

## Study Design

---

### 2.1 Week 2 Reading Guide: Sampling, Experimental Design, and Scope of Inference

Sections 2.2 (Observational studies), 2.3 (Experiments), and 2.4 (Scope of inference)

Videos

- 2.2to2.4

#### Reminders from Section 1.2

**Explanatory variable:** The variable researchers think *may be* affecting the other variable. What the researchers control/assign in an experiment. If comparing groups, the explanatory variable puts the observational units into groups.

**Response variable:** The variable researchers think *may be* influenced by the other variable. This variable is always observed, never controlled or assigned.

#### Vocabulary

Observational study:

Observational data:

Prospective study:

Retrospective study:

Confounding variable:

Experiment:

Randomized experiment:

Blocking:

Treatment group:

Control group:

Blinding:

Placebo:

Placebo effect:

Scope of inference:

Generalizability:

Causation:

## Notes

What are the four principles of a well-designed randomized experiment?

Fill in the appropriate scope of inference for each study design.

Selection of Cases	Study Type	
	Randomized experiment	Observational study
Random sample (and no other sampling bias)		
Non-random sample (or other sampling bias)		

True or False: Observational studies can show an association between two variables, but cannot determine a causal relationship.

True or False: In order for an experiment to be valid, a placebo must be used.

True or False: If random sampling of the target population is used, and no other types of bias are suspected, results from the sample can be generalized to the entire target population.

True or False: If random sampling of the target population is used, and no other types of bias are suspected, results from the sample can be inferred as a causal relationship between the explanatory and response variables.

## 2.2 Activity 2A: American Indian Address

### 2.2.1 Learning outcomes

- Explain why a sampling method is unbiased or biased.
- Identify biased sampling methods.
- Explain the purpose of random selection and its effect on scope of inference.

### 2.2.2 Terminology review

In today's activity, we will examine unbiased and biased methods of sampling. Some terms covered in this activity are:

- Random sample
- Unbiased vs biased methods of selection
- Generalization

To review these concepts, see Section 2.1 in the textbook.

### 2.2.3 American Indian Address

For this activity, you will read a speech given by Jim Becenti, a member of the Navajo American Indian tribe, who spoke about the employment problems his people faced at an Office of Indian Affairs meeting in Phoenix, Arizona, on January 30, 1947 (Moquin and Van Doren 1973). His speech is below:

**It is hard for us to go outside the reservation where we meet strangers. I have been off the reservation ever since I was sixteen. Today I am sorry I quit the Santa Fe [Railroad]. I worked for them in 1912-13. You are enjoying life, liberty, and happiness on the soil the American Indian had, so it is your responsibility to give us a hand, brother. Take us out of distress. I have never been to vocational school. I have very little education. I look at the white man who is a skilled laborer. When I was a young man I worked for a man in Gallup as a carpenter's helper. He treated me as his own brother. I used his tools. Then he took his tools and gave me a list of tools I should buy and I started carpentering just from what I had seen. We have no alphabetical language.**

**We see things with our eyes and can always remember it. I urge that we help my people to progress in skilled labor as well as common labor. The hope of my people is to change our ways and means in certain directions, so they can help you someday as taxpayers. If not, as you are going now, you will be burdened the rest of your life. The hope of my people is that you will continue to help so that we will be all over the United States and have a hand with you, and give us a brotherly hand so we will be happy as you are. Our reservation is awful small. We did not know the capacity of the range until the white man come and say "you raise too much sheep, got to go somewhere else," resulting in reduction to a skeleton where the Indians can't make a living on it. For eighty years we have been confused by the general public, and what is the condition of the Navajo today? Starvation! We are starving for education. Education is the main thing and the only thing that is going to make us able to compete with you great men here talking to us.**

### By eye selection

1. Circle ten words in Jim Becenti's speech which are a representative sample of the length of words in the entire text. Describe your method for selecting this sample.
2. Fill in the table below with your selected words from the previous question and the length of each word (number of letters/digits in the word):

Observation	Word	Length
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

3. Calculate the mean (average) word length in your selected sample. Is this value a parameter or a statistic?
4. Report your mean word length to your instructor. Your instructor will guide the class in creating a visualization of the distribution of results generated by your class. Draw a picture of the plot here. Include a descriptive  $x$ -axis label.
5. Based on the plot of sample mean word lengths in question 4, what is your best guess for the average word length of the population of all 359 words in the speech?

