Section 1B – Collecting Data

Vocabulary

Population: The collection of all people or objects you want to study.

Census: Collecting data from everyone in the population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not reflect the population.

Random: When everyone in the population has a chance to be included in the sample.

One of the most important goals in data science is to learn about the world around us (populations). It is very difficult to understand populations sometimes because data may be biased and not reflect the population very well. Bias can occur in many different ways, but certain ways people collect data have more bias than others do. Using a method for collecting data that increases bias is sometimes called "sampling bias".

It is important to be aware of various methods used to collect data, the good and the bad.

Method 1: Census

A census is the best way to collect data if it is possible. If our goal is to learn about the population, it makes sense to collect data from everyone in the population. There are ways for a census to be biased, but in terms of the collecting method, a census is the best. Unfortunately, it is almost impossible to collect a census if your population is large. Most statisticians and data scientists are only able to collect a sample, data collected from a small subgroup of the population.

Method 2: Simple Random Sample

If a statistician or data scientist cannot collect a census, the preferred method is to collect a random sample. A random sample is one where everyone in the population has a chance to be in the sample, so it tends to represent the population better than other non-random samples. It is nowhere near as good as a census, but as I said, a census is usually not possible.

A simple random sample is one where individuals in the population are selected randomly. This can be a difficult process. The usual method is to assign everyone in a population a number and then use a random number generator in a computer program to pick random numbers. Computer programs have many built in randomization functions for this purpose. If you have a spreadsheet of the entire population, a computer can also randomly select individuals from the list. The key with a "simple random sample" is that you are selecting people or objects one at a time. Collecting data randomly and one at a time gives greater flexibility to your sample. Almost any grouping is possible with a simple random sample, so it tends to represent populations better than other samples.

There are many examples of a simple random sample. Many statistics companies use a random phone number generator that randomly gives phone numbers. They then call the phone numbers randomly chosen and try to get information from people that answer the phone. The U.S. government may have a computer randomly select social security numbers to select individuals for a sample. A company may have a computer randomly select employee ID numbers to select individuals for a sample.

Method 3: Convenience Sample

People often find collecting a census or a simple random sample difficult, so they chose to collect data in whatever way seems easiest. A sample collected this way is often called a "convenience sample" and is popular with people not trained in statistics. A convenience sample usually has much more bias than a random sample and may not represent the population very well.

An example of a convenience sample is collecting data from your friends and family. This is fine if your population of interest is your friends and family, but will by no means represent a large population. Another example might be standing outside of a store or post office and collecting data from people that leave the store. Beginning statistics students may walk into a mall and collect data from whomever they bump into. They mistakenly think that these are random samples, but they are not. A random sample means everyone in the population has a chance to be included in the sample. Not everyone in the population has a chance to bump into you at a mall or come out of a store at 2:30 pm on a Tuesday afternoon. These are convenience samples and generally do not reflect the population very well.

Method 4: Voluntary Response Sample

Some say that all surveys are bad, but that is not the case. A survey is just a form to collect data from people. When a company takes a census of all its employees, it may require all of the employees to fill out a survey. That is a census. As long as no other forms of bias creep into the data, a census will probably be a very good representation of the population. The point is that giving a survey is not the issue. The issue is whom you give the survey to and who is allowed to fill out the survey.

A voluntary response sample puts a survey out into the world and allow anyone to respond. The usual method used today is to put a survey on a website and allow anyone that comes across the survey to answer. The survey can also be a mailed to every address in a given population. Again, those that fill it out self-select themselves to be in our data.

On the surface, a voluntary response sample may seem like a good way of collecting data. It usually gives a large amount of data. Does this really allow everyone in the population a chance to answer? It turns out the answer is no. Ask yourself the following question. When you are surfing the web and a survey pops up, do you fill it out? I have been asking my statistics classes that question for years and rarely have anyone that says that they do fill out surveys. The key problem is that only certain types of people will fill out a survey voluntarily. It may be a person who is bored and has nothing better to do. It is certainly not a person with three children, working a full time job and going to college full time. It may also be a person who is upset by or feels very passionate about the topic in the voluntary response survey. They are so upset by the lack of pay for teachers that they are willing to fill out a survey to tell you what they think. The point is that voluntary response surveys tend to over-sample people that are bored or upset and under-sample everyone else. For this reason, voluntary response samples can be very biased and may not represent the population very well.

