5 Sampling

The Basics

IN THIS CHAPTER

WHAT ARE SAMPLES AND WHY DO WE NEED THEM?

Informant accuracy, data validity, and even ethical issues (like whether it's all right to deceive people in conducting experiments) are all measurement problems in research. The other big class of problems involves sampling: Given that your measurements are credible, how much of the world do they represent? How far can you generalize the results of your research?

The answer depends, first of all, on the kind of data in which you're interested. There are two kinds of data of interest to social scientists: individual data and cultural data. These two kinds require different approaches to sampling.

Individual data are about *attributes of individuals in a population*. Each person has an age, for example; each person has an income; and each person has preferences for things like products, political positions, and characteristics for a mate. If the idea in collecting data is to estimate the average age, or income, or preference in a larger population—that is, to estimate some population parameters—then a scientifically drawn, unbiased sample is a must. By scientifically drawn, I mean random selection of cases so that every unit of analysis in your study has an equal chance of being chosen for study.

Cultural data are different. Cultural data require experts. If you want to understand a process—like how people in a factory work group decide on whether to lodge a complaint to management, or how the police in a squad car determine whether to stop someone on the street—then you want people who can offer expert explanations about the cultural norm and about variations on that norm (Handwerker et al. 1997). It's one thing to ask: "How many people did you stop on the street for questioning last week?" This requires an answer about individual behavior. It's another thing to ask: "How do people in your squad decide whether to stop someone for questioning on a street patrol?" This requires cultural experts.

Individual attribute data requires **probability sampling**; cultural data require **nonprobability sampling**. This chapter is about the basics of probability sampling, which will take us into a discussion of probability theory, variance, and distributions in Chapter 6. We'll get to nonprobability sampling in Chapter 7 (**Further Reading:** general sampling).

WHY THE UNITED STATES STILL HAS A CENSUS

If samples were just easier and cheaper to study but failed to produce useful data, there wouldn't be much to say for them. A study based on a random sample, however, is often *better* than one based on the whole population.

Since 1790, the United States has conducted a census once every 10 years in which every person in the country is supposed to be counted in order to apportion seats in the House of Representatives to the states. Lots of things can go wrong with counting. Heads of households are responsible for filling out and returning the census forms, but in 1990, only 63% of the mailed forms were returned, and that was down from 78% in 1970. The Bureau of the Census had to hire and train half a million people to track down all the people who had not been enumerated in the mailed-back forms.

Even then, there were problems with the final numbers. Some college students were counted twice: Their parents had counted them on the mailed-back census form and then, on census day, some of those same students were tracked down again by enumerators who canvassed the dorms. Meanwhile, lots of other people (like illegal immigrants and people living in places to which the census takers would rather not go) were not being counted at all.

In 1997, the Bureau of the Census asked the U.S. Congress to allow sampling instead of counting for at least some parts of the 2000 Census. This caused a serious political problem: If sampling produced more accurate (and, presumably, higher) estimates of the number of citizens who are, say, homeless or who are migrant farm workers, this would benefit only certain states and might benefit the Democratic Party over the Republican Party. So, Congress rejected the proposal, citing Article 1, Section 2 of the Constitution, which refers to the Census as an "actual Enumeration" (with a capital E, no less).

No getting around it: Actually enumerating means counting, not estimating, and the U.S. Supreme Court agreed, in 1999. To deal with the inaccuracies of a head count, the Bureau of the Census publishes adjustment tables, based on samples. In 2000, for example, the bureau determined that it had undercounted American Indians who live off reservations by about 53,000 (see U.S. Bureau of the Census n.d.). In May 2009, at his confirmation hearing to head the bureau, Robert Groves assured Republican senators that he would not use sampling methods in the then upcoming 2010 Census (Herszenhorn 2009).

IT PAYS TO TAKE SAMPLES AND STICK WITH THEM

If you are doing all the work yourself, it's next to impossible to interview more than a few hundred people. Even a small county school system might have 500 employees, including teachers, administrators, and staff. You'd need several interviewers to reach all those people within a reasonable amount of time. Interviewers may not use the same wording of questions; they may not probe equally well on subjects that require sensitive interviewing; they may not be equally careful in recording data on field instruments and in coding data for analysis. And, as you'll see in the section on telephone interviewing, in Chapter 9, some interviewers actually falsify data.

Most important, you have no idea how much error is introduced by these problems. A well-chosen sample, interviewed by people who have similarly high skills in getting data, has a known chance of being incorrect on any variable. Furthermore, studying an entire population may pose a history threat to the internal validity of your data. If you *don't* add interviewers it may take you so long to complete your research that events intervene that make it impossible to interpret your data.

For example, suppose you're interested in how the nursing staff at a midsize, private hospital feels about a reorganization plan. You decide to survey *all* 210 nurses on the staff, using a structured, 10-minute personal interview. You know that it's tough to track some nurses down—they are very busy and sometimes don't have even 10 minutes to stop and chat; they change shifts, forcing you to find them at 4 a.m.—but you have three months for the research and you figure you can do the survey a little at a time.

Two months into your work, you've gotten 160 interviews on the topic—only 50 to go. Just about that time, the hospital announces that it has been bought out by a big health maintenance corporation—one that's traded on the New York Stock Exchange. All of a sudden the picture changes. Your "sample" of 160 is biased toward those people whom it was easy to find, and you have no idea what *that* means. And even if you could now get those remaining 50 respondents, their opinions may have been radically changed by the new circumstances. The opinions of the 160 respondents who already talked to you may have also changed.

Now you're really stuck. You can't simply throw together the 50 and the 160 interviews; you have no idea what *that* will do to your results. Nor can you compare the 160 and the 50 as representing the nursing staff's attitudes before and after the buy-out. Neither sample is unbiased with regard to what you are studying.

