Section 1.2 Sampling and Bias

Get Started – What is a simple linear equation and how do you solve it?

- How do you identify the type of sampling used in a statistical study?
- What is bias and how do you identify a potential source of bias in a statistical study?

Get Started – What is a simple linear equation and how do you solve it?

Key Terms

Linear equation	Solve an equation

Summary

The root of our word "Algebra" comes from the Arabic "al-jabr" which is often translated as: restoring. We think of solving equations as restoring them to their original state. We must undo the mathematical operations that have been performed. To undo addition, we will subtract. To undo subtraction, we will add. To undo multiplication, we will divide. To undo division, we will multiply. Equations typically involve more than one operation at a time. Since we are undoing the math, we will often follow the order of operations in reverse order: add or subtract before we multiply or divide.

Notes

Guided Example 1

Solve and check $2x+5=19.7$.

Solution We will undo the + 5 first, then we will undo the multiplying by 2.

$$\begin{aligned} 2x+5&=19.7 && \text{Original equation} \\ 2x+5-5&=19.7-5 && \text{Subtract 5 from both sides} \\ 2x&=14.7 && \text{Simplify} \\ \frac{2x}{2}&=\frac{14.7}{2} && \text{Divide both sides by 2} \\ x&=7.35 && \text{Simplify} \end{aligned}$$

To check the solution, substitute the solution into the original equation:

$$\begin{aligned} 2x+5&=19.7 && \text{Original equation} \\ 2(7.35)+5&=19.7 && \text{Substitute } x = 7.35 \\ 14.7+5&=19.7 && \text{Simplify} \\ 19.7&=19.7 && \end{aligned}$$

When checking it is important to go back to the original equation, if we only checked from $2x=14.7$ we aren't checking our first step of subtracting the five.

Practice

Solve and check $5a-18=-12$.

Most equations involve more than one operation and often operations occur more than once, so we need some guidelines for dealing with more complicated equations. The most important thing to remember is that we need to maintain balance. We need to treat each side of the equation equally. We will be using all our properties of equality. Our goal when solving linear equations is to get the variable to equal a single number: we need to isolate the variable. And we are trying to undo the existing math, so we will often be following the order of operations backwards.

Guidelines for Solving Linear Equations

- Simplify the expressions on each side of the equal sign separately: use distributive property to remove any grouping symbols & combine like terms
- Use the addition/subtraction properties of equality to move all the variables to one side of the equal sign and the constant terms to the other side of the equal sign

- Use the multiplication/division properties of equality to solve for the variable.
- CHECK: plug in your solution to make sure it works.

Notes

Guided Example 2

Solve and check $5(a+2)=75$

Solution We will distribute the 5 to remove the parentheses and then isolate the variable.

$5(a+2)=75$	Original equation
$5a+10=75$	Remove the parentheses by distributing the 5.
$5a+10-10=75-10$	Subtract 10 from both sides.
$5a=65$	Simplify.
$\frac{5a}{5}=\frac{65}{5}$	Divide both sides by 5
$a=13$	Simplify.

To check the solution, substitute the solution into the original equation:

$5(a+2)=75$	Original equation
$5(13+2)=75$	Substitute a = 13
$5(15)=75$	Simplify
$75=75$	

Practice

Solve and check $3(x+7)=19$.

How do you identify the type of sampling used in a statistical study?

Key Terms

Census
Simple random sample
Stratified sampling
Systematic sampling
Cluster sampling
Quota sampling
Convenience sampling

Summary

Now that you know that you must take samples in order to gather data, the next question is how best to gather a sample? There are many ways to take samples. Not all of them will result in a representative sample. Also, just because a sample is large does not mean it is a good sample. As an example, you can take a sample involving one million people to find out if they feel there should be more gun control, but if you only ask members of the National Rifle Association (NRA) or the Coalition to Stop Gun Violence, then you may get biased results. This means that the results of the sample do not reflect the results of the population. You need to make sure that you ask a cross-section of individuals. Let’s look at the types of samples that can be taken. Do realize that no sample is perfect and may not result in a representation of the population.

