by discussing the use of lists and then broaden the discussion to include other types of frames.

2.2.1 Obtaining a List

The first and often most difficult step in using a list for sampling purposes is obtaining the list. Existing lists should be used whenever possible because they are cheaper than custom-made lists. For example, if you wish to do a mail survey of the local population to measure opinions for a friend's school board race, one way to get a list of the population is to send someone door-to-door to record the address of every occupied housing unit. A second way is to use the U.S. Postal Service Master Address File if possible. Obviously, it is much easier and cheaper to use the existing list as a sampling frame.

In fact, the availability of a list—or any other frame—can have a strong influence on the method of data collection used for a survey. For example, a survey of visitors to a particular website would typically be conducted online,

with a pop-up invitation for selected site visitors to participate in the survey, because it is far more efficient to identify members of the population through their visits to the site than, for example, calling people on the telephone and asking if they visited the website. Even if there is some reason why it is necessary to conduct the survey in person or by telephone, we probably would start with an online request to participate and ask respondents to provide the necessary contact information. On the other hand, a survey regarding general Internet usage might very well be conducted by telephone, because there is no efficient way to identify visitors to every site on the web, and it would be relatively easy to identify members of the population by calling people on the telephone and asking if they use the Internet.

We sometimes use the term *special population* to refer to groups for which a frame is readily available. Visitors to a particular website are an example of a special population, as are visitors to physical locations such as parks or zoos. Other examples include students at a college (who are listed in the college's directory), members of an association (identified through the membership list), inpatients or outpatients of a hospital or clinic (identified through patient records), schools or school districts (identified through state educational agency records), voters in a past election (identified through voter rolls), and many business or institutional entities (identified through directories). Such populations are likely to be sampled and contacted through whatever information is available in the frame; for example, if we have telephone numbers but no postal or e-mail addresses, then we are likely to do a telephone survey or at least start with telephone contacts.

Surveys of the "general population"[3] are most often conducted by telephone or mail because sampling frames with good population coverage are available. Regarding telephone, companies such as Survey Sampling International maintain records of listed landline and cellphone numbers, and it is possible to extend coverage to nonlisted households by incorporating random numbers in the sampling process (as discussed later in the chapter). These companies will draw a random sample of telephone numbers for a reasonable fee, nationally or within a defined area, and this is currently the most common way that telephone sampling is done. Nationally, it is possible to reach more than 95% of

3. This includes population subgroups for which no special frame is available, including demographic groups such as men, women, young people, old people, African Americans, whites, high-income people, low-income people, and so on. There are e-mail and postal mailing lists that provide this sort of information, but while mailing lists are useful to direct marketing organizations, they generally do not offer broad coverage. Usually, to get a high-quality sample of a population subgroup, you have to start with a general population frame and screen down to the targeted group.

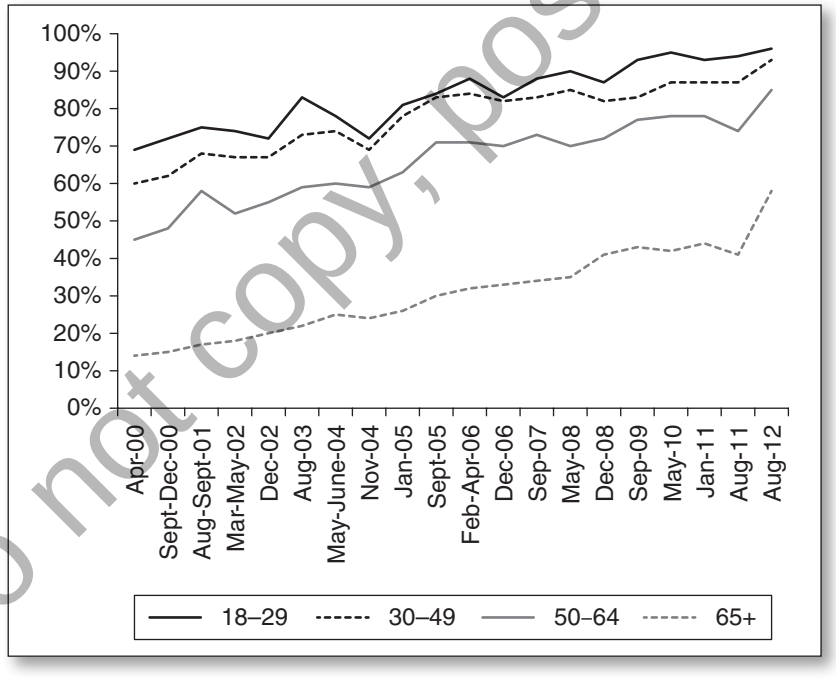
people via telephone (although, as we will discuss later, there are significant issues regarding landlines vs. cellphones).

For mail surveys, the U.S. Postal Service maintains a Master Address File, and this list can be accessed for sampling purposes. In principle, this list covers all households, as well as all other locations where mail is delivered, such as college dorms, residential hotels, and so on.

Online surveys face two coverage problems with respect to the general population. First, only 75% of the U.S. population currently uses the Internet (U.S. Census Bureau, 2012), and the coverage problem is more severe for some population subgroups. The figure drops to 68% for African Americans, 62% for Hispanics, and only 34% for those with less than high school education. Exhibit 2.1 shows that while Internet usage has grown among all age groups and is virtually universal among younger Americans, older age groups lag behind.

The second and more severe coverage problem for online surveys of the general population is that there is no general directory of e-mail addresses. As a

Exhibit 2.1 Internet Usage Differs Across Age Groups



Source: Data from Pew Research Center "Internet Use Over Time" Pew Research Center, Washington, D.C. (Published May 2013) http://www.pewinternet.org/data-trend/internet-use/internet-use-over-time/, Accessed on 9/23/14.

result, online surveys of the general population must rely on more limited sampling frames.

One such option is *opt-in panels* offered by a variety of companies such as YouGov. These panels consist of people who have agreed to respond to online questionnaires in return for incentives. Panelists are recruited by means of banner ads on websites, e-mail, direct mail, or word of mouth from other panelists (Baker et al., 2010). Samples drawn from these panels often are subjected to geographic and demographic quotas—that is, their distribution on geographic and demographic variables will be designed to match the U.S. Census—but this is not the same as saying that the samples are randomly drawn from the general population. The panels may have millions of members, but this is only a small fraction of the total population, so while users of these panels find them satisfactory for practical purposes, the theoretical potential for coverage bias is high. One interesting option is the KnowledgePanel (formerly Knowledge Networks) maintained by GfK Research; the premise of this panel is to recruit a random sample of the general U.S. population and provide them with incentives in return for participation in web surveys.

A less expensive option for researchers on a budget is Amazon's Mechanical Turk (*MTurk*), which has a large panel of people who participate in surveys, experiments, or other tasks in exchange for small payments. However, the MTurk panel cannot be viewed as a cross section of the general population, MTurk studies do not allow the outbound sample controls that are typical with the panels described above, and MTurk participants are volunteers who choose studies rather than being chosen, so MTurk samples have arguably more exposure to coverage and selection bias.