If you had taken a **random sample** of 60 people in a single week early in your project, you'd now be in much better shape because you'd know the potential sampling error in your study. If historical circumstances (the surprise buy-out, for example) require it, you could interview the same sample of 60 again (in what is known as a **panel study**) or take another representative sample of the same size and see what differences there are before and after the critical event. In either case, *you are better off with the sample than with the whole population.* There is no guarantee that a week is quick enough to avoid the problem described here. It's just less likely to be a problem. Less likely is better than more likely (Box 5.1).

Box 5.1 Probability samples are probably representative

Probability samples are based on taking a given number of units of analysis from a list, called a **sampling frame**, which represents some **population** under study. In a probability, or **unbiased sample**, each individual has exactly the same chance as every other individual of being selected. When this principle is violated, samples become biased.

A famous case of sampling bias occurred in 1970 while the United States was engaged in a very unpopular war in Vietnam. Men were selected to serve in the military by a supposedly random draw. Three hundred and sixty-six capsules (one for each day of the year, including leap year) were put in a drum and the drum was turned to mix the capsules. Then dates were pulled from the drum, one at a time. All the men whose birthdays fell on the days that were selected were drafted.

When enough men had been selected to fill the year's quota, the lottery stopped. Men whose birthdays hadn't been pulled were safe until the following year when the lottery would be run again. It turned out that men whose birthdays were in the later months had a better chance of being drafted than men whose birthdays were earlier in the year. This happened because the drum wasn't rotated enough to thoroughly mix the capsules (Williams 1978). Not a good sampling technique.

Sampling Frames

If you can get it, the first thing you need for a good sample is a good sampling frame. (I say "if you can get it" because a lot of social research is done on populations for which no sampling frame exists. More on this at the end of this chapter.) A sampling frame is a list of units of analysis, *from which* you take a sample and *to which* you generalize.

A sampling frame may be the tax rolls of a community or a geographic information system (GIS) map of housing units. Lists of local addresses may be available at libraries or at municipal buildings or online. (Careful: Lists of addresses can be out of date even before they are made public.) Telephone directories were used as sampling frames in the past, but are increasingly obsolete as sampling frames as more and more people across the world switch to cell-phone-only service. Professional survey researchers in the industrialized nations of the world often purchase samples from firms that keep up-to-date databases just for this purpose. For many projects, though, you just have to make your own census of the population you are studying. A census of a factory or a hospital or a small town gives you the opportunity to walk around a community and to talk with most of its members at least once. It lets you be seen by others and it gives you an opportunity to answer questions, as well as to ask them. It allows you to get information that official censuses don't retrieve. A list of the employees at a plant, for example, probably won't have information on all the variables that you need for your research.

A census of a community of actors gives you a sampling frame from which to take many samples during a research project. It also gives you a basis for comparison if you go back to the same population later.

Simple Random Sampling

To get a simple random sample of 200 out of 640 professors in a university, you number each individual from 1 to 640 and then take a random grab of 200 out of the numbers from 1 to 640. Most packages for statistical analysis have built in random-number generators, and you can create random samples by using one of the random-number generators on the Internet—like the one at http://www.randomizer.org/form.htm.

When you have your list of random numbers, then whoever goes with each one is in the sample. Period. If there are 1,230 people in the population, and your list of random numbers says that you have to interview person #212, then do it. No fair leaving out some people because they are members of the elite and probably wouldn't want to give you the time of day. No fair leaving out people you don't like or don't want to work with. None of that.

A common form of meddling with samples is when door-to-door interviewers find a sample selectee not at home and go to the nearest house for a replacement. This can have dramatically bad results (Box 5.2).

Box 5.2 The social research industry

In the real world of research, random samples are tampered with all the time. No snickering here about the "real world" of research. Social research is a major, worldwide industry. The American Community Survey of the U.S. Census surveys, by mail, around 250,000 people each month—that's around 3,000,000 interviews a year—plus telephone follow-ups with 85,000 people and face-to-face follow-ups with about 40,000 people each month (U.S. Bureau of the Census 2009:Ch. 2, p. 4). Over 300 firms are members of the Council of American Survey Research Organizations (Wright and Marsden 2010:17–19), and about $2 billion was spent in 2009 on online research alone (Baker et al. 2010:715). That's real enough for most people.

Suppose you go out to interview between 10 a.m. and 4 p.m. People who are home during these hours tend to be old, or sick, or mothers with small children. Those same people are home in the evening, too, but now they're joined by all the single people home from work, so the average family size goes down. As Tom Smith, director at the National Opinion Research Center, says, going to the nearest at-home household for a replacement interview introduces systematic bias to your data because you tend to replace nonrespondents with people who are like respondents rather than with people who are like nonrespondents (1989:53).

Telephone survey researchers typically call back from three to 10 times before replacing a member of a sample. When survey researchers suspect (from prior work) that, say, 25% of a sample won't be reachable within, say, three call-backs, they increase their original sample size by 25% so the final sample will be both the right size and representative. The reason we know this is because researchers report these kinds of compromises when they publish their results. You should, too.

Systematic Random Sampling

If you have a big, unnumbered sampling frame, like the 51,413 students at the University of Florida in 2008, then simple random sampling is nearly impossible. You would have to number all those names first. Instead, you can do systematic random sampling. For this, you need a random start and a sampling interval, N. You enter the sampling frame at the random start and take every Nth person (or item) in the frame. If you have a printout of 51,413 names, listed 400 to a page, select a single random number between 1 and 51,413. If the random number is 9,857, the listing will be 257 names down from the top of page 25.

The sampling interval depends on the size of the population and the number of units in your sample. If there are 51,413 people in the population, and you are sampling 400 of them, then after you enter the sampling frame (the list of 51,413 names) you need to take every 128th person (400 × 128 = 51,200) to ensure that every person has *at least one chance* of being chosen. If there are 640 people in a population, and you are sampling 200 of them, then you would take every 4th person. If you get to the end of the list and you are at number 2 in an interval of 4, just go to the top of the list, start at 3, and keep on going (Box 5.3).

