2.3: Representative Sample

Usually, the ideal method of picking out a sample to study is called random selection or sampling. The researcher starts with a **complete list of the population** and randomly selects some of them to study. . Random sampling is considered a fair way of selecting a sample from a given population since every member is given equal opportunities of being selected. This process also helps to ensure that the sample selected is more likely to be representative of the larger population. Theoretically, the only thing that can compromise its representativeness is luck. If the sample is not representative of the population, the random variation is called **sampling error.**

> **Example #1:** You have been hired by the National Election Commission to examine how the American people feel about the fairness of the voting procedures in the U.S. Who will you ask?

It is not practical to ask every single American how he or she feels about the fairness of the voting procedures. Instead, we query a relatively small number of Americans, and draw inferences about the entire country from their responses. The Americans actually queried constitute our sample of the larger population of all Americans. A **sample** is typically a small subset of the population. In the case of voting attitudes, we would sample a few thousand Americans drawn from the hundreds of millions that make up the country. In choosing a sample, it is therefore crucial that it not over-represent one kind of citizen at the expense of others. For example, something would be wrong with our sample if it happened to be made up entirely of Florida residents. If the sample held only Floridians, it could not be used to infer the attitudes of other Americans. The same problem would arise if the sample were comprised only of Republicans. Inferences from statistics are based on the assumption that sampling is representative of the population. If the sample is not representative, then the possibility of sampling bias occurs. Sampling bias means that our conclusions apply only to our sample and are not generalizable to the full population.

> **Example #2:** We are interested in examining how many math classes have been taken on average by current graduating seniors at American colleges and universities during their four years in school.

Whereas our population in the last example included all US citizens, now it involves just the graduating seniors throughout the country. This is still a large set since there are thousands of colleges and universities, each enrolling many students. (New York University, for example, enrolls 48,000 students.) It would be prohibitively costly to examine the transcript of every college senior. We therefore take a sample of college seniors and then make inferences to the entire population based on what we find. To make the sample, we might first choose some public and private colleges and universities across the United States. Then we might sample 50 students from each of these institutions. Suppose that the average number of math classes taken by the people in our sample were 3.2. Then we might speculate that 3.2 approximates the number we would find if we had the resources to examine every senior in the entire population. But we must be careful about the possibility that our sample is non-representative of the population. Perhaps we chose an overabundance of math majors, or chose too many technical institutions that have heavy math requirements. Such bad sampling makes our sample unrepresentative of the population of all seniors. To solidify your understanding of sampling bias, consider the following example. Try to identify the population and the sample, and then reflect on whether the sample is likely to yield the information desired.

> **Example #3:** A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well. What is the sample? What is the population? Can you identify any problems with choosing the sample in the way that the teacher did?

In Example #3, the population consists of all students in the class. The sample is made up of just the 10 students sitting in the front row. The sample is not likely to be representative of the population. Those who sit in the front row tend to be more interested in the class and tend to perform higher on tests. Hence, the sample may perform at a higher level than the population.

> **Example #4:** A coach is interested in how many cartwheels the average college freshmen at his university can do. Eight volunteers from the freshman class step forward. After observing their performance, the coach concludes that college freshmen can do an average of 16 cartwheels in a row without stopping.

In Example #4, the population is the class of all freshmen at the coach's university. The sample is composed of the 8 volunteers. The sample is poorly chosen because volunteers are more likely to be able to do cartwheels than the average freshman; people who can't do cartwheels probably did not volunteer! In the example, we are also not told of the gender of the volunteers. Were they all women, for example? That might affect the outcome, contributing to the non-representative nature of the sample (if the school is co-ed).

Simple Random Sampling

Researchers adopt a variety of sampling strategies. The most straightforward is simple random sampling. Such sampling requires every member of the population to have an equal chance of being selected into the sample. In addition, the selection of one member must be independent of the selection of every other member. That is, picking one member from the population must not increase or decrease the probability of picking any other member (relative to the others). In this sense, we can say that simple random sampling chooses a sample by pure chance. To check your understanding of simple random sampling, consider the following example.

What is the population? What is the sample? Was the sample picked by simple random sampling? Is it biased?

> **Example #5:** A research scientist is interested in studying the experiences of twins raised together versus those raised apart. She obtains a list of twins from the National Twin Registry, and selects two subsets of individuals for her study. First, she chooses all those in the registry whose last name begins with Z. Then she turns to all those whose last name begins with B. Because there are so many names that start with B, however, our researcher decides to incorporate only every other name into her sample. Finally, she mails out a survey and compares characteristics of twins raised apart versus together.

In Example #5, the population consists of all twins recorded in the National Twin Registry. It is important that the researcher only make statistical generalizations to the twins on this list, not to all twins in the nation or world. That is, the National Twin Registry may not be representative of all twins. Even if inferences are limited to the Registry, a number of problems affect the sampling procedure we described. For instance, choosing only twins whose last names begin with Z does not give every individual an equal chance of being selected into the sample. Moreover, such a procedure risks over-representing ethnic groups with many surnames that begin with Z. There are other reasons why choosing just the Z's may bias the sample.

Perhaps such people are more patient than average because they often find themselves at the end of the line! The same problem occurs with choosing twins whose last name begins with B. An additional problem for the B's is that the "every-other-one" procedure disallowed adjacent names on the B part of the list from being both selected. Just this defect alone means the sample was not formed through simple random sampling.

Sample size matters

Recall that the definition of a random sample is a sample in which every member of the population has an equal chance of being selected. This means that the sampling procedure rather than the results of the procedure define what it means for a sample to be random. Random samples, especially if the sample size is small, are not necessarily representative of the entire population. For example, if a random sample of 20 subjects were taken from a population with an equal number of males and females, there would be a nontrivial probability (0.06) that 70% or more of the sample would be female. Such a sample would not be representative, although it would be drawn randomly. Only a large sample size makes it likely that our sample is close to representative of the population. For this reason, inferential statistics take into account the sample size when generalizing results from samples to populations. In later chapters, you'll see what kinds of mathematical techniques ensure this sensitivity to sample size.