A third possibility for online surveys is to recruit respondents through venues such as social media and online forums. Participants in these venues are likely to differ in some ways from the general population, and further bias may be possible to the extent that respondents are attracted by the study topic (in contrast to opt-in panels or MTurk, where respondents are drawn by incentives that are unrelated to the topic of the research). However, social media may have useful sampling purposes, as discussed in Chapter 8.

Overall, online surveys of the general population face substantial frame limitations and, as a result, must be viewed as having some form of nonprobability sampling. This means that estimates drawn from these studies must rely on some form of model-based estimation, which we will discuss in Chapter 7. However, the samples obtained in online research may be as good as or better than the alternatives being considered; for example, a study conducted with a

sample of MTurk panelists may be at least as defensible as a study conducted with a sample of college students. Also, online surveys are wholly appropriate for special populations that are found online, such as visitors to a website, and for special populations that have a list of e-mail addresses. Finally, online samples may be good enough for the purposes at hand. We discuss the question "How good must the sample be?" in Chapter 9.

2.2.2 Problems With Lists

Having obtained a list, you must deal with problems in the list. Ideally, a list will contain every member of the population once and only once (i.e., there will be a one-to-one correspondence between the list and the population). Few lists meet this standard, and the sampling procedures must compensate for the list's deficiencies.

It is useful at the outset to point out that almost all lists one is likely to use for sampling have been constructed for some other specific purpose, such as organization membership lists, directories of professionals (such as physicians), or for the purpose of a business service to the population, such as listed telephone numbers. When this fact is noted, it becomes clear why certain characteristics of a list or changes in a list may be of concern for sampling: The processes that produce or change these list memberships do not occur at random but are related to the list's purpose. For example, if we consider the characteristics of categories of physicians who are included on a list, are in the process of being added to a list (newly licensed physicians), or are listed but should not be (deceased), it is clear how a sample chosen from that list may differ from the target physician population of interest and, furthermore, how those differences may be related to the survey variables about which we plan to collect information.

There are four ways that list and population elements can deviate from one-to-one correspondence:

- First, there can be population members that aren't listed. This is called *omission,* because the population element has been omitted from the list.
- Second, there can be listings that aren't in the population. This is called *ineligibility,* because the listed element is not a member of the population and therefore is ineligible for selection into the sample.

- Third, there can be two or more listings corresponding to a given population member. This is called *duplication,* because the population member is duplicated in the list.
- Fourth, there can be two or more population members corresponding to a given listing. This is called *clustering,* because the list member corresponds to a cluster of population elements.

The four list problems and their effects can be illustrated by thinking about a telephone directory (i.e., a list of landline telephones) as a list of adults, as follows.

- The directory *omits* people who have moved to town recently or have unlisted telephone numbers, no telephone, or only a cellphone not covered by the directory. As a result, these people will be omitted from any sample drawn from the directory. This will produce a sample that underrepresents groups such as new residents, schoolteachers with unlisted phones, poor people with no phones, and young people without landlines. More generally, it will create potential coverage bias due to undercoverage of the population.

- Many telephone directory listings are for businesses. These listings are *ineligible* for a sample of households or individuals. Furthermore, if we are interested in a population such as people who are likely to vote in an upcoming election, the telephone directory will contain many listings for people who are not registered or not likely to vote and hence *ineligible* for the purpose at hand. If these ineligibles are included in the survey, they may skew the results, resulting in coverage bias due to overcoverage.

- Some self-employed people such as doctors and lawyers may have two or more listings, without any clear indication that either is an office listing and therefore ineligible. If something is not done about these *duplications,* then the sample will contain a disproportionately large number of professionals. To see why this is true, consider the fact that selections are drawn from the list, not directly from the population. If professionals make up 5% of the population but 10% of the list, then a random sample from the list will be 10% professionals. This is a form of selection bias.

- Most households only have one landline listing, regardless of how many adults live in the household. Adults who live in two-adult households will have

only half the chance of being selected that single adults will have. This is because the *clustered* population members (such as married couples) must share one listing and therefore one chance of selection, while the single adult does not share a chance. To put it another way, if single adults make up 5% of the population but 10% of the telephone listings, then 10% of the sample will be single adults. If something is not done about clustering, a sample from the telephone directory will overrepresent single adults, many of whom are relatively young or relatively old. Again, this is a form of selection bias.

Fortunately, even though frame problems are unavoidable in most studies, we have methods for dealing with them, as shown in Exhibit 2.2 and discussed below.

Exhibit 2.2 Snapshot of Problems With Sampling Frames

Problem	*Picture*[a]	*Result*	*Possible solutions*
Omissions (population elements omitted from the list)	P –	Possible coverage bias from undercoverage of omitted elements	• Ignore omissions • Augment the list • Replace the list
Ineligibles (list elements that aren't in the population)	– L	Possible coverage bias from overcoverage of ineligible elements	• Drop ineligibles and adjust sample size
Duplications (multiple listings for population elements)	P < L, L	Possible selection bias from overrepresentation of duplicated elements	• Cross-check the list • Subsample duplicates • Weight the data
Clustering (multiple population elements per listing)	P, P > L	Possible selection bias from underrepresentation of clustered elements	• Take the whole cluster • Subsample within clusters • Weight the data
Ideal (one-to-one correspondence between population and list elements)	P – L		

a. P represents a population element and L represents a corresponding listing.

2.2.3 Coping With Omissions

The most common way of coping with omissions is to ignore them—or attempt to repair them[4]—and hope that the resulting bias is not serious. This approach usually makes sense if the list contains over 90% of the population and does not omit important subgroups. As list coverage declines, however, it becomes more and more necessary to compensate for list deficiencies, on the assumption that omissions do not occur at random; omitted elements are likely to have some characteristics in common. Thus, as the list's omissions increase, so may the potential for introducing bias into a sample selected from it, even though at the time of sampling, one may not know exactly why.

We will discuss various ways of compensating for omissions:

- Random-digit dialing
- Incorporating cellphones
- Address-based sampling
- Registration-based sampling
- Half-open intervals
- Dual-frame sampling

Of course, an evaluation of list coverage requires a reliable outside estimate of the size of the population. Estimates of this type often are made on the basis of U.S. Census reports, but sometimes it is necessary to do a test study to find members of the target population and determine what percentage of them are listed.