Box 5.3 Periodicity and systematic sampling

Systematic sampling *usually* produces a representative sample, but be aware of the **periodicity** problem. Suppose you're studying a big retirement community in South Florida. The development has 30 identical buildings. Each has six floors, with 10 apartments on each floor, for a total of 1,800 apartments. Now suppose that each floor has one big corner apartment that costs more than the others and attracts a slightly more affluent group of buyers.

If you do a systematic sample of every 10th apartment then, depending on where you entered the list of apartments, you'd have a sample of 180 corner apartments or no corner apartments at all.

David and Mary Hatch (1947) studied a sample of 413 wedding announcements taken from the Sunday society pages of the *New York Times* for June from 1932 to 1942. They found that 238, or about 58% of the announcements, were about weddings in an Episcopal church. They noted that only 2.5% of the population of New York City was Episcopalian at the time, and concluded that the Episcopalian church represented the elite of New York. Cahnman (1948) pointed out that the Hatches had studied only June issues of the *Times*. It seemed reasonable. After all, aren't most society weddings in June? Well, yes. Christian weddings. Upper-class Jews married in other months. The *Times* covered those weddings, but the Hatches missed them. The original article reported 25 weddings (6%) in a Catholic church. (At the time, Catholics were still mostly working-class Irish and Italians.) And Jews weren't on the Hatches' radar because they didn't even note the absence of Jewish weddings.

You can avoid the periodicity problem by doing simple random sampling, but if that's not possible, another solution is to make two systematic passes through the population using different sampling intervals. Then you can compare the two samples on a few independent variables, like age or years of education. Any differences should be attributable to sampling error. If they're not, then you might have a periodicity problem.

STRATIFIED SAMPLING

Stratified random sampling ensures that key subpopulations are included in your sample. You divide a population (a sampling frame) into subpopulations (subframes), based on key independent variables and then take a random (unbiased), sample from each of those subpopulations. You might divide the population into men and women, or into rural and urban subframes—or into key age groups (18–34, 35–49, etc.) or key income groups. As the main sampling frame gets divided by key *independent* variables, the subframes presumably get more and more homogeneous with regard to the key *dependent* variable in the study.

In 2009, for example, the Qunnipiac University Poll asked a representative sample of 2,041 registered voters in the United States the following question: Do you think abortion should be legal in all cases, legal in most cases, illegal in most cases or illegal in all cases? Across all voters, 52% said that abortion should be legal in all (15%) or most (37%) cases and 41% said it should be illegal in all (14%) or most (27%) cases. The remaining 7% had no opinion.

These facts hide some important differences across religious, political, and other subgroups. Among Catholic voters, 50% said that abortion should be legal in all (8%) or most (42%) cases; among Jewish voters, 86% said that abortion should be legal in all (33%) or most (53%) cases. Among registered Democrats, 66% favored legal abortion in all or most cases; among registered Republicans, 30% took that position (Quinnipiac University 2009). Sampling from smaller chunks (by age, gender, and so on) ensures not only that you capture the variation but that you also wind up understanding how that variation is distributed.

This is called **maximizing the between-group variance** and **minimizing the within-group variance** for the independent variables in a study. *It's what you want to do in building a sample* because it reduces sampling error and thus makes samples more precise.

This sounds like a great thing to do, but you have to *know what the key independent variables are*. Shoe size is almost certainly not related to what people think is the ideal number of children to have. Gender and generation, however, seem like plausible variables on which to stratify a sample for a study of ideal family size. So, if you are taking a poll to find this number, you might divide the adult population into, say, four generations: 15–29, 30–44, 45–59, and over 59.

With two genders, this creates a **sampling design** with eight strata: men 15–29, 30–44, 45–59, and over 59; women 15–29, 30–44, 45–59, and over 59. Then you take a random sample of people from each of the eight strata and run your poll. If your hunch about the importance of gender and generation is correct, you'll find the attitudes of men and the attitudes of women more homogeneous than the attitudes of men and women thrown together. Table 5.1 shows the distribution of gender and age **cohorts** for the United States in 2008. The numbers are in thousands. The numbers in parentheses are percentages of the total population 15 and older.

A **proportionate stratified random sample** of 2,400 respondents, 15 and older, would

Table 5.1 Gender and Age Cohorts for the U.S. in 2008

Age cohort	Males	Females	Total
15–29	33,132 (13%)	31,836 (13%)	64,968 (26%)
30–44	31,008 (13%)	30,668 (12%)	61,676 (25%)
45–59	31,442 (13%)	32,779 (13%)	64,221 (26%)
>59	25,316 (10%)	31,671 (13%)	56,987 (23%)
Total	120,898 (49%)	126,954 (51%)	247,852 (100%)

Source: Table 8, Statistical Abstract of the United States (2010).

include 312 men between the ages of 30 and 44 (13% of 2,400 = 312), but 288 women between the ages of 30 and 44 (12% of 2,400 = 288), and so on.

Watch out, though. We are accustomed to thinking in terms of gender on questions about family size, but gender-associated preferences are changing rapidly in late industrial societies, and we might be way off base in our thinking. Separating the population into gender strata might just be creating unnecessary work. Worse, it might introduce unknown error. If your guess about age and gender being related to desired number of children is wrong, then using Table 5.1 to create a sampling design will just make it harder for you to discover your error (Box 5.4).

Box 5.4 The rules on stratifying samples

Here are the rules on stratification: (1) If differences on a dependent variable are large across strata like age, sex, ethnic group, and so on, then stratifying a sample is a great idea. (2) If differences are small, then stratifying just adds unnecessary work. (3) If you are uncertain about the independent variables that could be at work in affecting your dependent variable, then leave well enough alone and don't stratify the sample. *You can always stratify the data* you collect and test various stratification schemes in the analysis instead of in the sampling.

