

## 5.1: Populations and Samples

The basis of hypothesis testing with statistical analysis is **inference**. In short, inference—and inferential statistics by extension—means deriving knowledge about a population from a sample of that population. Given that in most contexts it is not possible to have all the data on an entire population of interest, we, therefore, need to sample from that population.<sup>8</sup> However, in order to be able to rely on inference, the sample must cover the theoretically relevant variables, variable ranges, and contexts.

### 5.1.1 Populations and Samples

In doing the statistical analysis we differentiate between populations and samples. The population is the total set of items that we care about. The sample is a subset of those items that we study in order to understand the population. While we are interested in the population we often need to resort to studying a sample due to time, financial, or logistic constraints that might make studying the entire population infeasible. Instead, we use inferential statistics to make inferences about the population from a sample.

### 5.1.2 Sampling and Knowing

Take a relatively common – but perhaps less commonly examined – expression about what we “know” about the world around us. We commonly say we know “people, and some we know better than others. What does it mean to know someone? In part, it must mean that we can anticipate how that person would behave in a wide array of situations. If we know that person from experience, then it must be that we have observed their behavior across a sufficient variety of situations in the past to be able to infer how they would behave in future situations. Put differently, we have “sampled” their behavior across a relevant range of situations and contexts to be confident that we can anticipate their behavior in the future.<sup>9</sup> Similar considerations about sampling might apply to “knowing” a place, a group, or an institution. Of equal importance, samples of observations across different combinations of variables are necessary to identify relationships (or functions) between variables. In short, samples – whether deliberately drawn and systematic or otherwise – are integral to what we think we know of the world around us.

### 5.1.3 Sampling Strategies

Given the importance of sampling, it should come as little surprise that there are numerous strategies designed to provide useful inference about populations. For example, how can we judge whether the temperature of soup is appropriate before serving it? We might stir the pot, to assure uniformity of temperature across possible (spoon-sized) samples, then sample a spoonful. A particularly thorny problem in sampling concerns the practice of courtship, in which participants may attempt to put “their best foot forward” to make a good impression. Put differently, the participants often seek to bias the sample of relational experiences to make themselves look better than they might on average. Sampling in this context usually involves (a) getting opinions of others, thereby broadening (if only indirectly) the size of the sample, and (b) observing the courtship partner over a wide range of circumstances in which the intended bias may be difficult to maintain. Put formally, we may try to stratify the sample by taking observations in inappropriate “cells” that correspond to different potential influences on behavior – say high-stress environments involving preparation for final exams or meeting parents. In the best possible case, however, we try to wash out the effect of various influences on our samples through randomization. To pursue the courtship example (perhaps a bit too far!), observations of behavior could be taken across interactions from a randomly assigned array of partners and situations. But, of course, by then all bets are off on things working out anyway.

### 5.1.4 Sampling Techniques

When engaging in inferential statistics to infer the characteristics of a population from a sample, it is essential to be clear about how the sample was drawn. Sampling can be a very complex practice with multiple stages involved in drawing the final sample. It is desirable that the sample is some form of a **probability sample**, i.e., a sample in which each member of the population has a known probability of being sampled. The most direct form of an appropriate probability sample is a **random sample** where everyone has the same probability of being sampled. A random sample has the advantages of simplicity (in theory) and ease of inference as no adjustments to the data are needed. But, the reality of conducting a random sample may make the process quite challenging. Before we can draw subjects at random, we need a list of all members of the population. For many populations (e.g. adult US residents) that list is impossible to get. Not too long ago, it was reasonable to conclude that a list of telephone numbers was a reasonable approximation of such a listing for American households. During the era that landlines were ubiquitous, pollsters could randomly call numbers (and perhaps ask for the adult in the household who had the most recent birthday) to get a good approximation of a national random sample. (It was also an era before caller identification and specialized ringtones, which meant that calls were routinely answered, therefore decreasing – but not eliminating – concern with response bias.) Of course, telephone

habits have changed and pollsters find it increasingly difficult to make the case that random dialing of landlines serves as a representative sample of adult Americans.