Random-Digit Dialing

Many residential landline numbers are unlisted, especially in large cities and among certain types of people. To avoid the coverage bias that might result in telephone surveys, researchers use random-digit dialing (RDD) to reach unlisted as well as listed numbers. *Random-digit dialing* is the dialing of random

4. In addition to pure omissions, lists may suffer from incomplete or incorrect information that makes some listings inaccessible, which effectively turns them into omissions. For example, a list might have outdated addresses, or street addresses without telephone numbers or names without e-mail addresses. If the problems are not so severe as to make the list useless, we proceed to draw the sample, and for selected elements that lack the needed contact information, we turn to Google, social media, telephone calls, or whatever means possible to get that information.

numbers in working telephone exchanges so that unlisted numbers can be included.[5]

Researchers who use RDD for the first time often ask, "Won't people with unlisted numbers be upset about the call and refuse to participate?" Actually, cooperation rates for people with unlisted numbers are roughly as high as for people with listed numbers. This also is true for people who have registered for the U.S. Do Not Call list. Most people don't get unlisted or Do Not Call numbers to avoid research interviews, and most don't mind if they are called for this purpose. Some respondents will ask the interviewer how he or she got the number, so interviewers should be given a description of the sampling procedure to read to those who ask, but this is good practice in any survey.

The problem with random dialing is that, while it reaches unlisted numbers, it can be costly because it also reaches a large number of business and nonworking numbers. Several procedures are used to reduce the fraction of unusable numbers obtained in random dialing. A method developed by Mitofsky and Waksberg (Waksberg, 1978) starts with a screening call to a number selected at random from a working telephone exchange. If this first number is a working number, additional calls are made within the bank of 100 numbers until a prespecified number of households, usually three to five, are obtained. For example, suppose the number dialed is 217-555-1234 and it is a working household number. Additional calls will be made at random to the bank of numbers between 217-555-1200 and 1299 until the desired number of households is obtained. If the initial number is not a working number, no additional numbers will be used from that series. The major saving from this procedure is that banks with no working numbers are quickly eliminated. Mitofsky-Waksberg sampling usually yields about 50% working household numbers.

A drawback of the Mitofsky-Waksberg procedure is that it is cumbersome to implement. Having hit a working household number in a telephone bank, you will seek some number of additional households from the same bank: say, for purposes of example, three additional households. So you draw three random numbers within the bank and call them. Not all of the three will be working household numbers, so you replace the nonworking numbers. You continue to repeat this process until you have the desired number of working numbers. The process can be time-consuming, with a lot of back-and-forth between data collection and sampling.

5. This section focuses on procedures for sampling residential landlines. Cellphones are discussed in the next section.

A simpler and more common form of RDD is *list-assisted sampling*. Companies such as Survey Sampling International maintain a list of all listed telephone numbers in the United States, sorted into numeric sequence. These companies also have compiled a list of business telephone numbers and can eliminate known business numbers by matching the two lists. The resulting list is used to identify telephone banks that are known to have at least one working household number and to eliminate banks that have no listed household numbers. A sample of random telephone numbers is then drawn from the retained banks. The sample can be completely random from the retained banks or can be allocated in proportion to the number of listed numbers in each bank.

Exhibit 2.3 shows how list-assisted sampling works. Imagine that there are four telephone banks, each with 100 possible numbers. Of the 100 possible numbers in Bank A, 80 are working household numbers with 60 listed and 20 nonlisted; in Bank B, 60 are working household numbers with 40 listed and 20 nonlisted; in Bank C, 4 are working household numbers with none listed and 4 nonlisted; and in Bank D, none are working household numbers. Across the four banks, 144 of the 400 possible numbers (36%) are working household numbers, with 100 listed (69.4% of the 144 working household numbers) and 44 nonlisted (30.6% of the working household numbers). Now, assume that we will draw a sample of 100 numbers from these telephone banks. Here's what we will get under various systems:

- If we use simple RDD without list assistance and simply draw 100 random numbers in the four banks, we expect to draw 25 numbers per bank. Since 80% of the numbers in Bank A correspond to working household numbers (60% listed and 20% nonlisted), we expect 20 of the 25 selections to be working household numbers (15 listed and 5 nonlisted). Likewise, the 25 selections in Bank B should produce 15 working household numbers (10 listed and 5 nonlisted), the 25 selections in Bank C should produce 1 working household number (nonlisted), and the 25 selections in Bank D will produce no working household numbers. In total, we will get 36 working household numbers (36% of the 100 numbers drawn), with 25 listed (69.4% of the working household numbers) and 11 nonlisted (30.6% of the working household numbers). In other words, our sample will perfectly mirror the population (subject to sampling error in the exact composition of the sample).

- If we use simple list-assisted RDD, we will eliminate Bank C and Bank D because they have no listed household numbers, and our 100 selections will be made entirely in Bank A and Bank B. If we draw 100 random numbers in these

Exhibit 2.3 How List-Assisted RDD Works

		Bank A	*Bank B*	*Bank C*	*Bank D*	*Total*
Total population	Numbers in bank Working HH numbers Listed/nonlisted	100 80 60/20	100 60 40/20	100 4 0/4	100 0 0/0	400 144 100/44
Simple RDD sample	Numbers selected Working HH selected Listed/nonlisted	25 20 15/5	25 15 10/5	25 1 0/1	25 0 0/0	100 36 25/11
Simple list-assisted RDD	Numbers selected Working HH selected Listed/nonlisted	50 40 30/10	50 30 20/10	0	0	100 70 50/20
Proportional list-assisted RDD	Numbers selected Working HH selected Listed/nonlisted	60 48 36/12	40 24 16/8	0	0	100 72 52/20

two banks, we expect to draw 50 numbers per bank. The 50 numbers in Bank A should produce 40 working household numbers (30 listed and 10 nonlisted), and the 50 numbers in Bank B should produce 30 working household numbers (20 listed and 10 nonlisted). In total, we will get 70 working household numbers (70% of the 100 numbers drawn), with 50 listed (71.4% of the working household numbers) and 20 nonlisted (28.6% of the working household numbers). The sample is much more efficient than a simple RDD sample—it yields many more working household numbers—but nonlisted numbers are slightly underrepresented as a result of omitting working banks that don't have any listed numbers (Bank C in this example).

- If we use proportional list-assisted RDD, we will eliminate Bank C and Bank D because they have no listed household numbers, and our 100 selections will be made entirely in Bank A and Bank B. Since Bank A has 60 listed numbers and Bank B has 40, 60% of the sample will be allocated to Bank A (i.e., 60 random numbers will be chosen in this bank), and 40% of the sample will be allocated to Bank B (40 random numbers). The 60 selections in Bank A should produce 48 working household numbers (36 listed and 12 nonlisted), and the 40 selections in Bank B should produce 24 working household numbers (16 listed and 8 nonlisted). In total, we will get 72 working household numbers (72% of the 100 numbers drawn), with 52 listed (72.2% of the working household numbers) and 20 nonlisted (27.8% of the working household numbers). This sample is even more efficient than simple list-assisted sampling, as a result of shifting selections toward the more heavily populated bank, but it has another source of bias. To the extent that a bank has fewer listed household numbers because it has a higher rate of nonlisted households (rather than simply having fewer households), it will be unfairly penalized in the allocation scheme. In our example, Bank B should have received 3/4 as many selections as Bank A, because it has 3/4 as many working household numbers, but it only received 2/3 as many selections because it had a higher rate of nonlisted households. As with the exclusion of working banks that don't have listed numbers (such as Bank C), the effect is to slightly underrepresent nonlisted numbers.

