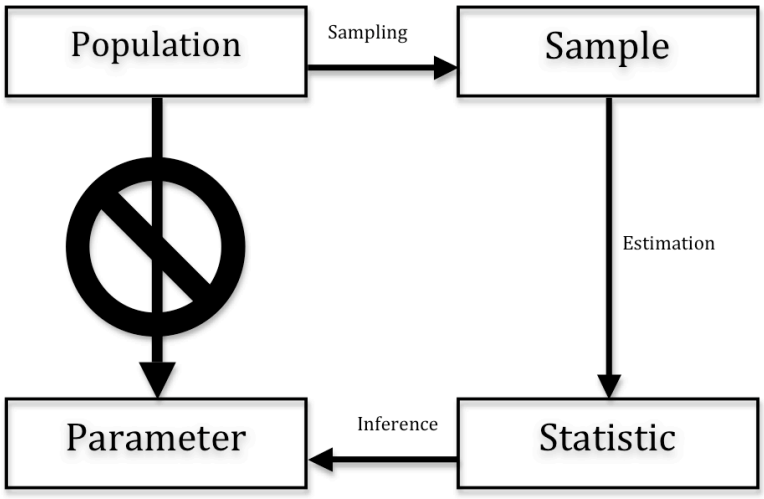
PLSC 502: "Statistical Methods for Political Research"

Sampling
October 4, 2016

Statistical Inference: The Idea

In conducting inference, the idea is to draw from information about a subgroup of a population to the population itself.

Figure 1: Research In A Single Figure



To the extent that a sample differs from its population, that difference can be due to two factors. The first, which we generically denote *bias*, is when some aspect of the sampling mechanism (or the research design in general) causes the sample to be *systematically* unrepresentative of the population in one or more relevant ways. The second, which we'll call *sampling error*, is any difference between the sample and the population that is nonsystematic, but instead is due to the randomness in the sample selection design.

As a rule, bias is a far bigger – and more complicated – threat to valid inference than is sampling error. In fact, in the end, sampling error winds up being terrifically easy to deal with, while bias (in all its forms) is a real PITA. Most of what you learn in a good Ph.D. program – theory, substance, methods, whatever – is designed to help you minimize bias in your own work.

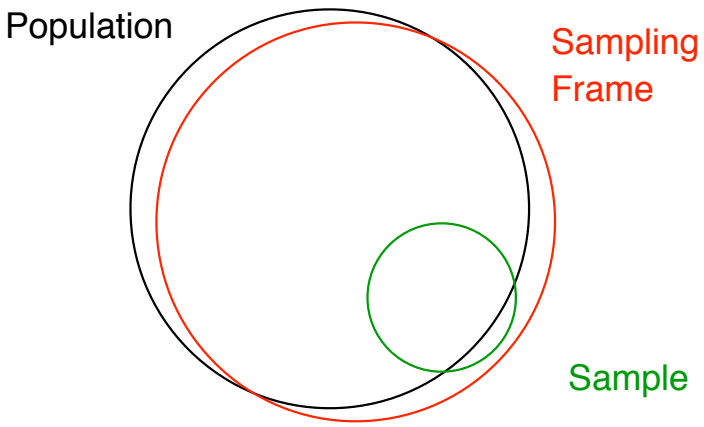
Sampling

Sampling simply means drawing a subgroup of units from some larger *population*. We'll refer to the units that make up the population as the *units of analysis*; these are the actual "things" (people, countries, boat-tailed grackles, whatever) we want to study. The population in question is often referred to as the *sampling frame*, and the sample size (the number of units in the sample) is almost always denoted by N. We'll refer to the population size (if it is known) as $\mathfrak{N}$, and the things – whatever they are – that are being sampled as the "primary sampling units" (PSUs).

Population vs. Sampling Frame

Ideally, the population and the sampling frame are identical. In practice, however, this is rarely the case; in most instances, there are units in the population that are omitted from the sampling frame, and – conversely – units in the sampling frame that are not in the population of interest. This general phenomenon is illustrated in the Figure below.