6. The true mean word length of the population of all 359 words in the speech is 3.95 letters. Is this value a parameter or a statistic?

Where does the value of 3.95 fall in the plot created in question 4? Near the center of the distribution? In the tails of the distribution?

7. If the class samples were truly representative of the population of words, what proportion of sample means would you expect to be below 3.95?
8. Using the graph created in question 4, estimate the proportion of students' computed sample means that were lower than the true mean of 3.95 letters?
9. Based on your answers to questions 7 and 8, would you say the sampling method used by the class is biased or unbiased? Justify your answer.
10. If the sampling method is biased, what type of sampling bias (selection, response, non-response) is present? What is the direction of the bias, i.e., does the method tend to overestimate or underestimate the population mean word length?
11. Should we use results from our by eye samples to make a statement about the word length in the population of words in Becenti's address? Why or why not?

### 2.2.4 Take-home messages

1. When we use a biased method of selection, we will over or underestimate the parameter.
2. To see if a method is biased, we compare the distribution of the estimates to the true value. We want our estimate to be on target or unbiased. When using unbiased methods of selection, the mean of the distribution matches or is very similar to our true parameter.
3. If the sampling method is biased, inferences made about the population based on a sample estimate will not be valid.

### **2.2.5 Additional notes**

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

## 2.3 Activity 2B: American Indian Address (continued)

### 2.3.1 Learning outcomes

- Explain the purpose of random selection and its effect on scope of inference.
- Select a simple random sample from a finite population using a random number generator.
- Explain why a sampling method is unbiased or biased.
- Explain the effect of sample size on sampling variability.

### 2.3.2 Terminology review

In today's activity, we will examine unbiased and biased methods of sampling. Some terms covered in this activity are:

- Random sample
- Unbiased vs biased methods of selection
- Generalization

To review these concepts, see Section 2.1 in the textbook.

#### Random selection

Today we will return to the American Indian Address introduced in Activity 2A. Suppose instead of attempting to select a representative sample by eye (which did not work), each student used a random number generator to select a simple random sample of 10 words. A **simple random sample** relies on a random mechanism to choose a sample, without replacement, from the population, such that every sample of size 10 is equally likely to be chosen.

To use a random number generator to select a simple random sample, you first need a numbered list of all the words in the population, called a **sampling frame**. You can then generate 10 random numbers from the numbers 1 to 359 (the number of words in the population), and the chosen random numbers correspond to the chosen words in your sample.

1. Use the random number generator at <https://istats.shinyapps.io/RandomNumbers/> to select a simple random sample from the population of all 359 words in the speech.
- Set “Choose Minimum” to 1 and “Choose Maximum” to 359 to represent the 359 words in the population (the sampling frame).
  - Set “How many numbers do you want to generate?” to 10 and ensure the “No” option is selected under “Sample with Replacement?”
  - Click “Generate”.



Fill in the table below with the random numbers selected and use the Bcenti.csv data file found on D2L to determine each number's corresponding word and word length (number of letters/digits in the word):

Observation	Number	Word	Length
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

- Calculate the mean word length in your selected sample in question 1. Is this value a parameter or a statistic?
- Report your mean word length to your instructor. Your instructor will guide the class in creating a visualization of the distribution of results generated by your class. Draw a picture of the plot here. Include a descriptive  $x$ -axis label.
- Where does the value 3.95, the true mean word length, fall in the distribution created in question 3? Near the center of the distribution? In the tails of the distribution?