Like the Mitofsky-Waksberg procedure, list-assisted RDD sampling gains efficiency by eliminating banks of nonworking numbers: Currently, in a national sample, simple list-assisted RDD yields about 55% working household numbers, and proportional list-assisted RDD yields about 65% households. Like Mitofsky-Waksberg, list-assisted sampling has the ability to reach nonlisted households, although it slightly underrepresents these households, with the bias being larger

for proportional list assistance. In preference to Mitofsky-Waksberg, list-assisted sampling yields a simple random sample and does not require you to dispose of one number before you can proceed to another.

Somewhat looser RDD methods combine the use of directories with random dialing. Two such methods are to (a) select a sample of numbers from a telephone directory and "add 1" to the last digit of each selected number, so that 555-1234 becomes 555-1235, and (b) select a sample of numbers from the directory and replace the last two digits of each selected number with a two-digit random number. These methods, like Mitofsky-Waksberg, yield about 50% working household numbers.

"Add 1" or "replace two" sampling may be useful for student projects done with local populations, because these methods are easy to implement and allow you to draw an RDD sample without having money to buy one (assuming you have access to a list of local telephone numbers). In general, though, we strongly recommend buying a list-assisted sample if resources are available. The quality of these samples is ensured, and cost efficiencies in doing the research will more than make up for the cost of the sample.

Incorporating Cellphones

Effective telephone sampling strategies must deal with the evolving population of households that are landline only, cell only, or both landline and cell. As of 2012, roughly 35% of American households were cell only, and another 15% received all or almost all of their calls on a wireless phone despite also having a landline (Blumberg & Luke, 2012). The use of "cell only" varies substantially across population subgroups, as shown in Exhibit 2.4. U.S. adults with only wireless service (no landline) are disproportionately younger, poorer, and renters. Clearly, failure to incorporate cellphones into telephone survey designs may lead to coverage bias.

This is not simply an American phenomenon. In explaining why Israeli pollsters badly miscalled the 2012 elections in that country, an editorial in the *Jewish Daily Forward* noted that "(Israeli) pollsters tend to only call landlines, effectively excluding huge swaths of the population … not coincidentally, mobile phone users are the type of younger voter more likely to support (the party that did better than expected)" (*Jewish Daily Forward,* February 8, 2013).

One conceptually straightforward approach to dealing with cellphones is to select and combine samples from both a landline frame and a cellphone frame. Brick et al. (2007) describe a study specifically designed to evaluate such dual-frame designs. Using samples drawn by Survey Sampling International, a

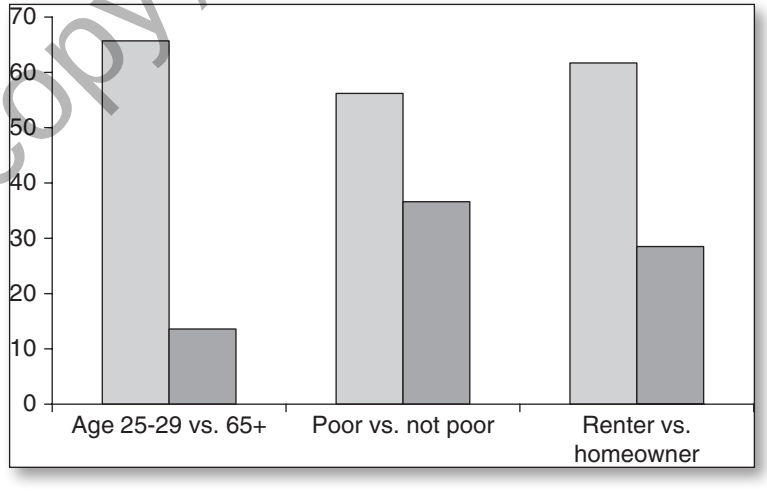
list-assisted RDD sample was combined with a cell sample from 1,000-number blocks identified as cellular in the commercial Telcordia database. In such designs, postsurvey weighting (as discussed in Chapter 7) is needed to adjust for overlap between the frames and any differences in selection probabilities.

Guterbock, Diop, Ellis, Le, and Holmes (2008) have proposed an alternative that dispenses with conventional RDD landline sampling. They argue that households with unlisted landlines and no cellphone are only a few percent of the total U.S. population and that researchers can obtain adequate population coverage by combining a sample of *listed* landline numbers with an RDD sample of cellphone numbers.

The most recent summary of cellphone survey methods is the report by the American Association for Public Opinion Research (AAPOR) Cell Phone Task Force (2010). That report concludes that "it remains premature to try to establish standards on the various issues as it is too soon in the history of surveying respondents in the U.S. reached via cell phone numbers to know with great confidence what should and should not be regarded as a best practice" (p. 16).

A sampling issue addressed in that report is whether to use (a) an overlapping dual-frame design in which respondents in the cellphone sample may also

Exhibit 2.4 Percentage of Adults With Cellphone Only (No Landline) Within Selected Population Groups



Source: Centers for Disease Control and Prevention. Retrieved from http://www.cdc.gov/nchs/nhis.htm

have landlines or (b) a dual-frame design with screening of the cellphone sample for cell-only status (and possibly for cell-mostly status). The two types of designs have different implications for how data from the landline and cell samples are combined into overall population estimates. The report does not recommend one over the other and concludes that either design might be the better choice based on the particulars of a given survey.

The AAPOR report notes that nonresponse in RDD cellphone surveys is somewhat greater than in comparable RDD landline surveys in the United States (although the difference has been shrinking over time), which makes cellphone samples somewhat more vulnerable to nonresponse bias. The report discusses various operational issues that affect response rates, including the best times of day for calling (which differ from landline phones), the maximum number and frequency of callbacks, transmitting accurate Caller ID information when dialing a cellphone, and keeping a Do Not Call list for cellphone owners who request that they not be called back.

Finally, the AAPOR report recommends that (1) researchers should explain the method by which the cellphone numbers used in a survey were selected; (2) if RDD telephone surveys do *not* sample cellphone numbers, then researchers should provide an explanation of how excluding cellphone owners might or might not affect the survey's results; and (3) researchers should explain any weighting of cellphone samples, including why the sample was not weighted if that is the case.

Address-Based Sampling

The main alternative to RDD designs for maximizing household coverage in surveys of the general population is address-based sampling (ABS), which uses a frame derived from the U.S. Postal Service (USPS) Computerized Delivery Sequence file (see Roth, Han, & Montaquila, 2013, for an examination of frame quality). It is possible to extend the coverage of this frame with a supplemental file available from the USPS, the No-Stat file, which includes additional rural addresses that do not get direct delivery of mail (Shook-Sa, 2013).