Figure 2: Population vs. Sampling Frame



For example: Suppose you wanted to know the opinions of adults (18+) in the U.S. about (say) Donald Trump, and you decide to do so by conducting a survey. You secure a list of all mobile and land line telephone numbers in the country, from which you can sample randomly (using, say, random-digit dialing). Note that:

1. There are (a small number of) adults who do not have telephones at all; those individuals are in the population, but not in the sampling frame.

2. There is also a (much larger) number of cell phones belonging to individuals under the age of 18; those individuals are in the sampling frame, but not in the population of interest.

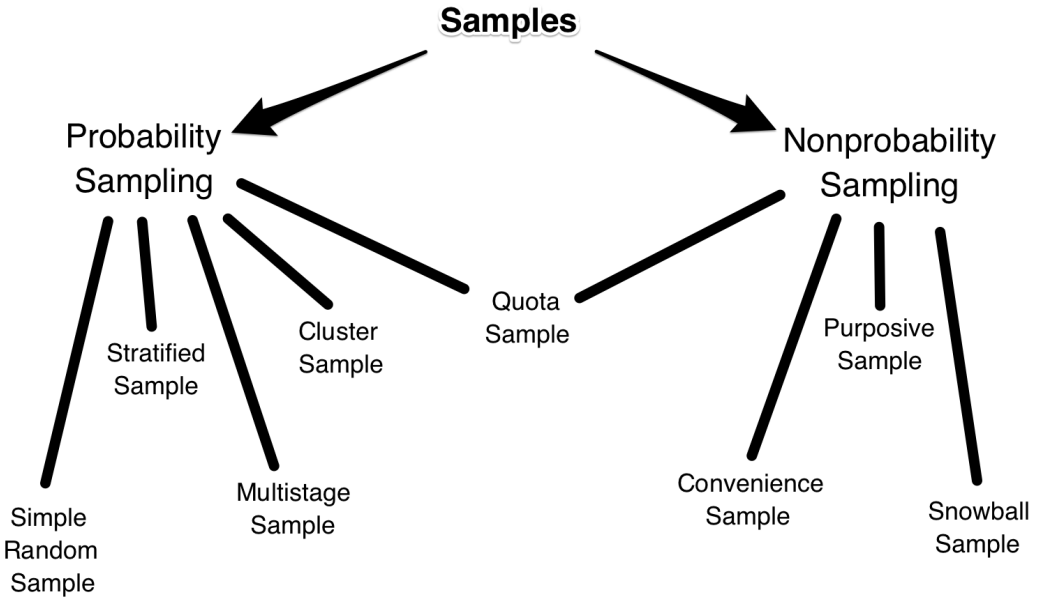
A sample is necessarily drawn from a sampling frame, *not* from the population itself. When the two are identical, no issues arise; to the extent that they are very close to identical, the issues that arise will be smaller. If, however, there is a significant difference between the sampling frame and the population, *and if that difference is non-random*, that can lead to *coverage bias* in the things you are trying to investigate.

Sampling

There are many ways to sample. We'll focus first on *probability sampling* methods of various sorts, then talk about some alternatives, as well as some other instances where randomization is valuable, and a few other things too.

In general, samples can be defined as either *probability samples* or *nonprobability samples*. The former is any sample where the probability of every unit's inclusion in a sample is *known*. If that probability is not known for every unit, then the sample is a nonprobability sample.

Among those different types, there are many variants; we'll discuss a few of the most important / widely-used ones today.



Simple Random Sampling

Simple random sampling begins with a population of size $\mathfrak{N}$, and randomly draws N respondents from the population in any fashion that ensures that *the probability of any one unit being drawn for the sample is* $1/\mathfrak{N}$. In a simple random sample, the PSUs are the same as the units of analysis themselves, be they people, mice, stars, or whatever.

In general, a simple random sample is the best sort of sample to have. It leads to the simplest means of inference, and will (in the long run) be the most representative type of sample vis-a-vis the population.

It will also, in general, be the hardest to draw. A simple random sample requires a lot on the part of the researcher; in particular, s/he must

1. Know every unit in his/her population (its existence, location, etc.), in order to assign it for sampling, and
2. Be able to actually include all selected units in the sample that is drawn.