5. How does the plot generated in question 3 compare to the plot generated in question 4 from Activity 2A?

Which features are similar?

Which features differ?

Why didn't everyone get the same sample mean?

One set of randomly generated sample mean word lengths from a single class may not be large enough to visualize the distribution results. Let's have a computer generate 1,000 sample mean word lengths for us.

- Navigate to the “One Variable with Sampling” Rossman/Chance web applet: <http://www.rossmanchance.com/applets/2021/sampling/OneSample.html?population=gettysburg>.
  - Click “Clear” below the text box containing data from the Gettysburg address to delete that data set.
  - Download the Becenti.csv file from D2L and open the spreadsheet on your computer.
  - Copy and paste the population of word lengths (column C) into the applet from the data set provided making sure to include the header. Click “Use Data”. Verify that the mean for the data set is 3.953 with a sample size of 359. If these are not the values you got, check with your instructor for help with copying in the data set correctly.
  - Click the check-box for “Show Sampling Options”
  - Select 1000 for “Number of samples” and select 10 for the “Sample size”.
  - Click “Draw Samples”.
6. The plot labeled “Statistics” displays the 1,000 randomly generated sample mean word lengths. Sketch this plot below. Include a descriptive  $x$ -axis label and be sure to write down the provided mean and SD (standard deviation) of the distribution.

7. What is the center value of the distribution created in question 6?

8. Explain why the sampling method of using a random number generator to generate a sample is a “better” method than choosing 10 words “by eye”.
9. Is random selection an unbiased method of selection? Explain your answer. Be sure to reference your plot from question 6.

## Effect of sample size

We will now consider the impact of sample size.

10. First, consider if each student had selected 20 words, instead of 10, by eye. Do you think this would make the plot from question 4 in Activity 2A centered on 3.95 (the true mean word length)? Explain your answer.
11. Now we will select 20 words instead of 10 words at random.
  - In the “One Variable with Sampling” Rossman/Chance web applet(<http://www.rossmanchance.com/applets/2021/sampling/OneSample.html?population=gettysburg>), change the Sample size to 20.
  - Click “Draw Samples”.

The plot labeled “Statistics” displays the 1,000 randomly generated sample mean word lengths. Sketch this plot below. Include a descriptive  $x$ -axis label and be sure to write down the provided mean and SD (standard deviation) of the distribution.

12. Compare the distribution created in question 11 to the one created in question 6.

Which features are similar?

Which features differ?

13. Compare the spreads of the plots in question 11 and in question 6. You should see that in one plot all sample means are closer to the population mean than in the other. Which plot shows this?

14. Using the evidence from your simulations, answer the following research questions:

Does changing the sample size impact whether the sample estimates are unbiased? Explain your answer.

Does changing the sample size impact the variability (spread) of sample estimates? Explain your answer

15. What is the purpose of random selection of a sample from the population?

### 2.3.3 Take-home messages

1. Random selection is an unbiased method of selection.
2. To determine if a sampling method is biased or unbiased, we compare the distribution of the estimates to the true value. We want our estimate to be on target or unbiased. When using unbiased methods of selection, the mean of the distribution matches or is very similar to our true parameter.
3. Random selection eliminates selection bias. However, random selection will not eliminate response or non-response bias.
4. The larger the sample size, the more similar (less variable) the statistics will be from different samples.
5. Sample size has no impact on whether a *sampling method* is biased or not. Taking a larger sample using a biased method will still result in a sample that is not representative of the population.

### 2.3.4 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

## 2.4 Week 2 Lab: Study Design

### 2.4.1 Learning outcomes

- Explain the purpose of random assignment and its effect on scope of inference.
- Identify whether a study design is observational or an experiment.
- Identify confounding variables in observational studies and explain why they are confounding.