Disproportionate Sampling

Disproportionate stratified random sampling is appropriate whenever an important subpopulation is likely to be underrepresented in a simple random sample or in a stratified random sample. Native Americans (including American Indians and Alaska Natives) comprise just 1.3% of the population of the United States. If you take 1,000 samples of 1,000 Americans at random, you expect to run into about 13 Native Americans, *on average, across all the samples*. (Some samples will have no Native Americans and some may have 20, but on average you'll get about 13.) Without disproportionate sampling, Native Americans would be underrepresented in any national survey in the United States.

Bachman et al. (2010) wanted to estimate the problem of sexual assault and rape among Native Americans. The U.S. National Crime Victimization Survey canvasses between 67,000 and 100,000 people every year and includes questions for women about rape and sexual assault. But even this massive survey didn't have enough cases for reliable analysis, so Bachman et al. aggregated all the surveys between 1992 and 2005.

Suppose you are doing a study of factors affecting grade-point averages among a population of 8,000 college students. You suspect that the independent variable called "race" is associated in some way with the dependent variable.

Suppose further that 5% of the student population is African American and that you have time and money to interview 400 students out of the population of 8,000. If you took 10,000 samples of 400 each from the population (replacing the 400 each time, of course), then the average number of African Americans in all the samples would approach 20—that is, 5% of the sample.

But you are going to take *one* sample of 400. It might contain exactly 20 (5%) African Americans; on the other hand, it might contain just five (1.25%) African Americans. To ensure that you have sufficient data on African American students and on White students, you put the African Americans and the Whites into separate *strata* and draw two random samples of 200 each. The African Americans are disproportionately sampled by a factor of 10 (200 instead of the expected 20).

This was the problem that Lieber and Fox (2005) faced in their study of how being African American in Iowa influences decisions in the juvenile courts system—from the decision to detain someone in the first place to the decision on length and type of sentence and the many decisions in between. The population of Iowa is 2.7% African American, compared to 13% in the United States as a whole. Lieber and Fox's sampling frame was all juvenile court referrals during a 21-year period, from 1980 to 2000. They selected 5,554 cases at random, of which 30% (1,666 cases) were African Americans.

This disproportionate random sampling procedure ensured that there would be a minimum number of African American cases for each of the decision stages. It also meant that there was a deliberately created, known source of bias—lack of proportionate representation in the subgroups—which had to be taken into account in data analysis. Which brings us to weighting of data.

Weighting Results

One popular method for collecting data about daily activities is called "experience sampling" (Csikszentmihalyi and Larson 1987; Hektner et al. 2007). You give a sample of people a beeper or a cell phone. They carry it around and you beep or call them at random times during the day. They fill out a little form (either on paper or on a smartphone or PDA that's been programmed with a form) about what they're doing at the time.

We'll look at these kinds of methods in Chapter 14. For now, suppose you want to contrast what people do on weekends and what they do during the week. If you beep people, say, eight times during each day, you'll wind up with 40 reports for each person for the five-day workweek but only 16 forms for each person for each two-day weekend because you've sampled the two strata—weekdays and weekends—proportionately.

If you want more data points for the weekend, you might beep people 12 times on Saturday and 12 times on Sunday. That gives you 24 data points, but you've disproportionately sampled one stratum. The weekend represents 2/7, or 28.6% of the week, but you've got 64 data points and 24 of them, or 37.5%, are about the weekend. Before comparing any data across the strata, you need to make the weekend data and the weekday data statistically comparable.

This is where **weighting** comes in. Multiply each weekday data point by 1.50 so that the 40 data points become worth 60 and the 24 weekend data points are again worth exactly 2/7 of the total.

Known sources of bias also occur by accident, like when you have unequal response rates in a stratified sample. Suppose you sample 200 men and 200 women for a survey in a factory that employs 60% women and 40% men. Of the 400 potential respondents, 178 men and 163 women respond to your questions. If you compare the answers of men and women on a variable, first, weight each man's data by 178/163 = 1.09 times each woman's data on that variable.

That takes care of the unequal response rates. Then weight each woman's data as counting 1.5 times each man's data on the variable. That takes care of the fact that there are half again as many women employees as there are men.

This may seem complicated, but weighting is a simple procedure available in all major statistical analysis packages.

CLUSTER SAMPLING AND COMPLEX SAMPLING DESIGNS

Cluster sampling is based on the fact that people act out their lives in more or less natural groups, or clusters, like geographic areas (counties, precincts, states), and institutions (like schools, churches, brotherhoods, credit unions, and so on). By sampling from these clusters, we narrow the sampling field from large, heterogeneous chunks to small, homogeneous ones that are relatively easy to find. This minimizes travel time in reaching scattered units of data collection. It also lets you sample populations for which there are no convenient lists or frames.

For example, there are no lists of schoolchildren in large cities, but children cluster in schools. There *are* lists of schools, so you can take a sample of them, and then sample children within each school selected.

Laurent et al. (2003) wanted to assess the rate of sexually transmitted diseases among unregistered female sex workers in Dakar, Senegal. By definition, unregistered means no list, so the researchers used a two-stage cluster sample. They created a sampling frame of all registered and all clandestine bars in Dakar, plus all the unregistered brothels and all the nightclubs. They did this over a period of several months with the help of some women prostitutes, some local physicians who had treated those women, and two social workers, each of whom had worked with female sex workers for over 25 years. Laurent et al. calculated that they needed 94 establishments, so they chose a simple random sample of places from the list of 183. Then they went in teams to each of the 94 places and interviewed all the unregistered prostitutes who were there at the time of the visit.

Sampling designs can involve several stages. If you are studying Haitian refugee children in Miami, you could take a random sample of schools, but if you do that, you'll almost certainly select some schools in which there are no Haitian children. A three-stage sampling design is called for.

In the first stage, you would make a list of the neighborhoods in the city, find out which ones are home to a lot of refugees from Haiti, and sample those districts. In the second stage, you would take a random sample of schools from each of the chosen districts. Finally, in the third stage, you would develop a list of Haitian refugee children in each school and draw your final sample.