Address-based sampling is a natural fit for mail or in-home surveys. Also, many addresses in the Delivery Sequence file can be linked to phone numbers, which creates the possibility of mixed-mode designs (Iannacchione, 2011). Boyle, Fleeman, Kennedy, Lewis, and Weiss (2012) used such a design in which the ABS addresses were matched to listed telephone numbers. Approximately half of a sample of 10,000 addresses was matched. The remaining addresses (assumed to include cell-only and landline-plus-cell households) were sent a

mail survey that included a cash incentive to provide a phone number by return mail. The returned questionnaires included cell-only and dual-use households in about equal numbers. This approach permitted weighted estimates for all population groups. Brick et al. (2012) used a similar approach in a two-state survey.

ABS survey methods are relatively new and continue to be refined. In an article summarizing current ABS methodologies, Iannacchione (2011) suggests that one "read [the] article quickly. By the time you're finished, parts of it will likely be out of date." However, ABS surveys are rapidly becoming more prevalent at research organizations with sufficient resources and expertise.

More generally, sample design for telephone or mixed-mode surveys is currently in flux, and the rapidly changing technical and behavioral context of modern surveys suggests that best practices are likely to remain a moving target.

Registration-Based Sampling

Registration-based sampling (RBS) is a variation on address-based sampling that is used in political surveys. In RBS, lists of registered voters are used as the sampling frame. A typical registration record contains the name and mailing address of each voter along with date of birth and date of registration (Green & Gerber, 2006). This information can be augmented with records of turnout in past elections, including turnout in party primaries.

Registration-based sampling can be implemented in two broad ways. First, the lists may be used "as is" to draw a random sample of registered voters. Second, if the research is being conducted to measure voter preferences in an upcoming election, the lists may be modified to reflect likely turnout in the election. For example, based on turnout records from previous elections, the listed elements may be divided into groups with different probabilities of voting in the upcoming election, and these groups may be sampled in proportion to their probability of voting. This will produce survey results that are implicitly weighted for likely turnout.

As with address-based sampling, registration-based sampling is a natural fit for mail or in-home surveys. However, political pollsters normally prefer data collection modes such as telephone or web that allow more timely results (including overnight polls to gauge public reaction to advertisements, debates, or news events). Using RBS for telephone or web surveys requires a transition from mailing addresses, which are given in voter registration records, to telephone numbers or e-mail addresses, which are not. For example, Green and

Gerber (2006) attempted to link telephone numbers to voter registration records and were able to do so for 65% to 70% of selected voters in Maryland, New York, Pennsylvania, and South Dakota (although some of the phone numbers were outdated). Barber, Mann, Monson, and Patterson (2014) sent a letter to selected voters in Colorado, Florida, and Utah asking them to access a survey website and got response rates of 5% to 7%.

An obvious problem in using RBS for telephone or web surveys is the number of potential respondents lost in translation from addresses to phone numbers or web: more than 30% for Green and Gerber (2006) and more than 90% for Barber et al. (2014). Green and Gerber also report evidence that the loss varied across population subgroups; for example, the Maryland sample underrepresented African Americans, and the New York and Pennsylvania samples underrepresented urban dwellers.

Another problem with RBS is that it may fail to capture newly registered voters. The number of voters missed for this reason will vary from place to place, depending on population characteristics such as age and mobility, how quickly local voting authorities update their lists, and whether the place allows same-day registration. Also, if turnout in prior elections is used to modify probabilities of selection, the result may improperly represent irregular voters who are energized by a particular candidate or issue.

The coverage problems that result from inability to match telephone numbers to registration records, or failure to capture new registrants, can be addressed by supplementing a registration-based sample with a random-digit dialing sample (cf. Mitofsky, Bloom, Lenski, Dingman, & Agiesta, 2005). This is an example of dual-frame sampling, which is discussed later in this chapter.

Half-Open Intervals

Next on our list of possible ways to compensate for omission is the use of *half-open intervals*. In this method, the set of potential population elements is divided into subgroups, and one element in each subgroup is designated as "open." If the open element is drawn for the sample, then the full subgroup is searched for unlisted population members, and any unlisted elements discovered in this process are included in the sample.

For example, in household surveys with personal interviews, you might divide the population into street blocks, and you might designate street numbers ending in "01" as the open elements on each block. So, if you draw the address 3701 Southmore Avenue, then you would search the 3700 block for unlisted housing units and add them to the sample.

The half-open interval gives unlisted elements the same chance of being drawn in the sample as the open element to which they are linked. Of course, if an open element is itself unlisted, then there is no chance of selecting it or other unlisted elements in its subgroup, which may cause some bias in the sample. This potential bias is not a practical problem if the list coverage rate is high but becomes more of a concern as coverage drops.

When using this method, keep the subgroups small, to limit the possibility of swamping the sample with a large number of unlisted elements from a single subgroup. In a household survey, we once had a situation in which an open element led to a new, unlisted high-rise building with hundreds of residents, in a study that was designed to study only 200 households in total. In such situations, the usual response is to take some but not all of the unlisted elements—for example, to take a maximum of three unlisted elements from any open unit and weight those elements as needed to adjust for the full number of unlisted elements in the unit.

While the half-open interval method works well in theory, it may not work so well in practice. Eckman and O'Muircheartaigh (2011) conducted an assessment of the procedure in a sample of housing units in neighborhoods of Seattle, Washington; Providence, Rhode Island; and Oakland, California. They found that interviewers had a low rate of success in reporting housing units that had been deliberately dropped from the area listing given to them. Their overall conclusion is that the half-open interval method does not produce satisfactory results and should not be used unless working from a frame with a "good deal of undercoverage which is believed to lead to undercoverage bias" (p. 130).

Dual-Frame Designs

We have mentioned dual-frame designs in which telephone numbers are drawn from two separate lists, one for landlines and one for cellphones. We also mentioned dual-frame designs in which prospective voters are drawn through a combination of registration-based sampling and random-digit dialing. More broadly, one might compensate for omissions by separating the population elements into two groups, listed and unlisted. Listed elements are drawn from the list, and unlisted elements are pursued by other means.

For example, if you are doing a study on politically active people in a community, and you have a list of members from a local political organization, then members of that organization might be sampled from the list, and politically active people who are not members of the organization might be sampled by screening

the general population.[6] Using the list allows you to sample the listed part of the population with high efficiency (hence lower costs), while screening the general population for unlisted members gets you beyond the potential coverage bias of the list. Since the unlisted elements will be more expensive to locate, they probably will be sampled at a lower rate, and overall estimates for the community will be calculated through a weighted combination of the two groups. This is an application of stratified sampling, which is discussed in detail in Chapter 5.

Here is another example of dual-frame sampling:

CASE STUDY 2.4

When Pope Benedict XVI resigned in early 2013—the first time in almost 600 years a Roman Catholic pope had resigned—the *New York Times* and CBS News conducted a survey to learn how U.S. Catholics felt about Benedict's papacy, the extent to which the Church was in touch with their needs, and what the next pope should be like. The survey was conducted with 1,585 adults throughout the United States.

