[RESOURCE TITLE] Understand Statistical Studies

Objective 4.1: Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

[START FLASHCARDS]

Population: A large group which you are interested in studying.

Sample: The individuals chosen from a population which you actually get data from.

Census: A sample that includes the entire population.

Sampling design: The method by which the sample is chosen.

Sampling frame: The "list" of individuals the sample is taken from.

Simple Random Sample: The equivalent of drawing names from a hat, individuals are chosen at random from a population.

[END FLASHCARDS]

[MAJOR HEAD]Populations and Samples

In previous sections, you have learned about how to interpret data that you get from samples. You've learned how to visualize data for a single categorical variable on bar charts and pie charts, and how to visualize data for a single quantitative variable on a histogram or dot plot.

You've learned how to visualize data for two associated categorical variables on a two-way table, and how to visualize data for two associated quantitative variables on a scatterplot. In addition, much attention has been given to fitting those scatterplots to a linear model, and there is even a particular value, called the Pearson correlation coefficient, denoted *r*, which quantifies just how closely related those two variables appear to be.

In all of these cases, you have been working with sample data. What does that mean, exactly? To learn that, you'll add two new terms to your vocabulary: *population* and *sample.*

Simply put, a population is a group of people you are interested in studying, and a sample is a group of individuals (people or objects) that you actually get data from.

Let's imagine you want to know the distribution of study habits for statistics students in your school district. There are a total of 10 statistics classes offered at the various schools in your district. You choose, at random, 4 students from each class and have them complete a study habits survey. The *population* in this case is *all* statistics students in your school district. The *sample* is the 40 students who answer your survey. The hope is that your sample are a good representation of the rest of the population.

Statistics is, as you are discovering, not really an "exact" science. Statistics almost always comes with some amount of uncertainty, and sampling is no different. There are all different kinds of statistics students. Some students study for 2 hours every day. Some students never study at all. It is entirely possible that your sample of students includes a little of each kind of student. It is also possible that your sample of students only includes the most diligent of students.

To put this another way, imagine that you pick, at random, 10 marbles from a bag of multi-colored marbles, and you happen to choose 10 red marbles. This is possible (if unlikely), and most importantly, it does not automatically invalidate your sample! When discussing sampling, it is more important to consider *how* you choose your sample than *what* ends up in the sample you choose. So, when we discuss populations and samples, it's important to talk about sampling methods.

[MAJOR HEAD]Sampling Methods

Sampling methods are, as the name implies, the different methods by which you could get a sample from a population. It could be as simple as putting up an online survey, as arbitary as choosing every 10th person in a room, or as convenient as going to the mall and asking the first 10 people you see a question.

[MINOR HEAD]Census

The best possible sample is the one that includes the entire population. After all, the population is the group we're interested in studying, so surveying, measuring, or otherwise getting data from every possible member of that population will guarantee accurate results. A sample that comprises an entire population is called a *census*. The United States conducts a census of its population once every ten years.

However, in most situations, a census is not going to be a viable sampling method. For most realistic cases, the best sampling method we can hope for is Simple Random Sampling.

[MINOR HEAD]Simple Random Sampling

The basic method by which most samples in this course are assumed to be taken, simple random sampling is the best method by which a sample can be taken from a population. You can think of simple random sampling as pulling names out of a hat, because it is essentially the same process.

In most studies, a Simple Random Sample (also called an SRS for short) is generated by random numbers. First, all individuals in a population are assigned an ID number. Then, using a random number generator (a random digit table, or software), individuals are chosen. This produces the same effect as pulling names from a hat.



INSERT FIGURE 4.1\_1. ALT TAG: A table of random digits. ]

[CAPTION] Random digit table.

In the above figure, you can see what a random number table, also called a random digit table, looks like. You'll use one of these to create your own SRS from a small population in a later section. For now, you can just remember that a sample that is chosen by a truly random method, whether it is by numbers or by randomly drawing a name, is a simple random sample.

[CALLOUT]

In a simple random sample, every possible set of *n* individuals has an equal chance of being the chosen sample.