Al-Nuaim et al. (1997) used multistage stratified cluster sampling in their national study of adult obesity in Saudi Arabia. In the first stage, they selected cities and villages from each region of the country so that each region's total population was proportionately represented. Then they randomly selected districts from the local maps of the cities and villages in their sample. Next, they listed all the streets in each of the districts and selected every third street. Then they chose every third house on each of the streets and asked each adult in the selected houses to participate in the study.

PROBABILITY PROPORTIONATE TO SIZE

The best estimates of a parameter are produced in samples taken from clusters of equal size. When clusters are not equal in size, then samples should be taken **PPS—with probability proportionate to size.**

Suppose you had money and time to do 800 household interviews in a city of 50,000 households. You intend to select 40 blocks, out of a total of 280, and do 20 interviews in each block. You want each of the 800 households in the final sample to have exactly the same probability of being selected.

Should each block be equally likely to be chosen for your sample? No, because census blocks never contribute equally to the total population from which you will take your final sample. A block that has 100 households in it *should* have twice the chance of being chosen for 20 interviews as a block that has 50 households, and half the chance of a block that has 200 households.

When you get down to the block level, each household on a block with 100 residences has a 20% (20/100) chance of being selected for the sample; each household on a block with 300 residences has only a 6.7% (20/300) chance of being selected.

Lené Levy-Storms wanted to talk to older Samoan women in Los Angeles County about mammography. The problem was not that women were reticent to talk about the subject. The problem was how do you find a representative sample of older Samoan women in Los Angeles County?

From prior ethnographic research, Levy-Storms knew that Samoan women regularly attend churches where the minister is Samoan. She went to the president of the Samoan Federation of America in Carson, California, and he suggested nine cities in L.A. County where Samoans were concentrated. There were 60 churches with Samoan ministers in the nine cities, representing nine denominations. Levy-Storms asked each of the ministers to estimate the number of female church members who were over 50 years old. Based on these estimates, she chose a PPS sample of 40 churches (so that churches with more or fewer older women were properly represented). This gave her a sample of 299 Samoan women over 50. This clever sampling strategy really worked: Levy-Storms contacted the 299 women and wound up with 290 interviews—a 97% cooperation rate (Levy-Storms and Wallace 2003).

PPS sampling is called for under three conditions: (1) when you are dealing with large, unevenly distributed populations (such as cities that have high-rise and single-family neighborhoods); (2) when your sample is large enough to withstand being broken up into a lot of pieces (clusters) without substantially increasing the sampling error; and (3) when you have data on the population of many small blocks in a population and can calculate their respective proportionate contributions to the total population.

PPS Samples in the Field—Space Sampling

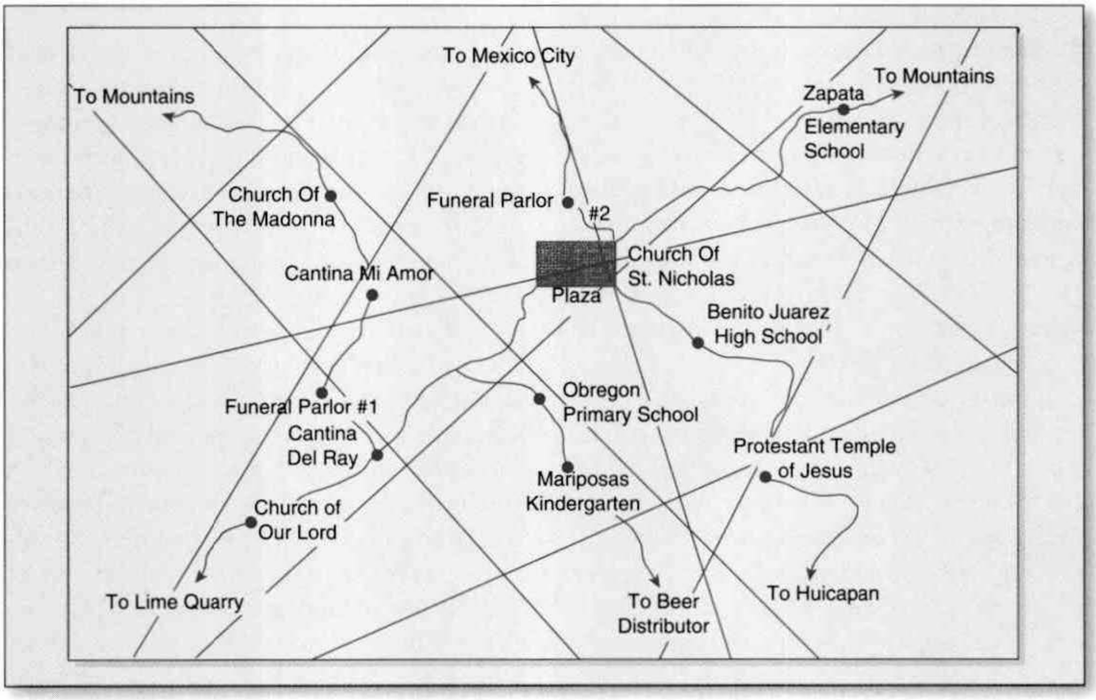
What do you do when you don't have neat clusters and neat sampling frames printed out on a computer by a reliable government agency? The answer is to place your trust in randomness and *create* maximally heterogeneous clusters from which to take a random sample using space sampling or map sampling.

The map sampling method is adapted from transect sampling in wildlife biology (Burnham et al. 1980). Draw or get a map of the area you are studying. Place 100 numbered dots around the edge of the map. Try to space the numbers equidistant from one another, but don't worry if they are not. Select a pair of numbers at random and draw a line between them. Now select another pair of numbers (be sure to replace the first pair before selecting the second), and draw a line between them. In the unlikely event that you choose the same pair twice, simply choose a third pair. Keep doing this, replacing the numbers each time. After you've drawn about 50 lines, you can begin sampling.

Notice that the lines drawn across the map in Figure 5.1 create a lot of wildly uneven spaces. Since you don't know the distribution of population density in the area you are studying, this technique maximizes the chance that you will properly survey the population, more or less PPS. By creating a series of (essentially) random chunks of different sizes, you distribute the error you might introduce by not knowing the density, and that distribution lowers the possible error.