I have had many students ask me if sample size is important. Isn't a voluntary response sample of five thousand people better than a random sample of fifty people?" I would tell them that though sample size is important, method is important also. The voluntary response sample of five thousand would tend to over-represent people that are bored or upset about the topic. It does not represent typical people in the population. The random sample of fifty people, while a small sample size, at least does not have that bias.

Method 5: Cluster Sample

A cluster sample is one where data is collected from groups of people in a population instead of one at a time. For example, a company that has 250 stores worldwide might have a computer randomly select ten stores and get data from those people that work at those ten stores. Notice this would be a random sample since every employee has a chance to be in the data. If their store was chosen, then they will be included in the sample. This is not a simple random sample however, since they are not choosing one at time. This example is sometimes called a "random cluster sample". While it is a good method for collecting data, it has less flexibility than a simple random sample. Think of it this way. In a simple random sample, any grouping is possible, but in this random cluster example, only groups of people that work at the same store can be chosen. It is still a random sample though, and would tend to be more representative of the population than non-random samples like convenience or voluntary response.

It is good to note that the goal of a cluster sample should be to choose the groups of people randomly. If we choose groups of people that are convenient to collect data from, our cluster sample will have more sampling bias and will not represent the population nearly as well.

Method 6: Stratified Sample

One of the most common studies done in statistics is to compare groups. We may compare data from 2016 to data from this year. We may compare people living in Canada to people living in Australia. To compare groups, you need to collect a stratified sample.

Some people in statistics explain a stratified sample as comparing two or more groups in one population. I like to think of it as comparing two or more populations. Whether you explain a stratified sample as comparing groups in one population or comparing populations, the key is that you are comparing.

For example, we may want to compare the percentage of adults in the U.S. with diabetes to the percentage of children in the U.S. with diabetes. Some statistics authors think of this as comparing adults and children in the one population of all people in the U.S. I like to think of it as comparing the population of U.S. adults to the population of U.S. children.

Another example may be to compare the mean average salary of people working in London, England to the mean average salary of people working in Toronto, Canada. Again, a stratified sample is needed because we are comparing.

To do a stratified sample, we often take a simple random sample from each group. I like to think of it as taking a simple random sample from each population you want to compare. In the previous example, we may collect a simple random sample of adults in the U.S. and another simple random sample of children in the U.S. We then can calculate the sample percentages that are diabetic from each sample and use statistical methods to compare them. For the salary example, we can collect a simple random sample of salaries for people working in London, and another simple random sample for people working in Toronto. The goal is then to use statistical methods to compare the mean average salaries.

It should be noted that when taking a stratified sample, we should use randomization. Again, if we just take a convenience sample from each group or voluntary response sample from each group, we will likely have a lot more bias and the data will not reflect the population (or populations) as well as we would like.

Many people confuse a cluster sample with a stratified sample because they both involve groups. The goal of a cluster is to get data on and analyze one population, not to compare. You are just collecting data from groups of people from that one population instead of one at a time. The goal of a stratified sample is to compare two or more populations so we need to collect data from each population.

Method 7: Systematic Sample

A systematic sample is one where we use a system to collect the sample. Usually it involves collecting data from every fifth person that comes in your store or every twentieth person on a list.

For example, let us suppose we want to collect a sample of students from our college. We could look at an alphabetical list of the names of all students that attend our college and then chose every 50^{th} person on the list. Is this a random data set? Ask yourself this question. Does everyone on this list have a chance to be chosen? No. Only the 50^{th}, 100^{th}, 150^{th}, 200^{th} and so forth have a chance. People from 1-49 have no chance. People from 51-99 have no chance. Therefore, it is not a random sample. This may not be random, but we may make the argument that it is representative of the population. This method would have less bias than convenience or voluntary response samples. There is a way to incorporate randomization into the method. Many data scientists have a computer chose a random number between 1 and 50. Suppose it is 33. Then they collect data from the 33^{rd} person on the alphabetical list. Now, from there use the system of choosing every 50^{th} person. Therefore, they would choose the 33^{rd} person, then the 83^{rd} person, then the 133^{rd} person and so on. Making the first choice random, makes the whole data set random, because everyone on the list now has a chance to be chosen.

Summary

So let us summarize the various methods.

- An unbiased census is the best way to collect data to represent a population, because we are collecting data from everyone in the population.
- If you cannot do a census, then use a random sample of some sort. It may be a simple random sample, random cluster, or a random systematic sample. The main thing is that if you are collecting a sample, randomization needs to be involved.
- Voluntary response samples and convenience samples tend to be very biased and should be avoided if possible.