> **Census:** An attempt to gather measurements or observations from all the objects in the entire population.

A true census is very difficult to do in many cases. However, for certain populations, like the net worth of the members of the U.S. Senate, it may be relatively easy to perform a census. We should be able to find out the net worth of each member of the Senate since there are only 100 members. But, when our government tries to conduct the national census every 10 years, you can believe that it is impossible for them to gather data on each American.

The best way to find a sample that is representative of the population is to use a random sample. There are several different types of random sampling. Though it depends on the task at hand, the best method is often simple random sampling which occurs when you randomly choose a subset from the entire population.

> **Simple Random Sample:** Every sample of size *n* has the same chance of being chosen, and every individual in the population has the same chance of being in the sample.

An example of a simple random sample is to put all the names of the students in your class into a hat, and then randomly select five names out of the hat.

> **Stratified Sampling:** This is a method of sampling that divides a population into different groups, called strata, and then takes random samples inside each strata.

An example where stratified sampling is appropriate is if a university wants to find out how much time their students spend studying each week; but they also want to know if different majors spend more time studying than others. They could divide the student body into the different majors (strata), and then randomly pick several people in each major to ask them how much time they spend studying. The number of people asked in each major (strata) does not have to be the same.

> **Systematic Sampling:** This method is where you pick every k^{th} individual, where k is some whole number. This is used often in quality control on assembly lines.

For example, a car manufacturer needs to make sure that the cars coming off the assembly line are free of defects. They do not want to test every car, so they test every 100th car. This way they can periodically see if there is a problem in the manufacturing process. This makes for an easier method to keep track of testing and is still a random sample.

> **Cluster Sampling:** This method is like stratified sampling, but instead of dividing the individuals into strata, and then randomly picking individuals from each strata, a cluster sample separates the individuals into groups, randomly selects which groups they will use, and then takes a census of every individual in the chosen groups.

Cluster sampling is very useful in geographic studies such as the opinions of people in a state or measuring the diameter at breast height of trees in a national forest. In both situations, a cluster sample reduces the traveling distances that occur in a simple random sample. For example, suppose that the Gallup Poll needs to perform a public opinion poll of all registered voters in Colorado. To select a good sample using simple random sampling, the Gallup Poll would have to have all the names of all the registered voters in Colorado, and then randomly select a subset of these names. This may be very difficult to do. So, they will use a cluster sample instead. Start by dividing the state of Colorado up into categories or groups geographically. Randomly select some of these groups. Now ask all registered voters in each of the chosen groups. This makes the job of the pollsters much easier, because they will not have to travel over every inch of the state to get their sample, but it is still a random sample.

Quota Sampling: This is when the researchers deliberately try to form a good sample by *creating* a cross-section of the population under study.

For an example, suppose that the population under study is the political affiliations of all the people in a small town. Now, suppose that the residents of the town are 70% Caucasian, 25% African American, and 5% Native American. Further, the residents of the town are 51% female and 49% male. Also, we know information about the religious affiliations of the townspeople. The residents of the town are 55% Protestant, 25% Catholic, 10% Jewish, and 10% Muslim. Now, if a researcher is going to poll the people of this town about their political affiliation, the researcher should gather a sample that is representative of the entire population. If the researcher uses quota sampling, then the researcher would try to artificially create a cross-section of the town by insisting that his sample should be 70% Caucasian, 25% African American, and 5% Native American. Also, the researcher would want his sample to be 51% female and 49% male. Also, the researcher would want his sample to be 55% Protestant, 25% Catholic, 10% Jewish, and 10% Muslim. This sounds like an admirable attempt to create a good sample, but this method has major problems with selection bias.

The main concern here is when does the researcher stop profiling the people that he will survey? So far, the researcher has cross-sectioned the residents of the town by race, gender, and religion, but are those the only differences between individuals? What about socioeconomic status, age, education, involvement in the community, etc.? These are all influences on the political affiliation of individuals. Thus, the problem with quota sampling is that to do it right, you have to take into account all the differences among the people in the town. If you cross-section the town down to every possible difference among people, you end up with single individuals, so you would have to survey the whole town to get an accurate result. The whole point of creating a sample is so that you do not have to survey the entire population, so what is the point of quota sampling?