Number the uneven spaces created by the lines and choose some of them at random. Go to those spaces, number the households, and select an appropriate number at random. Remember, you want to have the same number of households from *each* made-up geographic cluster, no matter what its size. If you are doing 400 interviews, you would select 20 geographic chunks and do 20 interviews or behavioral observations in each.

Figure 5.1 Sampling Map



My colleagues and I used this method in 1986 to find out how many people in Mexico City knew someone who died in that city's monster earthquake the year before (Bernard, Johnson et al. 1989). Instead of selecting households, though, my interview team went to each geographic chunk we'd selected and stopped the first 10 people they ran into on the street at each point. This is called a street-intercept survey (Box 5.5).

Box 5.5 Street- and mall-intercept sampling

K. W. Miller et al. (1997) sampled blocks of streets in a city and did a street-intercept survey of African American men. They compared the results to a random-digit dialing telephone survey in the same city. The street-intercept survey did a better job of representing the population than did the telephone survey. For one thing, the response rate for the street intercept survey was over 80%.

Compare that to the typical telephone survey, where half or more of the respondents may refuse to be interviewed. Also, with telephone surveys, the socioeconomic profile of respondents is generally higher than in the population (partly because more affluent people agree more often to be interviewed on the telephone). A variant of this method is mall-intercept sampling, used widely in marketing (**Further Reading:** street and mall intercept surveys).

Handwerker (1993) used a map-sampling method in his study of sexual behavior on Barbados. In his variation of map sampling, you generate 10 random numbers between 0 and 360 (the degrees on a compass). Next, put a dot in the center of a map that you will use for the sampling exercise, and use a protractor to identify the 10 randomly chosen compass points. You then draw lines from the dot in the center of the map through all 10 points to the edge of the map and interview people (or observe houses, or whatever) along those lines. (See Duranleau [1999] for an empirical test of the power of map sampling.)

If you use this technique, you may want to establish a sampling interval (like every fifth case, beginning with the third case). If you finish interviewing along the lines and don't have enough cases, you can take another random start, with the same or a different interval and start again. Be careful of periodicity, though (Box 5.6).

Box 5.6 Combining map sampling and cluster sampling

In Chapter 4, I mentioned a study in which Lambros Comitas and I compared Greeks who had returned from what was then West Germany as labor migrants with Greeks who had never left their country (Bernard and Comitas 1978). There were no lists of returned migrants, but we thought we could do a cluster sample by locating the children of returned migrants in the Athens schools and then use the children to select a sample of their parents.

The problem was that we couldn't even get a list of schools in Athens. So we took a map of the city and divided it into small bits by laying a grid over it. Then we took a random sample of the bits and sent interviewers to find the school nearest each bit selected. The interviewers asked the principal of each school to identify the children of returned labor migrants. (It was easy for the principal to do, by the way. The principal said that all the returned migrant children spoke Greek with a German accent.) That way, we were able to make up two lists for each school: one of children who had been abroad, and one of children who had not. By sampling children randomly from those lists at each school, we were able to select a representative sample of parents.

Camilla Harshbarger (1995) used another variation of map sampling in her study of farmers in North West Province, Cameroon (1993). To create a sample of 400 farmers, she took a map of a rural community and drew 100 dots around the perimeter. She used a random number table to select 50 pairs of dots and drew lines between them. She numbered the points created by the crossing of lines and chose 80 of those points at random. Then Harshbarger and her field assistants interviewed one farmer in each of the five compounds they found closest to each of the 80 selected dots. (If you use this dot technique, remember to include the points along the edges of the map in your sample or you'll miss households on those edges.)

There are times when a random, representative sample is out of the question. After she did those interviews with 400 randomly selected farmers in North West Province, Cameroon, Harshbarger set out to interview Fulani cattle herders in the same area. Here's what Harshbarger wrote about her experience in trying to interview the herders:

> It was rainy season in Wum and the roads were a nightmare. The graziers lived very far out of

> town and quite honestly, my research assistants were not willing to trek to the compounds because it would have taken too much time and we would never have finished the job. I consulted X and he agreed to call selected people to designated school houses on certain days. We each took a room and administered the survey with each individual grazier.
>
> Not everyone who was called came for the interview, so we ended up taking who we could get. Therefore, the Wum grazier sample was not representative and initially that was extremely difficult for me to accept. Our team had just finished the 400-farmer survey of Wum that *was* representative, and after all that work it hit me hard that the grazier survey would not be. To get a representative sample, I would have needed a four-wheel drive vehicle, a driver, and more money to pay research assistants for a lengthy stay in the field. Eventually, I forgave myself for the imperfection. [personal communication]

The lessons here are clear. (1) If you are ever in Harshbarger's situation, you, too, can forgive yourself for having a nonrepresentative sample. (2) Even then, like Harshbarger, you should feel badly about it.

Maximizing Between-Group Variance: The Wichita Study

Whenever you do multistage cluster sampling, be sure to take as large a sample as possible from the largest, most heterogeneous clusters. The larger the cluster, the larger the **between-group variance**; the smaller the cluster, the higher the **within-group variance**.

Counties in the United States are more like each other on any variable (income, race, average age, whatever) than states are; towns within a county are more like each other than counties are; neighborhoods in a town are more like each other than towns are; blocks are more like each other than neighborhoods are. In sampling, the rule is: *maximize between-group variance*.

What does this mean in practice? The following is an actual example of multistage sampling from John Hartman's study of Wichita, Kansas (Hartman 1978; Hartman and Hedblom 1979:160ff.). At the time of the study, in the mid-1970s, Wichita had a population of about 193,000 persons over 16. This was the population to which the study team wanted to generalize. The team decided that they could afford only 500 interviews. There were 82 census tracts in Wichita, from which they randomly selected 20. These 20 tracts then became the actual population of their study. We'll see in a moment how well their actual study population simulated (represented) the study population to which they wanted to generalize.