[END CALLOUT]

[MINOR HEAD]Stratified Random Sampling

In stratified random sampling, a population is separated into several groups or categories, and a simple random sample is taken from each group. For example, consider a hospital. Hospital staff includes medical staff like Doctors and Nurses, office staff like Receptionists and Accountants, and general staff like Janitorial and Food Service. If you wanted to conduct a survey of all hospital students, you could divide the hospital's staff into three groups: Group A for medical staff, Group B for office workers, and Group C for everyone else.



INSERT FIGURE 4.1\_1. ALT TAG: Stratified Random Sample. ]

[CAPTION] Stratified Random Sample.

[MINOR HEAD]Multi-Stage Random Sampling

In a multi-stage sample, a sample is taken first from a population, and then a further simple random sample is taken from that.



INSERT FIGURE 4.1\_1. ALT TAG: Multi-Stage Random Sample. ]

[CAPTION] Multi-Stage Random Sample.

[MAJOR HEAD]Bias and "Bad" Sampling Methods

As mentioned, Simple Random Sampling is the best method of making samples that are available to those who work in statistics. However, many statistics you see on the web, in the news, and in other media don't come from simple random samples at all. How many times have you seen some statistical results based on a web poll, on a phone survey, or based on "man on the street" type polling? These are conducted all the time, and often different publications report wildly different numbers for the same questions - why is that? The answer is *bias*.

Simple random sampling is so successful because it avoids bias. A study is biased when it systematically favors one outcome over another. Here's an example: a local talk radio station has a crew of mostly conservative hosts, and their audience are typically conservative voters. If that radio hosts a call-in poll, it shouldn't be surprising that the results lean toward the conservative side of the political spectrum. If a phone survey only calls individuals who have a landline and are listed in the telephone directory, this is not going to be a good representation of the population - many people do not have a land line at all today. Like these examples, many sampling methods used in polls and studies you see in the media are "bad" sampling methods - that is, they are potentially prone to bias.

[MINOR HEAD]Convenience Sampling

Convenience sampling is a sampling method in which the individuals making up the sample are merely the first individuals you run across. Consider this example: a journalist wants to take a poll among fans of a local sports team, so he stands outside the stadium's east entrance and surveys the first 30 people who exit.

Can you think of some potential sources of bias? First, the people may either all be fans of the home team, or may all be fans of the opposition, depending upon which side of the stadium is the "Home" side. Second, since the survey happens right after the game, the outcome of the game could potentially influence the responses - if your favorite team just lost the big game, wouldn't that influence the way you responded to a survey on local sports?

Other examples of convenience sampling include taking a poll of the first 10 callers into a radio station, and going to the beach and picking up the first 25 seashells you see. The recurring theme is that a convenience sample is made up of those individuals whom you encounter *first*, not necessarily at random.

[MINOR HEAD]Voluntary Response Sampling

Another common sampling method which is potentially biased is voluntary response. One prevalent example of this is the web poll. Many websites put up reader polls where visitors can voice their opinions on a wide variety of matters. However, in this case, it is up to the reader to voluntarily take the survey. This is potentially biased because those who are most likely to respond to a voluntary response survey or study can have extreme opinions one way or the other, while the more moderate opinions may opt not to participate.

Think of this example: a website wants to know how its readers like their steaks cooked. There are many people who strongly believe that steaks should be cooked rare and never more than rare. Conversely, there are many people who strongly believe that steaks should be cooked well done. Finally, there are also many people who aren't terribly picky about how their steaks are cooked. As a result, the reader poll seems diametrically opposed between rare and well done, with other options having fewer responses. The importance attached to an issue among individuals is a source of bias in voluntary response sampling.

[MINOR HEAD]Other sources of bias

There are many potential sources of bias in a study or survey. Here are a few of the most common ones not already mentioned:

**Non-response** - if a large part of a population simply does not respond to a poll or survey, this is an example of non-response.