Note: The Gallup Poll did use quota sampling in the past, but does not use it anymore.

Convenience Sampling: As the name of this sampling technique implies, the basis of convenience sampling is to use whatever method is easy and convenient for the investigator. This type of sampling technique creates a situation where a random sample is not achieved. Therefore, the sample will be biased since the sample is not representative of the entire population.

For example, if you stand outside the Democratic National Convention to survey people exiting the convention about their political views. This may be a convenient way to gather data, but the sample will not be representative of the entire population.

Of all the sampling types, a random sample is the best type. Sometimes, it may be difficult to collect a perfect random sample since getting a list of all the individuals to randomly choose from may be hard to do.

Notes

Guided Example 3	Practice
Determine if the sample type is simple random sample, stratified sample, systematic sample, cluster sample, quota sample, or convenience sample.	Determine if the sample type is simple random sample, stratified sample, cluster sample, systematic sample, or convenience sample. A study to determine the opinion of Americans about the use of marijuana for medical purposes is being conducted using the following designs.
a. A researcher wants to determine the different species of trees that are in the Coconino National Forest. She divides the forest using a grid system. She then randomly picks 20 different sections and records the species of every tree in each of the chosen sections. **Solution** This is a cluster sample, since she randomly selected some of the groups, and all individuals in the chosen groups were surveyed.	a. The researchers attend a festival in a town in Kansas and ask all the people they can what their opinions are.
b. A pollster stands in front of an organic foods grocery store and asks people leaving the store how concerned they are about pesticides in their food. **Solution** This is a convenience sample, since the person is just standing out in front of one store. Most likely the people leaving an organic food grocery store are concerned about pesticides in their food, so the sample would be biased.	b. The researchers divide Americans into groups based on the person's race, and then take random samples from each group.
c. The Pew Research Center wants to determine the education level of mothers. They randomly ask mothers to say if they had some high school, graduated high school, some college, graduated from college, or advance degree. **Solution** This is a simple random sample, since the individuals were picked randomly.	c. The researchers number all Americans and call the 50th person on the list. Then they call every 10,000th person after the 50th person.

d. Penn State wants to determine the salaries of their graduates in the majors of agricultural sciences, business, engineering, and education. They randomly ask 50 graduates of agricultural sciences, 100 graduates of business, 200 graduates of engineering, and 75 graduates of education what their salaries are. **Solution** This is a stratified sample, since all groups were used, and then random samples were taken inside each group.	d. The researchers call every person in each of 10 area codes that were randomly chosen.
e. For the Ford Motor Company to ensure quality of their cars, they test every 130th car coming off the assembly line of their Ohio Assembly Plant in Avon Lake, OH. **Solution** This is a systematic sample since they picked every 130th car.	e. The researchers number every American, and then call all randomly selected Americans.
f. A town council wants to know the opinion of their residents on a new regional plan. The town is 45% Caucasian, 25% African American, 20% Asian, and 10% Native American. It also is 55% Christian, 25% Jewish, 12% Islamic, and 8% Atheist. In addition, 8% of the town did not graduate from high school, 12% have graduated from high school but never went to college, 16% have had some college, 45% have obtained bachelor's degree, and 19% have obtained a post-graduate degree. So the town council decides that the sample of residents will be taken so that it mirrors these breakdowns. **Solution** This is a quota sample, since they tried to pick people who fit into these subcategories.	

What is bias and how do you identify a potential source of bias in a statistical study?

Key Terms

Bias

Voluntary response bias

Perceived lack of anonymity

Selection bias

Self-interest study

Loaded questions

Non-response bias

Response bias

Summary

When we collect data, we often sample a population to measure a statistic. We hope that the statistic from the sample matches the corresponding parameter from the population.

> **Bias** is the tendency for a statistic from a sample to underestimate or overestimate a parameter from a population.