The sample had three broad elements. First, a list-assisted RDD sample of landline telephone exchanges was drawn. Second, this landline sample was supplemented through random dialing of cellphone numbers. Third, self-identified Catholics from previous national polls by the *New York Times* and CBS News were called again for the new poll, and their responses were added to those of Catholics in the new sample. The three samples were combined and weighted as needed to make estimates.

This led to a front-page story, "U.S. Catholics in Poll See a Church Out of Touch." Among the findings, "three-fourths of those polled said they thought it was a good idea for Benedict to resign. Most wanted the next pope to be someone younger, with new ideas" (*New York Times,* March 5, 2013, p. 1).

This survey illustrates two uses of dual-frame sampling (as well as list-assisted random-digit dialing). First, a cellphone frame was used to address omissions in the landline frame. Second, a frame of known Catholics—identified through previous random samples of the general population, and hence presumably a random sample of U.S. Catholics—was used to obtain additional Catholic respondents without the cost of screening the general population.

6. In doing so, it will be important to identify whether respondents selected from the general population frame are members of the organization, for two reasons. First, these respondents will be counted with the organization sample. Second, since we probably don't have a good measure of how many people in the community are politically active, the rate of occurrence for organization members in the general population may be used to estimate the size of the nonlisted population. This will tell us how to weight the two groups in generating overall estimates. Similarly, in a dual-frame telephone survey, we will ask members of the landline sample if they have cellphones, and members of the cellphone sample if they have landlines, so we can assign respondents to landline-only, cell-only, and landline-plus-cell groups and weight each group appropriately.

Dual-frame sampling also may be used to address the fact that there is no general frame of online users. For example, Blair and Blair (2006) discuss the possibility of using an online panel to efficiently locate members of a rare population, in conjunction with RDD telephone screening to obtain broader population coverage.

General Comments on Coping With Omissions

All of the procedures that we have described for coping with omissions fall into two broad categories. In some cases, we augment the list (i.e., we use the list and supplement it in some way). Half-open intervals fall into this category, as do dual-frame designs. In other cases, we create or find a better list. Pure random-digit dialing and address-based sampling for telephone surveys fall into this category; we abandon listed telephone numbers for purely random numbers or mailing addresses. The extent to which we augment or replace the list—or simply ignore the problem—depends on practical judgments regarding the extent to which omissions might cause coverage bias, the extent to which the proposed method will mitigate those problems, and the cost-effectiveness of the method.

A review of these procedures also shows the special challenge that omissions pose for online surveys of the general population. In planning an in-home or mail survey of the general population, one can obtain a relatively complete list of addresses or, worst case, send people to list the housing units or establishments in some area. In planning a telephone survey, one can use random dialing to cover the population. In surveys of visitors to a place, one can use a counting frame and count people as they pass a sampling location. But in online surveys, there is no way to cover the general population completely. If the population of interest is a group for which a list of e-mail addresses is available, this is not an issue, but for general populations, the inability to assemble a complete frame means that online surveys ultimately must rely on some form of model-based estimates as opposed to probability-based estimates. We discuss model-based estimates further in Chapter 7.

2.2.4 Coping With Ineligibles

The next framing problem is ineligibles. These are elements that are contained in the frame that do not represent eligible population members. For example, if you are doing a study of political preferences among people who are likely to vote in an upcoming election, and you are using RDD telephone numbers

as a sampling frame, those numbers will include many people who are not likely to vote.

Coping with ineligibles is straightforward—don't select them. Since they aren't in the population, they shouldn't be in the sample.

There are two ways to keep ineligibles out of a sample. First, the entire list can be screened for eligibility before sampling, with all ineligible listings being deleted. This often is not practical because the eligibility factor is not visible in the list; for example, a list of RDD numbers will not show whether people intend to vote in an upcoming election. The other approach is to screen selected elements for eligibility *after* sampling, with ineligibles dropped at this time. This is the usual method employed. In some cases, both methods may be used if the population is defined on multiple criteria, and some are visible and some are not. For example, if a population is defined as undergraduate students at a particular university who exercise at least three times per week, it may be possible to drop graduate students when drawing names from the campus directory, then screen selected undergraduates for exercise frequency.

Dropping ineligibles means just that—dropping them, not replacing them. Inexperienced researchers sometimes think that ineligibles should be replaced by taking the next name on the list, but this procedure gives extra chances of selection to the population members who are listed after ineligibles and may cause sample bias. The proper method is simply to adjust the sample size for shrinkage due to ineligibility and drop ineligible listings when encountered.

Adjusting the sample size is done as follows: If e is the proportion of the list that is eligible, the adjusted sample size is n/e, where n is the desired sample size. For example, if you desire a sample of 300 freshmen at a given university, and only 20% of the names in the college directory are freshmen, the adjusted sample size to be chosen from the directory is $300 \div [.20] = 1{,}500$. A sample of 1,500 names from this directory should yield 300 freshmen.

Estimates of eligibility rates may be obtained either from prior experience or by studying a small pilot sample. Since these estimates may not be exactly accurate, it is a good idea to estimate eligibility on the low side so you are sure to get a large enough sample. Say, for example, that you think that freshmen will account for 20% of the listings drawn from a college directory, but you think the eligibility rate for any given sample could be anywhere from 15% to 25%. Use the 15% figure; in other words, if you want 300 eligibles, draw an initial sample of $300 \div [.15] = 2{,}000$. Then, if this sample yields 400 eligibles (an eligibility rate of 20%), you can randomly choose 300 for retention or 100 for deletion. The result will still be a random sample: A random sample of a random sample

is a random sample, and a random sample minus a random sample is a random sample.

When eligibility will be determined through screening interviews, you should use both the low and high eligibility estimates in planning sample size. In our example, the lowest eligibility estimate, 15%, means that a sample size as large as 2,000 may be needed to produce 300 eligibles. The highest estimate, 25%, means that a sample size as small as 1,200 may be needed (300 ÷ [.25] = 1,200). To protect yourself in this situation, draw a sample of 2,000 but do not release the entire sample for screening. Instead, draw a random subsample of 1,200 from the 2,000; release this subsample and hold the other 800 selections for use if all or part of it is needed. If the first 1,200 selections yield 240 interviews (an eligibility rate of 20%), then draw 300 of the holdout selections to get the remaining 60 interviews (60 ÷ [.20] = 300). This procedure gives you as many selections as you need without producing expensive and unnecessary data.

In studies of rare populations, the potential difference in sample size (depending on variations in the eligibility rate) may be so large that you prefer to draw the smaller sample, work it, and then use the resulting information about eligibility rates to resample as needed. For example, an eligibility rate of 1% would require 100,000 selections to provide 1,000 observations, but an eligibility rate of 2% would require only 50,000 selections. In a recent study of gay urban males that one of the authors worked on, initial samples in various geographic areas were drawn with an assumption that self-identified gay males would be no more than 4% of the population; in general, the correct rate was found to be approximately 2%, so many additional observations were needed, but the first-stage results were very useful in designing the second-stage sample.