Other forms of probability sampling are frequently used to overcome some of the difficulties that pure random sampling presents. Suppose our analysis will call upon us to make comparisons based on race. Only 12.6% of Americans are African-American. Suppose we also want to take into account religious preference. Only 5% of African-Americans are Catholic, which means that only .6% of the population is both. If our sample size is 500, we might end up with three Catholic African-Americans. A **stratified random sample** (also called a quota sample) can address that problem. A stratified random sample is similar to a simple random sample but will draw from different subpopulations, strata, at different rates. The total sample needs to be weighted, then, to be representative of the entire population.

Another type of probability sample that is common in face-to-face surveys relies on **cluster sampling**. Cluster sampling initially samples based on clusters (generally geographic units, such as census tracts) and then samples participants within those units. In fact, this approach often uses multi-level sampling where the first level might be a sample of congressional districts, then census tracts, and then households. The final sample will need to be weighted in a complex way to reflect varying probabilities that individuals will be included in the sample.

**Non-probability samples**, or those for which the probability of inclusion of a member of the population in the sample is unknown, can raise difficult issues for statistical inference; however, under some conditions, they can be considered representative and used for inferential statistics.

**Convenience samples** (e.g., undergraduate students in the Psychology Department subject pool) are accessible and relatively low cost but may differ from the larger population to which you want to infer in important respects. Necessity may push a researcher to use a convenience sample, but inference should be approached with caution. A convenience sample based on “I asked people who came out of the bank” might provide quite different results from a sample based on “I asked people who came out of a payday loan establishment”.

Some non-probability samples are used because the researcher does not want to make inferences to a larger population. A **purposive or judgmental sample** relies on the researcher’s discretion regarding who can bring useful information to bear on the subject matter. If we want to know why a piece of legislation was enacted, it makes sense to sample the author and co-authors of the bill, committee members, leadership, etc., rather than a random sample of members of the legislative body.

**Snowball sampling** is similar to a purposive sample in that we look for people with certain characteristics but rely on subjects to recommend others who meet the criteria we have in place. We might want to know about struggling young artists. They may be hard to find, though, since their works are not hanging in galleries so we may start with one or more that we can find and then ask them who else we should interview.

Increasingly, various kinds of non-probability samples are employed in social science research, and when this is done it is critical that the potential biases associated with the samples be evaluated. But there is also growing evidence that non-probability samples can be used inferentially - when done very carefully, using complex adjustments. Wang, et al. (2014) demonstrate that a sample of Xbox users could be used to forecast the 2012 presidential election outcome.<sup>10</sup> An overview of their technique is relatively simple, but the execution is more challenging. They divided their data into cells based on politically and demographically relevant variables (e.g., party id, gender, race, etc.) and ended up with over 175,000 cells - post stratification. (There were about three-quarters of a million participants in the Xbox survey). Basically, they found the vote intention within each cell and then weighted each cell based on a national survey using multilevel regression. Their final results were strikingly accurate. Similarly, Nate Silver, with FiveThirtyEight, has demonstrated remarkable ability to forecast based on his weighted sample of polls taken by others.

Sampling techniques can be relatively straightforward, but as one moves away from simple random sampling, the sampling process either becomes more complex or limits our ability to draw inferences about a population. Researchers use all of these techniques for good purposes and the best technique will depend on a variety of factors, such as budget, expertise, need for precision, and what research question is being addressed. For the remainder of this text, though, when we talk about drawing inferences, the data will be based upon an appropriately drawn probability sample.

### 5.1.5 So How is it That We Know?

So why is it that the characteristics of samples can tell us a lot about the characteristics of populations? If samples are properly drawn, the observations taken will provide a range of values on the measures of interest that reflect those of the larger population. The connection is that we expect the phenomenon we are measuring will have **distribution** within the population, and a sample of

observations drawn from the population will provide useful information about that distribution. The theoretical connection comes from probability theory, which concerns the analysis of random phenomena. For present purposes, if we randomly draw a sample of observations on a measure for an individual (say, discrete acts of kindness), we can use probability theory to make inferences about the characteristics of the overall population of the phenomenon in question. More specifically, probability theory allows us to make inference about the shape of that distribution – how frequent are acts of kindness committed, or what proportion of acts evidence kindness?

In sum, samples provide information about **probability distributions**. Probability distributions include all possible values and the probabilities associated with those values. The **normal distribution** is the key probability distribution in inferential statistics.

---

This page titled [5.1: Populations and Samples](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Jenkins-Smith et al. \(University of Oklahoma Libraries\)](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.