Hartman and Hedblom added up the total population in the 20 tracts and divided the population of *each tract* by the total. This gave the percentage of people that each tract, or cluster, contributed to the new population total. Since the researchers were going to do 500 interviews, each tract was assigned that percentage of the interviews. If there were 50,000 people in the 20 tracts, and one of the tracts had a population of 5,000, or 10% of the total, then 50 interviews (10% of the 500) would be done in that tract.

Next, the team numbered the blocks in each tract and selected blocks at random until they had enough for the number of interviews that were to be conducted in that tract. When a block was selected it stayed in the pool, so that in some cases more than one interview was to be conducted in a single block. This did not happen very often, and the team wisely left it up to chance to determine this.

This study team made some excellent decisions that maximized the heterogeneity (and hence the representativeness) of their sample. As clusters get smaller and smaller (as you go from tract to block to household, or from village to neighborhood to household), the homogeneity of the units of analysis within the clusters gets greater and greater. People in one census tract or village are more like each other than people in different tracts or villages. People in one census block or barrio are more like each other than people across blocks or barrios. And people in households are more like each other than people in households across the street or over the hill.

This is very important. Most researchers would have no difficulty with the idea that they should only interview one person in a household because, for example, husbands and wives often have similar ideas about things and report similar behavior with regard to kinship, visiting, health care, child care, and consumption of goods and services. Somehow, the lesson becomes less clear when new researchers move into clusters that are larger than households.

But the rule stands: Maximize heterogeneity of the sample by taking as many of the biggest clusters in your sample as you can, and as many of the next biggest, and so on, always at the expense of the number of clusters at the bottom where homogeneity is greatest. Take more tracts or villages, and fewer blocks per tract or barrios per village. Take more blocks per tract or barrios per village, and fewer households per block or barrio. Take more households and fewer persons per household.

Many survey researchers say that, as a rule, you should have no fewer than five households in a census block. The Wichita group did not follow this rule but only had enough money and person power to do 500 interviews and they wanted to maximize the likelihood that their sample would represent faithfully the characteristics of the 193,000 adults in their city.

The Wichita study group drew two samples—one main sample and one alternate sample.

Whenever they could not get someone on the main sample, they took the alternate. That way, they maximized the representativeness of their sample because the alternates were chosen with the same randomized procedure as the main respondents in their survey. They were not forced to take next-door neighbors when a main respondent wasn't home. (This kind of "winging it" in survey research has a tendency to clobber the representativeness of samples. In the United States, at least, interviewing only people who are at home during the day produces results that represent women with small children, shut-ins, telecommuters, and the elderly—and not much else.)

Next, the Wichita team randomly selected the households for interview within each block. This was the third stage in this multistage cluster design. The fourth stage consisted of flipping a coin to decide whether to interview a man or a woman in households with both. Whoever came to the door was asked to provide a list of those in the household over 16 years of age. If there was more than one eligible person in the household, the interviewer selected one at random, conforming to the decision made earlier on sex of respondent.

Table 5.2 shows how well the Wichita team did. All in all, they did very well. In addition to the variables shown in the table here, the Wichita sample was a fair representation of marital status, occupation, and education, although there were some pretty large discrepancies on this last independent variable. For example, according to the 1970 census, 8% of the population of Wichita had less than eight years of schooling, but only 4% of the sample had this characteristic. Only 14% of the general population had completed one–three years of college, but 22% of the sample had that much education.

All things considered, though, the sampling procedure followed in the Wichita study was a model of technique, and the results show it. Whatever they found out about the 500 people they interviewed, the researchers could be very confident that the results were generalizable to the 193,000 adults in Wichita.

In sum: If you don't have a sampling frame for a population, try to do a multistage cluster sample, narrowing down to natural clusters that do have lists. Sample heavier at the higher levels in a multistage sample and lighter at the lower stages.

Table 5.2 Comparison of Survey Results and Population Parameters for the Wichita Study by Hartman and Hedblom

	Wichita in 1973	Hartman and Hedblom's Sample for 1973 (in Percentages)
White	86.8	82.8
African	9.7	10.8
Chicano	2.5	2.6
Other	1.0	2.8
Male	46.6	46.9
Female	53.4	53.1
Median age	38.5	39.5

Source: J. J. Hartman and J. H. Hedblom, *Methods for the Social Sciences: A Handbook for Students and Non-Specialists*, p. 165, 1979, Greenwood Publishing Company.

HOW BIG SHOULD A SAMPLE BE?

There are two things you can do to get good samples. You can ensure sample accuracy by making sure that every element in the population has an equal chance of being selected—that is, you can make sure the sample is unbiased. You can ensure sample precision by increasing the size of unbiased samples. We've already discussed the importance of how to make samples unbiased. The next step is to decide how big a sample needs to be.

Sample size depends on: (1) the heterogeneity of the population or chunks of population (strata or clusters) from which you choose the elements; (2) how many population subgroups (that is, independent variables) you want to deal with simultaneously in your analysis; (3) the size of the phenomenon that you're trying to detect; and (4) how precise you want your sample statistics (or parameter estimators) to be.

1. Heterogeneity of the population. When all elements of a population have the same score on some measure, a sample of 1 will do. Ask a lot of people to tell you how many days there are in a week and you'll soon understand that a big sample isn't going to uncover a lot of heterogeneity. But if you want to know what the average ideal family size is, you may need to cover a lot of social ground. People of different ethnicities, religions, incomes, genders, and ages may have very different ideas about this. (In fact, these independent variables may interact in complex ways. Multivariate analysis tells you about this interaction. We'll get to this in Chapter 22.)

2. The number of subgroups in the analysis. Remember the factorial design problem in Chapter 4 on experiments? We had three independent variables, each with two attributes, so we needed eight groups ($2^3 = 8$). It wouldn't do you much good to have, say, one experimental subject in each of those eight groups. If you're going to analyze all eight of the conditions in the experiment, you've got to fill each of the conditions with some reasonable number of subjects. If you have only 15 people in each of the eight conditions, then you need a sample of 120.