Two types of bias are commonly encountered when we collect data.

> **Sample bias** (selection bias) occurs when the sample chosen from the population is not representative of the population.
>
> **Nonresponse bias** occurs when the intended objects in the sample do not respond for many different reasons. Those who feel strongly about an issue will be more likely to participate.

The Literary Digest was a magazine that was founded in 1890. Starting with the 1916 U.S. presidential election, the magazine had predicted the winner of each election. In 1936, the Literary Digest predicted that Alfred Landon would win the election in a landslide over Franklin Delano Roosevelt with fifty-seven percent of the popular vote. The process for predicting the winner was that the magazine sent out ten million mock ballots to its subscribers and names of people who had automobiles and telephones. Two million mock ballots were sent back. Roosevelt won the election with 62% of the popular vote. (“Case Study 1: The 1936 Literary Digest Poll,” n.d.)

A side note is that while the Literary Digest was publishing its prediction, a man by the name of George Gallup also conducted a poll to predict the winner of the election. Gallup only polled about fifty thousand voters using random sampling techniques, yet his prediction was that

Roosevelt would win the election. His polling techniques were shown to be the more accurate method and have been used to present-day.

Because of the people whom the Literary Digest polled, they created a sample bias. The poll asked ten million people who owned cars, had telephones, and subscribed to the magazine. Today, you would probably think that this group of people would be representative of the entire U.S. However, in 1936 the country was during the Great Depression. The people polled were mostly in the upper middle to upper class. They did not represent the entire country. It did not matter that the sample was very large. The most important part of a sample is that it is representative of the entire population. If the sample is not, then the results could be wrong, as demonstrated in this case. It is important to collect data so that it has the best chance of representing the entire population.

When looking at the number of ballots returned, two million appears to be a very large number. However, ten million ballots were sent out. So that means that only about one-fifth of all the ballots were returned. This is known as a nonresponse bias. The only people who probably took the time to fill out and return the ballot were those who felt strongly about the issue. So, when you send out a survey, you must pay attention to what percentage of surveys are returned. If possible, it is better to conduct the survey in person or through the telephone. Most credible polls conducted today, such as Gallup, are conducted either in person or over the telephone. Do be careful though, just because a polling group conducts the poll in person or on the telephone does not mean that it is necessarily credible.

There are many other types of bias that may be encountered when data is collected for a sample.

> **Voluntary response bias** often occurs when the sample is volunteers. For example, suppose a survey is conducted among callers to a radio show to determine their attitudes towards vaccinations. The sample members are volunteers who might tend to have strong opinions regarding vaccinations. This overrepresentation might lead to statistics that do not represent the attitudes of the population.
>
> **Self-interest bias** may occur when the researchers have an interest in the outcome. Consider a recent study which found that chewing gum may raise math grades in teenagers. This study was conducted by the Wrigley Science Institute, a branch of the Wrigley chewing gum company. This is an example of a self-interest study; one in which the researches have a vested interest in the outcome of the study. While this does not necessarily ensure that the study was biased, it certainly suggests that we should subject the study to extra scrutiny.
>
> **Response bias** may occur when the responder gives inaccurate responses for any reason. Suppose a survey asks people “when was the last time you visited your doctor?” This might suffer from response bias, since many people might not remember exactly when they last saw a doctor and give inaccurate responses. Sources of response bias may be innocent, such as bad memory, or as intentional as pressuring by the pollster.

Perceived lack of anonymity is possible when the responder fears giving an honest answer might negatively affect them. Suppose a survey is being conducted to learn more about illegal drug use among college students. If a uniformed police officer is conducting the survey, then the results will very likely be biased since the college students may feel uncomfortable telling the truth to the police officer.