**Undercoverage** - if a survey is only seen by a small segment of the population (which is not representative of the population as a whole), this is most definitely a potential source of bias. Our previous example of the conservative talk radio station would be an example of undercoverage.

**Leading questions** - if the questions in a survey are asked in a way that influences the person answering, this is a potential source of bias.

**Dishonesty in responses** - Surveys and studies that ask questions that may display the responder in a negative light, such as asking if they have ever committed a crime, can introduce bias because no one wants to portray themselves as a bad person.

[RESOURCE TITLE] Key Points and Links

[MAJOR HEAD] Key Points

[BEGIN BULLETED LIST]

In statistics, we are interested in learning things about populations - all citizens of a country, city, or town, for example.

However, since it is seldom possible to conduct a census (in which the entire population is studied), we use sampling.

When we take a sample, we select certain individuals from a population and collect data from them.

Sampling is prone to bias, in the case of polling a limited audience,or in voluntary response and convenience sampling.

The best kind of sampling we can use is simple random sampling, because it tends to not be prone to bias.

In a simple random sample (SRS), individuals may be assigned ID numbers, and then chosen via a random number generator.

In a stratified random sample, a population is split into groups, and an SRS is taken from each.

In a multi-stage random sample, a sample is taken from a population, and then a further SRS is taken from that sample.

[END BULLETED LIST]

[END KEY POINTS AND LINKS]

[START DISCOVER MORE]

A marketing company is interested in which grade level of US high school students (Freshmen, Sophomores, Juniors, Seniors) watch the most television. A sample of 25 students is taken from your high school. Answer the following questions:

1. What is the population?

2. What is the sample?

3. What is the sampling frame?

4. Suppose the marketing company first divides the student body at your school into four groups (one for each grade level), and then takes a random sample from each. What is the sampling method used here?

5. Suppose the marketing company simply visits your high school and interviews the first 25 students who walk out of a classroom. What is the sampling method used here?

[START DISCOVER MORE ANSWER]

1. All US High School Students.

2. The 25 students polled.

3. The list of all students at your high school.

4. Stratified random sample.

5. Convenience sample.

[END DISCOVER MORE ANSWER]

[END DISCOVER MORE]

[START SELF CHECK]

1. A study or survey which systematically (whether intentional or not) favors certain outcomes is called what?

A. Biased

B. Lurking

C. Random

D. Multistage.

Explanation: Bias is the systematic favoring of one outcome or another.

2. A local television station wishes to know whether local voters are opposed to, or in favor of, a state sales tax increase. The station asks viewers to phone in and indicate their support or opposition. Which of the following is true about the sample?

A. The sample is potentially biased, because they are selected by voluntary response.

B. This is a stratified random sample.

C. This is a multi-stage random sample.

D. This is a simple random sample.

Explanation: Since this survey asks viewers to call in, it relies on voluntary response, a potential source of bias.

3. A local police department performs a study on city residents' cell phone use while driving. Using a list of all residents within the city limits who have a valid driver's license, they choose 1,000 names at random and conduct a telephone survey of those individuals.

Identify the population of interest:

A. All city residents.

B. The 1,000 people called.

C. The police officers.

D. Everyone who has a driver's license in the US.

Explanation: The police department is interested in learning about all city residents.

4. A local police department performs a study on city residents' cell phone use while driving. Using a list of all residents within the city limits who have a valid driver's license, they choose 1,000 names at random and conduct a telephone survey of those individuals.

Identify the sampling frame:

A. The list of all residents who have a valid driver's license.

B. The 1,000 people called.

C. The police officers.

D. All city residents.

Explanation: The sampling frame is the list that the sample is actually taken from.

5. A local police department performs a study on city residents' cell phone use while driving. Using a list of all residents within the city limits who have a valid driver's license, they choose 1,000 names at random and conduct a telephone survey of those individuals.

Identify the sampling method

A. Simple random sampling.

B. Convenience sampling.

C. Voluntary response sampling.

D. Multi-stage random sampling.

Explanation: Since the names are chosen at random, this is a simple random sample.

[END SELF-CHECK]