The first of these can be a challenge any time the population in question is (a) large, (b) ever-changing, (c) amorphous (for example, what is the population of "Beatles fans"?), and/or (d) "hidden" in some way. Any of them can result in a particular form of sampling bias, one due to the inability to ensure that the probability of each unit's being sampled is $1/\mathfrak{N}$.

The second can be a problem as well. What if a survey respondent is simply unreachable (away on vacation, or a hermit on a mountain, or the like)? Or a particular lion in a pride can't be darted for tagging? Etc. This latter problem goes to the issue of *nonresponse bias*, the bias resulting from the inability of a researcher to include all selected units in their sample.

So, while simple random sampling is great, it's also hard to do in practice. Moreover, there are times when we can "do better" than a simple random sample if we are interested in particular subpopulations within the population.

Stratified Sampling

A *stratified sample* is one in which the population is divided *a priori* into two or more groups, and individuals are sampled randomly from within those groups. The groups are known as *strata*, and they are often selected on the basis of the study's subject matter. Importantly, the PSUs remain the units themselves; the researcher is simply dividing the units up by strata before doing the simple random sampling described above.

Stratified sampling is often done when there is interest in some particular characteristics of the population. Suppose we divide the population into two subgroups; call them A and B. If the proportion of group A in the sample is equal to its proportion in the population, then

we say we are taking a *proportional stratified sample*.

By contrast, if the proportion of some group A in the sample is different from its proportion in the population, then we say we have *oversampled* (or *undersampled*) that group. Oversamples are often done when one or more groups of interest are a relatively small fraction of the total population. For example, if we were interested in the political attitudes of native Americans in the United States, a simple random sample of 1,000 people in the U.S. would probably have about seven such individuals in it – hardly enough to learn anything generalizable about Native American attitudes. In such a situation, the researcher might stratify potential respondents by ethnicity / heritage, and then oversample that group to ensure adequate numbers for study.

Of course, when there is an over- or undersample, the probability of selection for any given unit is no longer $1/\mathfrak{N}$; accordingly, we have to adjust our sample to reflect the differential probabilities, by giving different *weights* to observations from different strata. More on this a bit later.

Stratified sampling can therefore improve on simple random sampling by allowing us to over- or undersample particular groups of interest. That can let us learn about groups that are a relatively small fraction of the population without requiring prohibitively large sample sizes. At the same time, stratified sampling requires (if anything) that we know *even more* about our population than does simple random sampling. That can make it somewhat harder to implement in practice.

Cluster Sampling

Cluster sampling is different from stratified sampling, in a somewhat subtle way. In a clustered sample, the units if analysis are grouped into "clusters." However, *the clusters themselves are then sampled randomly*, and all units in each selected cluster are used in the sample / study.

Cluster sampling thus changes the identity of the PSU, from the unit of analysis to the cluster (whatever that is). This means that the probability of an individual unit of analysis being sampled is no longer equal to all others; moreover, it may not even be known directly at all.

Cluster samples are used extensively; in fact, most major surveys are done via cluster sampling (often by telephone area code and exchange).

Multistage Sampling

This is a generalization of cluster sampling. The process, in a nutshell, is:

1. Select a type of "cluster," and identify subclusters of units within the cluster, etc. until we get to the "lowest" level cluster in which units are located.

2. Select – randomly or in a stratified way – some number of top-level clusters.

3. Within each selected cluster, select – again, randomly or stratifying – some number of subclusters.

4. Within subclusters, select sub-subclusters, etc.

5. At the "lowest" subcluster level, select some number of units from each sub-cluster.

The first-stage clusters are the "primary sampling units," the second stage are the "secondary sampling units," and so forth.

As an example, Agresti and Finlay talk about sampling survey respondents by first selecting blocks, then selecting houses within blocks, then selecting residents within each (selected) house. This is a three-stage design; the blocks are clusters, the houses are subclusters, and the individuals are the end units being sampled.

Multistage sampling can be useful because it can can allow us to take a probability sample without having to know the "identity" of each potential unit in the sampling frame. In the house example, we need not know exactly who lives in each house, as long as we have a rule that says (e.g.) "select one person from among those in each house with equal probability" at the last stage. Many (most) large, national surveys are actually conducted using multistage sampling – either by addresses/locations, or by telephone exchanges.