Loaded questions are questions where wording influences the responses. A question regarding the environment may ask "Do you think that global warming is the most important world environmental issue, or pollution of the oceans?" Alternatively, the question may be worded "Do you think that pollution of the oceans is the most important world environmental issue, or global warming?" The answers to these two questions will vary greatly simply because of how they are worded. The best way to handle a question like this is to present it in multiple choice format as follows:

> What do you think is the most important world environmental issue?
>
> a. Global warming
>
> b. Pollution of the oceans
>
> c. Other

Non-response bias may be an issue when people refusing to participate in the study can influence the validity of the outcome. If a telephone poll asks the question "Do you often have time to relax and read a book?", and 50% of the people called refused to answer the survey. It is unlikely that the results will be representative of the entire population. When people refuse to participate, we can no longer be so certain that our sample is representative of the population.

Notes

Guided Example 4	Practice
In each situation, identify a potential source of bias.	In each situation, identify a potential source of bias.
a. A survey asks how many sexual partners a person has had in the last year. **Solution** This survey suffers from response bias where the responder might give inaccurate responses. In this case, men are likely to over-report the number of sexual partners and women are likely to under-report the number of sexual partners.	a. A survey asks the following: Should the mall prohibit loud and annoying rock music in clothing stores catering to teenagers?
b. A radio station asks readers to phone in their choice in a daily poll. **Solution** The readers are volunteers and will be more likely to respond with strong opinions. The survey has the potential to suffer from voluntary response bias.	b. A survey asks people to report their actual income and the income they reported on their IRS tax form.
c. High school students are asked if they have consumed alcohol in the last two weeks. **Solution** Since students are asked a question whose response might impact them negatively, this is an example of perceived lack of anonymity.	c. A survey asks the following: Should the death penalty be permitted if innocent people might die?
d. The Beef Council releases a study stating that consuming red meat poses little cardiovascular risk. **Solution** The Beef Council has an interest in the results of the study so the study might suffer from self-interest bias.	d. To determine opinions on voter support for a downtown renovation project, a surveyor randomly questions people working in downtown businesses.

e. A poll asks “Do you support a new transportation tax, or would you prefer to see our public transportation system fall apart?” **Solution** The question uses the words “fall apart” in describing the potential failure of the public transportation system. This choice of words might influence responses so this is an example of a loaded question.	

Notes

How Do You Conduct a Study?

Now you know how to collect a sample, next you need to learn how to conduct a study. We will discuss the basics of studies, both observational studies and experiments.

Observational Study: This is where data is collected from just observing what is happening. There is no treatment or activity being controlled in any way. Observational studies are commonly conducted using surveys, though you can also collect data by just watching what is happening such as observing the types of trees in a forest.

Survey: Surveys are used for gathering data to create a sample. There are many kinds of surveys, but overall, a survey is a method used to ask people questions when interested in the responses. Examples of surveys are Internet and T.V. surveys, customer satisfaction surveys at stores or restaurants, new product surveys, phone surveys, and mail surveys. Most surveys are some type of public opinion poll.

Experiment: This is an activity where the researcher controls some aspect of the study and then records what happens. An example of this is giving a plant a new fertilizer, and then watching what happens to the plant. Another example is giving a cancer patient a new medication, and monitoring whether the medication stops the cancer from growing. There are many ways to do an experiment, but a clinical study is one of the more popular ways, so we will look at the aspects of this.

Clinical Study: This is a method of collecting data for a sample and then comparing that to data collected for another sample where one sample has been given some sort of treatment and the other sample has not been given that treatment (control). Note: There are occasions when you can have two treatments, and no control. In this case you are trying to determine which treatment is better.

Here are examples of clinical studies.

a. A researcher may want to study whether or not smoking increases a person's chances of heart disease.
b. A researcher may want to study whether a new antidepressant drug will work better than an old antidepressant drug.
c. A researcher may want to study whether taking folic acid before pregnancy will decrease the risk of birth defects.

Participants in a clinical study are broken into two groups.

Treatment Group: This is the group of individuals who are given some sort of treatment. The word treatment here does not necessarily mean medical treatment. The treatment is the cause, which may produce an effect that the researcher is interested in.

Control Group: This is the group of individuals who are not given the treatment. Sometimes, they may be given some old treatment, or sometimes they will not be given anything at all. Other times, they may be given a placebo (see below).