A common error is to draw the larger sample (2,000 in our example), release the entire sample for data collection, and simply stop the research when the desired number of eligibles is obtained. This procedure is incorrect, because the first observations obtained will be the observations that are easiest to get; for example, in a survey of the general public, these observations will be homemakers and retired people. Any sample released for data collection should be worked completely to avoid bias in favor of the easy observations.

When adjusting the sample size to allow for ineligibles, it is also appropriate to adjust for the expected response rate. The adjusted sample size is $n/(e^*r)$, where n is the desired sample size, e is the eligibility rate, and r is the expected response rate. For example, if you want 300 usable observations, and you expect 20% of the selections to be eligible and 20% to cooperate, then you need an initial sample size of 300 ÷ [.20 × .20] = 7,500.

2.2.5 Coping With Duplications

If a sample is selected from a list that contains duplications of some elements, but none of the duplicated elements are selected more than once, has the duplication caused any problems? The answer is yes.

The basic problem with duplication is that it gives *groups* of population members disproportionate chances for selection into the sample. Even if individual duplicated elements aren't chosen more than once, their *group* is chosen at a higher rate. The group of population members who appear twice on the list is overrepresented by a factor of 2 compared with the group of members who appear once, the group of population members who appear three times on the list is overrepresented by a factor of 3, and so on. This overrepresentation will cause sample bias if the duplicated elements are different from the nonduplicated elements on some variable of interest in the research.

For example, imagine that you are drawing a sample of students at a college, and the procedure is to randomly select classes and then students within classes. Students who are in four classes have four chances that their classes will be drawn, while students who are in two classes have only two chances. The result will be to overrepresent full-time versus part-time students. (Notice that you would not encounter this problem if you sampled students directly from the student directory. The directory might have other problems such as omissions or ineligibles, but each student should be listed only once. This illustrates the fact that different frames for the same population may present different problems.)

There are three ways to correct for duplicate listings. The brute-strength method is to cross-check the list, identify duplicates, and remove them. This will be done if the list is computerized.

A second method is to draw the sample and check only the selected elements to determine how many times they are duplicated in the total population list. Then, to restore equal probabilities of selection, sample members that are listed k times would be retained at the rate of $1/k$ (i.e., members that appear twice in the list would be retained at the rate of 1/2, members that appear three times at the rate of 1/3, etc.). This method is appropriate for noncomputerized lists. It will cause some shrinkage in the sample size, which can be handled in the same way as shrinkage for eligibility.[7]

7. This method also provides a way to assess the seriousness of duplication in a list. If the problem is minor, you might ignore it.

The third method, which is usable in surveys, is to ask selected population members how many times they appear in the list. Under this method, all of the data gathered are retained, because it would be wasteful to discard completed interviews, but observations are weighted by the inverse of their number of times in the list. That is, sample members who say they are listed k times are weighted by $1/k$.

This last method obviously requires that sample members know how many times they are listed (e.g., how many classes they are taking). This method should be used only when there is good reason to believe that this assumption is true and when checking the list is difficult or impossible.

Similarly, cross-checking a list works best if the population members have unique identifiers, such as a telephone number. Otherwise, the same household might be represented by people with different first names, and the same person might be represented with variations in the name or address. If you look at the junk mail that comes to your home, you'll probably see these types of variations. Because of these variations, computerized cross-checking usually doesn't remove all of the duplications in a list, but it should reduce them to a level where they cause negligible sample bias.

2.2.6 Coping With Clustering

Our final list problem is clustering. Clustering is similar to duplication in that it involves unfair representation of some group of population members, rather than complete exclusion or inclusion. The difference is that clustered elements are underrepresented in the sample while duplicated elements are overrepresented.

Here is an example of clustering. A town has 100,000 adults: 50,000 married and 50,000 single. The 50,000 married adults form 25,000 households, each with one landline telephone number. The 50,000 single adults form 50,000 households, each with one landline number. A sample of 300 of these numbers for a telephone survey will produce 100 "married" numbers and 200 "single" numbers, because the single people account for 2/3 of the telephone numbers. If one adult is interviewed at each number, the sample will contain 100 married adults and 200 single adults. This is a fair sample of households but not of individuals. Married people account for 1/2 of the adults in town but, because they are clustered on the list, they account for only 1/3 of the adults in the sample.

Clustering arises in mail, telephone, or in-home surveys whenever the desired population unit is individuals (e.g., potential voters) but the sampling frame is a list of households (e.g., a list of residential addresses or landline

telephone numbers). Clustering typically does not arise in lists of cellphone numbers or, for the most part, lists of e-mail addresses in online surveys because these lists correspond to individuals. On the other hand, if you have a list of individuals but the desired population unit is households, then individuals from the same household will represent duplicates in the list.

Clustering also arises in surveys of organizations whenever the desired population unit is nested within organizations. An example is Case Study 2.2 in which individual businesspeople gave answers to a survey but the desired population was sales dollars.

There are three basic ways to cope with clustering:

- First, you can gather data from all population elements in the selected clusters. For example, in our married/single example, you would gather data from both adults at each "married" number, which would give you data from 200 married people and 200 single people. Some adjustment to the initial sample size would be needed if the data collection budget allows only 300 observations.

This method provides a fair chance of selection for every member of the population. Unfortunately, as the married/single example shows, it also produces a sample that contains related cluster members. Of 400 population members in that sample, 200 are husbands and wives. This can cause problems of contamination from respondents talking to each other. It also means that the sample is less heterogeneous and hence contains less information than an equal-sized sample of unrelated observations (see the discussion of cluster sampling in Chapter 6). Because of these issues, taking entire clusters is only a good idea when (a) clusters are relatively small and relatively few in number, or (b) it is simply impossible to separate the elements, as with sales dollars within companies.

- Second, you can sample population members within clusters at some fixed rate. The usual rate for sampling individuals within households is 1/2. In the married/single example, you would retain only half of the "single" households (and drop the others), and you would gather data from one randomly chosen person in each of the "married" households. This would produce a sample size of 100 people from the 100 "married" households and 100 people from the 200 "single" households. A sample size adjustment would be needed to get 300 interviews.

This method provides a fair chance of selection for every member of the population. However, it is a relatively expensive procedure, because you pay to locate some number of "single" households, and then you discard half of them.

• Third, you can compensate for clustering by randomly selecting one population member from each selected cluster and weighting that observation by the size of the cluster. In the married/single example, one observation would be obtained from each of the 100 "married" and 200 "single" households in the original sample, but the married observations would be weighted by 2, giving them the weight of 200 people in total.

Sampling Within Households

The original method of randomly selecting individuals within households was developed by Kish (1949), who developed a procedure that requires listing all household members and then selecting one individual based on a preset rule called a Kish Table. This procedure became the standard method for sampling within households for door-to-door surveys.