Nonprobability Samples

All of the above-mentioned means of sampling are known as *probability samples*, because one can calculate the probability of a given unit in the population being selected for the sample. That probability may be complex (for example, in multi-stage, cross-stratified designs), but it can at least be known. That, in turn, means that the researcher can adjust the resulting sample (typically by using *weighting*) so that it has the statistical properties of a simple random sample.

Nonprobability samples are samples where the probability that every unit is in the sample cannot be known. Typically, this lack of information arises either because the sample is not probabilistic at all, or because it depends on probabilities that the researcher does not know or control. As a result, the researcher cannot "correct" for any unrepresentativeness in the sampling mechanism after the fact. This means that results using nonprobability samples will be questionably representative at best, and disastrously wrong at worst. A few such sampling methods are:

1. **Convenience Sampling**, where the researcher simply samples whatever units come most readily to hand. In this case, probability is not used in the sampling at all.

2. **Purposive Sampling**, where the researcher selects units on the basis of whether s/he believes they ought to be in the sample (or to otherwise achieve some purpose). Again, here there is no use of probability at all in the sampling method.

3. **Snowball Sampling**, where the researcher selects a unit, and then other units with some relationship to that first unit are sampled, and so forth. In such a design, even if the initial sample is random / probabilistic, the probabilities of selection for subsequent units cannot be known by the researcher (even if, for example, they are known to the units themselves, as might be the case in a survey).

A borderline case is **quota sampling**. In quota sampling, the researcher selects units of various types up to some quota (for example, s/he might question 100 men and 100 women) and then stops, no longer selecting units of that type. If quotas are combined with (say) convenience sampling – as they often are in commercial marketing efforts, say – then they generate a nonprobability sample. On the other hand, if quota sampling is combined with a probability sampling method, it can be a potentially viable (if messy) way of sampling.

Quota sampling was used a *lot* in older times, before it was appreciated that it didn't lead to samples with nice properties. As a result, there are lots of surveys, scientific observational studies, and the like that have quota-sampled data. This glut of data has led to the development (and the continuing development) of methods for analyzing quota-sampled data. I won't go into all that here; the point is just that not all quota samples are necessarily bad, but rather than they can require more sophisticated techniques to draw valid inferences from them.

Sampling Error (or "Margin of Error")

Sampling error (sometimes called the "margin of error," and abbreviated MOE) is just the (random) difference between the thing you want to know in the population and its respective value in the sample. Of course, most of the time, we don't know the former, but that doesn't stop us from being able to get "bounds" on sampling error anyway. Formally,

$$\text{Standard error} = \sqrt{\frac{q(1-q)}{N}} \tag{1}$$

where N is (as usual) the sample size and q is the calculated quantity (proportion) of interest. More often, we calculate a relative sampling error for a particular *level of confidence*.[1]

Three things affect the margin of error:

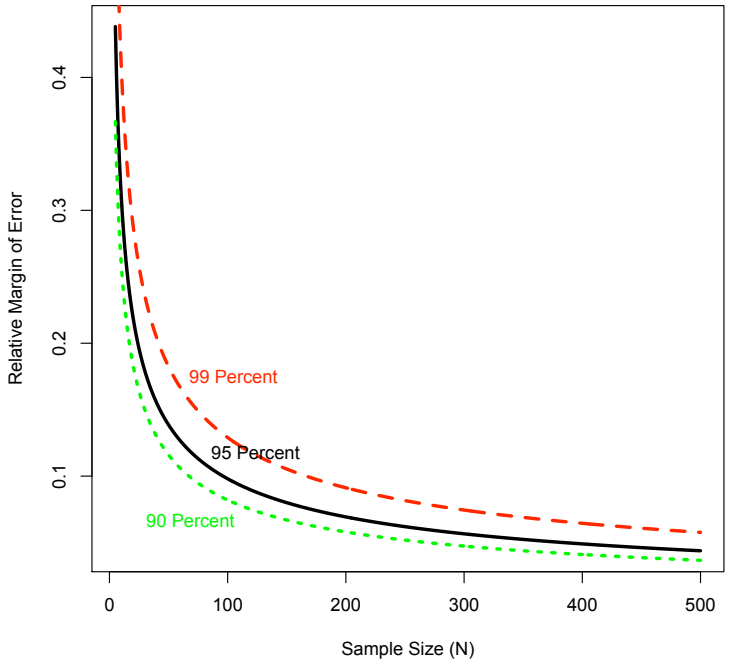
[1] Don't worry about exactly what this means right now, if you don't already know; we'll get to it.