Any clinical study where the researchers compare the results of a treatment group versus a control group is called a **controlled study**. Any clinical study in which the treatment group and the control group are selected randomly from the population is called a **randomized controlled study**.

Notes

Guided Example 5	Practice
Determine the treatment group, control group, treatment, and control for each clinical study. a. A researcher may want to study whether or not smoking increases a person's chances of heart disease. **Solution** The treatment group is the people in the study who smoke, and the treatment is smoking. The control group is the people in the study who do not smoke, and the control is not smoking. b. A researcher may want to study whether a new antidepressant drug will work better than an old antidepressant drug. **Solution** The treatment group is the people in the study who take the new antidepressant drug and the treatment is taking the new antidepressant drug. The control group is the people in the study who take the old antidepressant drug and the control is taking the old antidepressant drug. *Note: In this case the control group is given some treatment since you should not give a person with depression a non-treatment.*	Determine the treatment group, control group, treatment, and control for each clinical study. A researcher may want to study whether taking folic acid before pregnancy will decrease the risk of birth defects.

There are other possible causes that may produce the effect of interest rather than the treatment under study. These causes are called **confounding variables**. Researchers minimize the effect of confounding variables by comparing the results from the treatment group versus the control group.

A **placebo** is sometimes used on the control group in a study to mimic the treatment that the treatment group is receiving. The idea is that if a placebo is used, then the people in the control group and in the treatment group will all think that they are receiving the treatment. However, the control group is merely receiving something that looks like the treatment but should have no effect on the outcome. An example of a placebo could be a sugar pill if the treatment is a drug in pill form.

Guided Example 6	Practice
For each situation, identify if a placebo is necessary to use.	For each situation, identify if a placebo is necessary to use.
a. A researcher may want to study whether smoking increases a person's chances of heart disease. **Solution** In this example, it is impossible to use a placebo. The treatment group is comprised of people who smoke, and the control group is comprised of people who do not smoke. There is no way to get the control group to think that they are smoking as well as the treatment group.	a. A researcher may want to study whether taking folic acid before pregnancy will decrease the risk of birth defects.
b. A researcher may want to study whether a new antidepressant drug will work better than an old antidepressant drug. **Solution** In this example, a placebo is not needed since we are comparing the results of two different antidepressant drugs.	b. A researcher wants to determine if morphine reduces pain during dental tooth extractions.

Usually, when a placebo is used in a study, the people in the study will not know if they received the treatment or the placebo until the study is completed. In other words, the people in the study do not know if they are in the treatment group or in the control group. This type of study is called a **blind study**. *Note: When researchers use a placebo in a blind study, the people in the study are told ahead of time that they may be getting the actual treatment, or they may be getting the placebo.*

Sometimes when researchers are conducting a very extensive study using many healthcare workers, the researchers will not tell the people in the study or the healthcare workers which patients will receive the treatment and which patients will receive the placebo. In other words, the healthcare workers who are administering the treatment or placebo to the people in the study do not know which people are in the treatment group and which people are in the control group. This type of study is called a **double-blind study**.

Whether you are doing an observational study or an experiment, you need to figure out what to do with the data. You will have many data values that you collected, and it sometimes helps to calculate numbers from these data values. Whether you are talking about the population or the sample, determines what we call these numbers. As mentioned in an earlier section, a parameter is a numerical value calculated from a population. A statistics is a numerical value calculated from a sample, and used to estimate the parameter.

Some examples of parameters that can be estimated from statistics are the percentage of all people who strongly agree to a question and mean net worth of all Americans. The statistic would be the percentage of people asked who strongly agree to a question, and the mean net worth of a certain number of Americans.

Parameters are usually denoted with Greek letters. This is not to make you learn a new alphabet. It is because there just are not enough letters in our alphabet. Also, if you see a letter you do not know, then you know that the letter represents a parameter. Examples of letters that are used are μ (mu), and σ (sigma). Statistics are usually denoted with our alphabet. In some cases, we try to use a letter that would be equivalent to the Greek letter. Examples of letter that are used are $\overline{x}$ (x-bar), s, and r.