Chapter 8

Sampling

Sampling is the statistical process of selecting a subset (called a "sample") of a population of interest for purposes of making observations and statistical inferences about that population. Social science research is generally about inferring patterns of behaviors within specific populations. We cannot study entire populations because of feasibility and cost constraints, and hence, we must select a representative sample from the population of interest for observation and analysis. It is extremely important to choose a sample that is truly representative of the population so that the inferences derived from the sample can be generalized back to the population of interest. Improper and biased sampling is the primary reason for often divergent and erroneous inferences reported in opinion polls and exit polls conducted by different polling groups such as CNN/Gallup Poll, ABC, and CBS, prior to every U.S. Presidential elections.

The Sampling Process

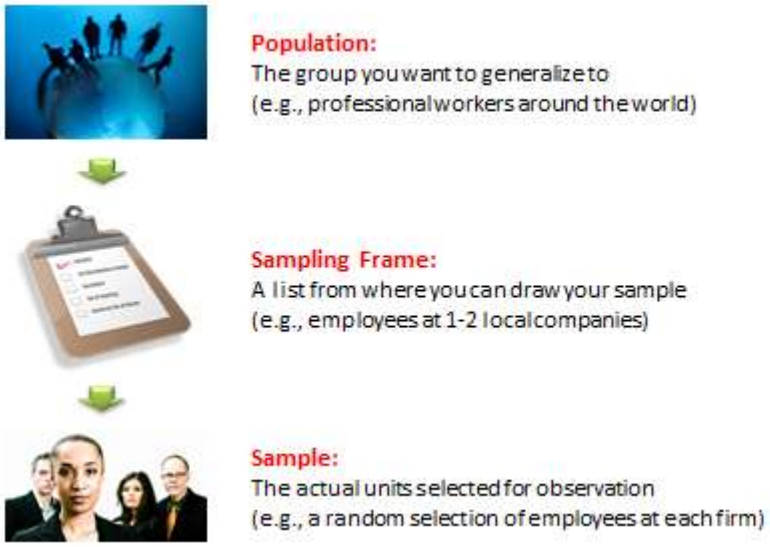
Population:
The group you want to generalize to
(e.g., professional workers around the world)
Sampling Frame:
A list from where you can draw your sample
(e.g., employees at 1-2 local companies)
Sample:
The actual units selected for observation
(e.g., a random selection of employees at each firm)

Figure 8.1. The sampling process

The sampling process comprises of several stage. The first stage is defining the target population. A **population** can be defined as all people or items (**unit of analysis**) with the characteristics that one wishes to study. The unit of analysis may be a person, group, organization, country, object, or any other entity that you wish to draw scientific inferences about. Sometimes the population is obvious. For example, if a manufacturer wants to determine whether finished goods manufactured at a production line meets certain quality requirements or must be scrapped and reworked, then the population consists of the entire set of finished goods manufactured at that production facility. At other times, the target population may be a little harder to understand. If you wish to identify the primary drivers of academic learning among high school students, then what is your target population: high school students, their teachers, school principals, or parents? The right answer in this case is high school students, because you are interested in their performance, not the performance of their teachers, parents, or schools. Likewise, if you wish to analyze the behavior of roulette wheels to identify biased wheels, your population of interest is not different observations from a single roulette wheel, but different roulette wheels (i.e., their behavior over an infinite set of wheels).

The second step in the sampling process is to choose a **sampling frame**. This is an accessible section of the target population (usually a list with contact information) from where a sample can be drawn. If your target population is professional employees at work, because you cannot access all professional employees around the world, a more realistic sampling frame will be employee lists of one or two local companies that are willing to participate in your study. If your target population is organizations, then the Fortune 500 list of firms or the Standard & Poor's (S&P) list of firms registered with the New York Stock exchange may be acceptable sampling frames.

Note that sampling frames may not entirely be representative of the population at large, and if so, inferences derived by such a sample may not be generalizable to the population. For instance, if your target population is organizational employees at large (e.g., you wish to study employee self-esteem in this population) and your sampling frame is employees at automotive companies in the American Midwest, findings from such groups may not even be generalizable to the American workforce at large, let alone the global workplace. This is because the American auto industry has been under severe competitive pressures for the last 50 years and has seen numerous episodes of reorganization and downsizing, possibly resulting in low employee morale and self-esteem. Furthermore, the majority of the American workforce is employed in service industries or in small businesses, and not in automotive industry. Hence, a sample of American auto industry employees is not particularly representative of the American workforce. Likewise, the Fortune 500 list includes the 500 largest American enterprises, which is not representative of all American firms in general, most of which are medium and small-sized firms rather than large firms, and is therefore, a biased sampling frame. In contrast, the S&P list will allow you to select large, medium, and/or small companies, depending on whether you use the S&P large-cap, mid-cap, or small-cap lists, but includes publicly traded firms (and not private firms) and hence still biased. Also note that the population from which a sample is drawn may not necessarily be the same as the population about which we actually want information. For example, if a researcher wants to the success rate of a new "quit smoking" program, then the target population is the universe of smokers who had access to this program, which may be an unknown population. Hence, the researcher may sample patients arriving at a local medical facility for smoking cessation treatment, some of whom may not have had exposure to this particular "quit smoking" program, in which case, the sampling frame does not correspond to the population of interest.

The last step in sampling is choosing a sample from the sampling frame using a well-defined sampling technique. Sampling techniques can be grouped into two broad categories: probability (random) sampling and non-probability sampling. Probability sampling is ideal if generalizability of results is important for your study, but there may be unique circumstances where non-probability sampling can also be justified. These techniques are discussed in the next two sections.