The same principle holds when you're trying to figure out how big a sample you need for a survey. If you have four age groups and two genders, you wind up with an eight-cell sampling design.

If all you want to know is a single proportion—like what percentage of people in a population approve or disapprove of something—then you need about 100 respondents to be 95% confident, within plus or minus three points, that your sample estimate is within two standard deviations of the population parameter (more about confidence limits, normal distributions, standard deviations, and parameters coming up in the next chapter). But if you want to know whether retired widowers who have less than $3,000 per month in total income have different opinions from, say, working, married mothers who have more than $3,000 per month in total income, then you'll need a bigger sample.

3. The size of the subgroup. If the population you are trying to study is rare and hard to find, and if you have to rely on a simple random sample of the entire population, you'll need a very large initial sample. A needs assessment survey of people over 75 in Florida took 72,000 phone calls to get 1,647 interviews—about 44 calls per interview (Henry 1990:88). This is because only 6.5% of Florida's population was over 75 at the time of the survey. By contrast, the monthly Florida survey of 600 representative consumers in that state takes about 5,000 calls (about eight per interview). That's because just about everyone in the state 18 and older is a consumer and is eligible for the survey (Christopher McCarty, personal communication).

The smaller the difference on any measure between two populations, the bigger the sample you need to detect that difference. Suppose you suspect that Blacks and Whites in a prison system have received different sentences for the same crime. Henry (1990:121) shows that a difference of 16 months in sentence length for the same crime would be detected with a sample of just 30 in each racial group (if the members of the sample were selected randomly, of course). To detect a difference of three months, however, you need 775 in each group.

4. Precision. This one takes us into sampling theory.

Key Concepts in This Chapter

individual data
cultural data
population parameters
probability sampling
nonprobability sampling
random sample
panel study
sampling frame
population
unbiased sample
simple random sample
systematic random sample
random start
sampling interval
periodicity
stratified random sapling
maximizing the between-group variance
minimizing the within-group variance
sampling design
age cohort
proportionate stratified random sample
disproportionate stratified random sample
weighting
cluster sampling
PPS-probability proportionate to size
space sampling
map sampling
street-intercept survey
between-group variance
within-group variance
sample accuracy
sample precision
sample statistics
parameter estimators
population heterogeneity
factorial design

Summary

- There are two kinds of data of interest to social scientists: individual data and cultural data. These two kinds of data require different approaches to sampling.
  - Individual data are about attributes of individuals in a population. To estimate the parameters of these attributes in a population requires probability sampling.
  - Cultural data requires experts, which means relying on nonprobability sampling.
- There are several ways to take probability samples.
  - Simple random sampling involves generating a list of random numbers and applying that list to a numbered sampling frame. (Most researchers actually take systematic, rather than simple random samples.)
  - Stratified random samples are used to ensure that key subpopulations are included in a study. Disproportionate stratified random sampling is used to ensure that important but relatively small subpopulations are included in a sample.
  - Cluster sampling is used when there is no overall sampling frame. Cluster sampling is based on the fact that people live in natural clusters (counties, states, etc.) and they participate in the activities of institutions (schools, churches, credit unions, etc.).
  - The best estimates of a parameter are produced in samples taken from clusters of equal size. When clusters are not equal in size, then samples should be taken PPS—with probability proportionate to size.
- Sample size depends on: (1) the heterogeneity of the population from which you choose the elements; (2) how many population subgroups you want to deal with simultaneously in your analysis; (3) the size of the phenomenon that you're trying to detect; and (4) how precise you want your parameter estimators to be. Precision involves sampling theory.

Exercises

1. Record in a spreadsheet, like Excel®, as many variables as you can about your contacts on a networking site, like Facebook—things like age, sex, ethnicity, region where they live, etc. This gives you a sampling frame and parameters for the variables you've recorded. Take a random sample of the friends in your spreadsheet and estimate the parameter values of the variables for which you have measures. Try to create other sampling frames—for businesses in your town or for members of a church or other organization, like a sorority or fraternity, to which you belong.

2. Consider a study in which we will do 150 interviews in a town of 23,000 inhabitants. There are neighborhoods in the town, and we want an unbiased sample that represents all of the neighborhoods. One of the neighborhoods, with 12,000 residents, is much larger than the other four. If we do a PPS sample (one that takes account of the different sizes of the neighborhoods), then what is the probability that any individual in the big neighborhood will wind up in our sample?

3. Answer the following questions about sampling:

   a. What is the danger in systematic random sampling?

   b. Why are telephone books usually poor sampling frames?

   c. What is a stratified, random disproportionate sample?

d. Why do we sample more heavily among hierarchically higher (more heterogeneous) units than among lower (more homogeneous) ones?

e. What is the relation among "parameter," "estimator," and "sampling error"?

f. What are the two ways in which sampling error are reduced?

4. A multinational corporation asks you to survey their 840 midlevel managers. There are 590 men and 250 women in the cohort of managers. You decide to take a stratified random sample of 100 from each of the gender groups. What is the sampling weight for each of the strata? Hint: Find the probability, p, of sampling each man and each woman, given that you are sampling 100 of each. The weight is the inverse of p, or $1/p$.

Further Reading

General sampling. Ardilly and Tillé (2006), Czaja and Blair (2005), Dattalo (2008, 2010), Handwerker (2003), Hoyle (1999), Nardi (2003), Onwuegbuzie and Collins (2007), Stine (1990), Teddlie and Yu (2007).

Street intercept and mall intercept sampling. Bruwer and Haydam (1996), Bush and Hair (1985), Choi et al. (2008), Daley et al. (2001), Gates and Solomon (1982), Hemphill et al. (2007), Hew and Wesley (2008), Oxford et al. (2004), Ross et al. (2006), WalterMaurer et al. (2003).