1. The sample size,
2. The sampling design, and
3. The size of the population.

Sample Size

The margin of error is always decreasing in sample size, albeit at a decreasing rate. For a simple random sample, for example, the relationship between sample size and MOE looks like this:

Figure 3: Sampling Error vs. Sample Size



Sampling Design

Complicated sampling designs make calculation of the margin of error a bit more complicated. In general:

1. Simple random sampling is the most straightforward, and will have a relatively small amount of sampling error. However,

2. Stratified samples can actually make the margin of error *even smaller*, relative to simple random sampling, if the stratification is done in a way that "balances" the sample across the strata.

3. Cluster samples and their variants will have higher margins of error than those for which the units of analysis are the PSUs

Population Size

Finally, note that (in a marginal way) the size of the population affects the margin of error. Holding sample size constant, the margin of error is increasing in the population size, though it also does so with a high degree of diminishing returns. That is, a random sample of $N = 1000$ will do an almost-equally good job of representing a population of 50,000 as it will a population of 50,000,000. That said, when the population size gets *very* small – that is, when it approaches the sample size – then the sample necessarily does a better job of being representative. Because of that, we use *finite population corrections* in situations where the sample N is ≥ 5 percent of the population $\mathfrak{N}$; we'll talk about those a bit later. (In the limit, of course, when the sample size equals the population size, the margin of error is zero).

Randomization

Beyond random sampling, the practice of *randomization* is an incredibly powerful tool for learning about the world. In fact, at some level, nearly all of science is based on randomization, or at least ideas tied to randomization. The value of randomization is largely to prevent *confounding* effects from biasing our results. Confounding can occur in a number of ways, but appears most commonly when an extraneous (third) variable has an influence on both the "cause" and the "effect" that we are trying to study.

In the social sciences, we can randomize in a number of ways, but (forgetting about sampling for the moment) the two most common are random *treatment assignments* and randomization as a means of *orthogonalization*. The two are related, of course, with the former being a special case of the latter; in practice, at least in observational studies, the latter is much more common than the former.

Randomization of Treatments

Randomizing a treatment is the single best way to assess the causal relationship between some factor and some outcome, period. It involves (a) randomly assigning subjects to "treatment" and "control" groups, (b) administering the treatment or placebo, respectively, and (c) measuring the outcomes.

Beyond this simple design, randomized treatment designs have a host of different forms. For example, medical studies often utilize "crossover" designs, where those randomly selected to

receive the treatment initially are (at some point in the study) "crossed over" and receive the placebo, and vice-versa.

Of course, we typically don't have the luxury of randomized treatment in the social sciences (or, at least, not in those not named "psychology"). Which leads us to...

Randomization As Orthogonalization

Randomization can also be useful even if no "treatment" per se is administered. Consider U.S. Courts of Appeals. They decide cases in three-judge panels; the process by which they do so begins when (a) three judges from a circuit are randomly assigned to a panel, and then (b) cases are randomly assigned to that panel.

Notice that, here, there is no "treatment" per se, at least not in the orthodox experimental sense. At the same time, the fact of randomization allows us to get a handle on a range of interesting questions. For example, because judges are assigned to cases without regard to gender, we have a mix of all-male, all-female, and mixed-gender three-judge panels. If we are interested in whether female judges decide cases different than do male judges, the random design gives us leverage on that question, since we can be sure that a given case (say, one on Equal Protection rights for women) is no more likely to be decided by a panel of male judges than a pane of female ones.[2]

More generally, randomization of this sort makes it possible that all those potentially confounding factors in our study are no longer related to the variables we care about. This, in turn, simplifies our life considerably, and makes the inferences we can draw much stronger.

[2]Of course, this assumes that men and women are equally represented on the courts; but even if they are not, we *do* know how uneven that representation is, and so can correct for any differences after the fact.