Probability Sampling

Probability sampling is a technique in which every unit in the population has a chance (non-zero probability) of being selected in the sample, and this chance can be accurately determined. Sample statistics thus produced, such as sample mean or standard deviation, are unbiased estimates of population parameters, as long as the sampled units are weighted according to their probability of selection. All probability sampling have two attributes in common: (1) every unit in the population has a known non-zero probability of being sampled, and (2) the sampling procedure involves random selection at some point. The different types of probability sampling techniques include:

Simple random sampling. In this technique, all possible subsets of a population (more accurately, of a sampling frame) are given an equal probability of being selected. The probability of selecting any set of n units out of a total of N units in a sampling frame is $^{N}C_{n}$. Hence, sample statistics are unbiased estimates of population parameters, without any weighting. Simple random sampling involves randomly selecting respondents from a sampling frame, but with large sampling frames, usually a table of random numbers or a computerized random number generator is used. For instance, if you wish to select 200 firms to survey from a list of 1000 firms, if this list is entered into a spreadsheet like Excel, you can use Excel's RAND() function to generate random numbers for each of the 1000 clients on that list. Next, you sort the list in increasing order of their corresponding random number, and select the first 200 clients on that sorted list. This is the simplest of all probability sampling techniques; however, the simplicity is also the strength of this technique. Because the sampling frame is not subdivided or partitioned, the sample is unbiased and the inferences are most generalizable amongst all probability sampling techniques.

Systematic sampling. In this technique, the sampling frame is ordered according to some criteria and elements are selected at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every k^{th} element from that point onwards, where $k = N/n$, where k is the ratio of sampling frame size N and the desired sample size n, and is formally called the *sampling ratio*. It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first k elements on the list. In our previous example of selecting 200 firms from a list of 1000 firms, you can sort the 1000 firms in increasing (or decreasing) order of their size (i.e., employee count or annual revenues), randomly select one of the first five firms on the sorted list, and then select every fifth firm on the list. This process will ensure that there is no overrepresentation of large or small firms in your sample, but rather that firms of all sizes are generally uniformly represented, as it is in your sampling frame. In other words, the sample is representative of the population, at least on the basis of the sorting criterion.

Stratified sampling. In stratified sampling, the sampling frame is divided into homogeneous and non-overlapping subgroups (called "strata"), and a simple random sample is drawn within each subgroup. In the previous example of selecting 200 firms from a list of 1000 firms, you can start by categorizing the firms based on their size as large (more than 500 employees), medium (between 50 and 500 employees), and small (less than 50 employees). You can then randomly select 67 firms from each subgroup to make up your sample of 200 firms. However, since there are many more small firms in a sampling frame than large firms, having an equal number of small, medium, and large firms will make the sample less representative of the population (i.e., biased in favor of large firms that are fewer in number in the target population). This is called *non-proportional* stratified sampling because the proportion of sample within each subgroup does not reflect the proportions in the sampling frame (or the population of interest), and the smaller subgroup (large-sized firms) is *over-sampled*. An alternative technique will be to select subgroup samples in proportion to their size in the population. For instance, if there are 100 large firms, 300 mid-sized firms, and 600 small firms, you can sample 20 firms from the "large" group, 60 from the "medium" group and 120 from the "small" group. In this case, the proportional distribution of firms in the population is retained in the sample, and hence this technique is called *proportional* stratified sampling. Note that the non-proportional approach is particularly effective in representing small subgroups, such as large-sized firms, and is not necessarily less representative of the population compared to the proportional approach, as long as the findings of the non-proportional approach is weighted in accordance to a subgroup's proportion in the overall population.

Cluster sampling. If you have a population dispersed over a wide geographic region, it may not be feasible to conduct a simple random sampling of the entire population. In such case, it may be reasonable to divide the population into "clusters" (usually along geographic boundaries), randomly sample a few clusters, and measure *all* units within that cluster. For instance, if you wish to sample city governments in the state of New York, rather than travel all over the state to interview key city officials (as you may have to do with a simple random sample), you can cluster these governments based on their counties, randomly select a set of three counties, and then interview officials from *every* official in those counties. However, depending on between-cluster differences, the variability of sample estimates in a cluster sample will generally be higher than that of a simple random sample, and hence the results are less generalizable to the population than those obtained from simple random samples.

Matched-pairs sampling. Sometimes, researchers may want to compare two subgroups within one population based on a specific criterion. For instance, why are some firms consistently more profitable than other firms? To conduct such a study, you would have to categorize a sampling frame of firms into "high profitable" firms and "low profitable firms" based on gross margins, earnings per share, or some other measure of profitability. You would then select a simple random sample of firms in one subgroup, and match each firm in this group with a firm in the second subgroup, based on its size, industry segment, and/or other matching criteria. Now, you have two matched samples of high-profitability and low-profitability firms that you can study in greater detail. Such matched-pairs sampling technique is often an ideal way of understanding bipolar differences between different subgroups within a given population.

Multi-stage sampling. The probability sampling techniques described previously are all examples of single-stage sampling techniques. Depending on your sampling needs, you may combine these single-stage techniques to conduct multi-stage sampling. For instance, you can stratify a list of businesses based on firm size, and then conduct systematic sampling within each stratum. This is a two-stage combination of stratified and systematic sampling. Likewise, you can start with a cluster of school districts in the state of New York, and within each cluster, select a simple random sample of schools; within each school, select a simple random sample of grade levels; and within each grade level, select a simple random sample of students for study. In this case, you have a four-stage sampling process consisting of cluster and simple random sampling.