In telephone surveys, the Kish procedure was found to be less effective, because respondents could not see the interviewer and were less willing to provide a listing of all household members. One alternative was to use a method proposed by Troldahl and Carter (1964). This method required only two pieces of information: the number of adults in the household and the number of males (or females). An individual was then selected through a rotating series of four selection tables. The Troldahl-Carter method eliminated the need to list the entire household, but some bias remained because not all household members had equal probabilities of selection within the four tables and also because of disparities in the gender composition of single-adult households. There tend to be more single-female households, due to widows and single mothers who don't take a roommate or a live-in partner (on the other hand, single-female households tend to be undermeasured, because females who live alone may not tell a caller that there is no male in the household).

A subsequent method has been to select the household member who most recently had a birthday (O'Rourke & Blair, 1983) or the person who will have the next birthday (Lind, Link, & Oldendick, 2000). While unbiased in theory, the last (or next) birthday method encounters some bias in practice because initial respondents disproportionately claim themselves as the proper respondent. Studies have found that the proper respondent is identified 75% to 90% of the time (Lavrakas, Bauman, & Merkle, 1993; Lavrakas, Stasny, & Harpruder, 2000; Lind et al., 2000; O'Rourke & Blair, 1983). While not perfect, this method is now the most widely used procedure for sampling within households in telephone surveys because it is easy to administer.

Most recently, Rizzo, Brick, and Park (2004) have proposed a method of within-household sampling that is minimally intrusive, reducing the effect of the selection process on household response. Their approach takes advantage of the fact that single and two-adult households account for about 85% of all U.S. households. Selection is not at issue in the single-adult households. In the two-adult households, half the time, the adult answering the phone is selected, and half the time the other adult is chosen. More complicated or intrusive methods then have to be used only in the 15% of remaining households.

Weighting Data to the Proper Population Unit

Earlier in this chapter, we gave an example in which individual businesspeople gave answers to a survey, but the desired population unit was sales dollars: see Case Study 2.2. This is a common problem in market research: The unit of interest is expenditures, but the reporting units are people or companies.

From a framing point of view, the problem can be viewed as a form of clustering. Just as there are two adults in each "married" household in our married/single example, there are $X of potential sales at each company in Case Study 2.2. In the married/single example, we can randomly select adults within households if our desired population unit is individuals rather than households, but we cannot sample dollars within companies in Case Study 2.2 because the dollars cannot speak. Instead, we let one respondent speak for all of the dollars in a company, and we weight those answers by the cluster size (i.e., the number of dollars).

2.2.7 Framing Populations Without a List

In some research projects, sampling must be done without a list of the population. For example, if you want to survey visitors to a particular website or shopping mall, you won't find a list of visitors. Sampling without lists is done from "counting frames" as follows:

- Estimate the size of the population.
- Select a sample of numbers between 1 and *N,* where *N* is the population size.
- Count the population and gather data from the appropriately numbered members.

For example, in a sample of visitors to a shopping mall, if you expect 10,000 shoppers to enter the mall during the interviewing period, and you wish to select 500 of them, you can randomly select 500 numbers between 1 and 10,000. Alternately, you can take every 20th number after some random start (10,000 ÷ 500 = 20). You then will count shoppers as they enter the center and approach the shoppers with appropriate numbers.

Counting frames are subject to the same problems as lists: omission, ineligibility, duplication, and clustering. Omission results from underestimating the population size (or the sizes of population subgroups). For example, if you estimate the number of visitors to a shopping mall at 10,000, and this estimate is too small, then all shoppers beyond the 10,000th have no chance of selection because they don't appear in the sampling frame.

Ineligibility results from some of the counted elements not meeting population criteria. For example, counting the visitors to a mall is easiest if you count every visitor, but some of the visitors might not meet requirements that have been set for the population (regarding age, gender, product usage, or whatever). Ineligibility also results from overestimating the size of the population (or of population subgroups). Your sample may be smaller than expected because only 9,412 shoppers visited the mall during the interviewing period, and you expected 10,000. In effect, the numbers 9,413 through 10,000 were ineligible elements in your sampling frame.

Duplication and clustering usually result from a mismatch between the counting units and population units. For example, you might want a sample of *people* who shop at some mall, but the implicit counting unit is visits to the mall. Some people visit the mall more than others, and the extra visits constitute duplications in the counting frame if the desired unit is people (Blair, 1983).

In general, the available solutions for problems in counting frames are more limited than the solutions for problems in lists. Omission is solved by simply estimating population size on the high side. Ineligibility is solved by screening for ineligibles. Duplication and clustering are usually solved by weighting data after they are gathered, because the absence of list documentation makes it impossible to clean or check the sampling frame prior to data collection.

2.3 CHAPTER SUMMARY

This chapter discussed issues in defining and framing populations. Regarding population definition, we noted that a sampling population is the set of elements about which you would like to draw conclusions. To define a population,

you have to specify (a) the population units and (b) the population boundaries. The boundaries must be stated in specific operational terms.

Regarding sample frames, we noted that a frame is a list or system that identifies every member of the population symbolically. Ideally, the frame should have a one-to-one correspondence with the members of the population. This may not occur because of omissions, ineligibles, duplications, or clustering.

The response to omissions is to ignore them if the problem is small and address them if the problem is likely to cause bias in the sample. Random-digit dialing, dual-frame designs to incorporate cellphones, and possibly address-based sampling or registration-based sampling are used to handle unlisted numbers in telephone surveys. Other methods for dealing with omissions are the use of half-open intervals and more general dual-frame designs.

The response to ineligibles is to drop them when encountered. If ineligible elements cannot be recognized in the frame, it may be necessary to screen for them in the field. Either way, the initial sample size should allow for losses due to ineligibility.

Possible responses to duplicate listings are (1) to delete them from the frame prior to sampling, (2) to retain selected elements at a rate that is inverse to the number of times each selected unit is listed, or (3) to ask each selected participant how many times he or she is listed and weight the observations by the inverse of the number of listings.

Possible responses to clustering are (1) including every member of selected clusters, (2) sampling within clusters, or (3) randomly selecting one element in each cluster and weighting for cluster size.

EXERCISES AND DISCUSSION QUESTIONS

Exercise 2.1

A researcher wishes to study key business leaders to learn their opinions about issues facing a metropolitan area. Define this population in specific operational terms.

Exercise 2.2

The municipal government of a "college town" wishes to survey area residents regarding their park and recreation needs. Define this population in

specific operational terms. Should children be eligible to respond? People who live outside the city boundaries? Students at the local university who have access to university facilities? Students who live in dormitories? Students in fraternities or sororities? People in the local jail? People in a homeless shelter? People in nursing homes?

Exercise 2.3

A friend of yours is running for a place on the local school board, and you agree to help her by surveying local voters to learn which issues are most important and what they would like the school board to do. Can you get a list of registered voters who live in the school district? If so, does it contain mailing addresses? Telephone numbers? E-mail addresses? Apart from this, is there a directory of telephone numbers that might be usable for your purposes? A directory of mailing addresses? A directory of e-mail addresses?