More complex sampling

Sometimes it is not feasible to build a sample using simple random sampling. To see the problem, consider the fact that both Dallas and Houston are competed to be hosts of the 2012 Olympics. Imagine that you are hired to assess whether most Texans prefer Houston to Dallas as the host, or the reverse. Given the impracticality of obtaining the opinion of every single Texan, you must construct a sample of the Texas population. But now notice how difficult it would be to proceed by simple random sampling. For example, how will you contact those individuals who don't vote and don't have a phone? Even among people you find in the telephone book, how can you identify those who have just relocated to California (and had no reason to inform you of their move)? What do you do about the fact that since the beginning of the study, an additional 4,212 people took up residence in the state of

Texas? As you can see, it is sometimes very difficult to develop a truly random procedure. For this reason, other kinds of sampling techniques have been devised. We now discuss two of them.

Stratified Sampling

Since simple random sampling often does not ensure a representative sample, a sampling method called **stratified random sampling** is sometimes used to make the sample more representative of the population. This method can be used if the population has a number of distinct "strata" or groups. In stratified sampling, you first identify members of your sample who belong to each group. Then you randomly sample from each of those subgroups in such a way that the sizes of the subgroups in the sample are proportional to their sizes in the population.

Let's take an example: Suppose you were interested in views of capital punishment at an urban university. You have the time and resources to interview 200 students. The student body is diverse with respect to age; many older people work during the day and enroll in night courses (average age is 39), while younger students generally enroll in day classes (average age of 19). It is possible that night students have different views about capital punishment than day students. If 70% of the students were day students, it makes sense to ensure that 70% of the sample consisted of day students. Thus, your sample of 200 students would consist of 140 day students and 60 night students. The proportion of day students in the sample and in the population (the entire university) would be the same. Inferences to the entire population of students at the university would therefore be more secure.

Convenience Sampling

Unfortunately, it is often impractical or impossible to study a truly random sample. Much of the time, in fact, studies are conducted with whoever is willing or available to be a research participant – this is commonly referred to as **convenience sampling**. At best, as noted, a researcher tries to study a sample that is not systematically unrepresentative of the population in any known way. For example, suppose a study is about a process that is likely to differ for people of different age groups. In this situation, the researcher may attempt to include people of all age groups in the study. Alternatively, the researcher would be careful to draw conclusions only about the age group studied. Remember that one of the goals of research is to make conclusions about the population from the sample results. An unbiased random sample and a representative sample are important when drawing conclusions from the results of a study.

"WEIRD" Culture Samples

Psychologists have been guilty of largely recruiting samples of convenience from the thin slice of humanity—students—found at universities and colleges (Sears, 1986). This presents a problem when trying to assess the social mechanics of the public at large. Aside from being an overrepresentation of young, middle-class Caucasians, college students may also be more compliant and more susceptible to attitude change, have less stable personality traits and interpersonal relationships, and possess stronger cognitive skills than samples reflecting a wide range of age and experience (Peterson & Merunka, 2014; Visser, Krosnick, & Lavrakas, 2000). Put simply, these traditional samples (college students) may not be sufficiently representative of the broader population. Furthermore, considering that 96% of participants in psychology studies come from *western, educated, industrialized, rich, and democratic countries* (so-called **WEIRD cultures**; Henrich, Heine, & Norenzayan, 2010), and that the majority of these *are also psychology students*, the question of non-representativeness becomes even more serious. Of course, when studying a basic cognitive process (like working memory capacity) or an aspect of social behavior that appears to be fairly universal (e.g., even cockroaches exhibit social facilitation), a non-representative sample may not be a big deal. Over time research has repeatedly demonstrated the important role that individual differences (e.g., personality traits, cognitive abilities, etc.) and culture (e.g., individualism versus collectivism) play in shaping social behavior. For instance, even if we only consider a tiny sample of research on aggression, we know that narcissists are more likely to respond to criticism with aggression (Bushman &Baumeister, 1998); conservatives, who have a low tolerance for uncertainty, are more likely to prefer aggressive actions against those considered to be 'outsiders'(de Zavala et al., 2010); countries, where men hold the bulk of the power in society, have higher rates of physical aggression directed against female partners (Archer, 2006); and males from the southern part of the United States are more likely to react with aggression following an insult (Cohen et al., 1996).

Why does random sampling work?

Below is an example showing how many credit hours students are currently enrolled in at a community college. This data represents the entire population of interest, all students currently enrolled in classes at Chandler-Gilbert Community College. Let's say we randomly selected one student out of the population and asked them how many credit hours they are currently taking. How likely would it be for this one student to represent the entire population? This is the first line showing 1 student reported taking 12 hours while the average credit hours for a CGCC student was 8 (population average).

Sample size (n)	Sample average	Population average	Difference between Sample & Population
1	12	8	4
2	15	8	7
5	9.8	8	1.8
25	9.5	8	1.5
250	8.3	8	0.3
2500	7.9	8	0.1

The larger the sample size, the more closely it represents the population.

As we can see from this activity, the larger our sample is, the more accurately it will represent the population from which it was drawn. This brings up a very important rule in research design. The larger the sample size is, the more accurately the sample will represent the population from which it was drawn. Also, if you are comparing groups, consider that the more diverse, or variable, individuals in each group are, the larger the sample needs to be to detect real differences between groups. We will further dive into the importance of sample sizes with inferential statistics, but for now, consider that the larger the sample, the more likely the researcher will represent the population.