### 2.4.2 Terminology review

In this activity, we will examine different study designs, confounding variables, and how to determine the scope of inference for a study. Some terms covered in this activity are:

- Scope of inference
- Explanatory variable
- Response variable
- Confounding variable
- Experiment
- Observational study

To review these concepts, see Sections 2.2 through 2.5 in the textbook.

### 2.4.3 General information labs

Remember that each Friday you will complete a lab. Questions are selected from each lab to be turned in on Gradescope. The questions to be submitted on Gradescope are bolded in the lab. As you work through the lab have the Gradescope lab assignment open so that you can answer those questions as you go.

## Study design

The two main study designs we will cover are **observational studies** and **experiments**. In observational studies, researchers have no influence over which subjects are in each group being compared (though they can control other variables in the study). An experiment is defined by assignment of the treatment groups of the *explanatory variable*, typically via random assignment. In today's activity we will discover the purpose behind random assignment.

For the next exercises, identify the explanatory variable, the response variable, and the study design (observational study or experiment).

1. The pharmaceutical company Moderna Therapeutics, working in conjunction with the National Institutes of Health, conducted Phase 3 clinical trials of a vaccine for COVID-19 last fall. US clinical research sites enrolled 30,000 volunteers without COVID-19 to participate. Participants were randomly assigned to receive either the candidate vaccine or a saline placebo. They were then followed to assess whether or not they developed COVID-19. The trial was double-blind, so neither the investigators nor the participants knew who was assigned to which group.

Explanatory variable:

Response variable:

Study design:

2. **In another study, a local health department randomly selected 1000 US adults without COVID-19 to participate in a health survey. Each participant was assessed at the beginning of the study and then followed for one year. They were interested to see which participants elected to receive a vaccination for COVID-19 and whether any participants developed COVID-19.**

Explanatory variable:

Response variable:

Study design:

## Atrial fibrillation

Atrial fibrillation is an irregular and often elevated heart rate. In some people, atrial fibrillation will come and go on its own, but others will experience this condition on a permanent basis. When atrial fibrillation is constant, medications are required to stabilize the patient's heart rate and to help prevent blood clots from forming. Pharmaceutical scientists at a large pharmaceutical company believe they have developed a new medication that effectively stabilizes heart rates in people with permanent atrial fibrillation. They set out to conduct a trial study to investigate the new drug. The scientists will need to compare the proportion of patients whose heart rate is stabilized between two groups of subjects, one of whom is given a placebo and the other given the new medication.

3. Identify the explanatory and response variable in this trial study.

Explanatory variable:

Response variable:

Suppose 24 subjects with permanent atrial fibrillation have volunteered to participate in this study:

Self-identified males: Paul, Antonio, Davieon, Chao, Aryan, Jabari, Tong, Andres, John, Liu, Lucas, Rashidi, Shiwoo, Jihoon, Alejandro, Daniel

Self-identified females: An, Nailah, Jasmine, Ka Nong, Keyaina, Mary, Adah, Sassandra

4. Is this a simple random sample or a convenience sample? How do you know?
5. Based on the sampling method, to what population should the results of this study be generalized?
6. One way to separate into two groups would be give all the males the placebo and all the females the new drug. Would this be a reasonable strategy? Explain your answer.
7. Could the scientists fix the problem with the strategy presented in question 6 by creating equal sized groups by putting 4 males and 8 females into the drug group and the remaining 12 males in the placebo group? Explain your answer.
8. A third strategy would be to **block** on sex. In this type of study, the scientists would assign 4 females and 8 males to each group. Using this strategy, what proportion of males is in each group?
9. **Assume the scientists used the strategy in question 8, but they put the four tallest females and eight tallest males into the placebo group and the remaining subjects into the control group. They found that the proportion of patients whose heart rate stabilized is higher in the drug group than the placebo group.**

Could that difference be due to the sex of the subjects? Explain your answer.

Could it be due to other variables? Explain your answer.

While the strategy presented in question 9 controlled for the sex of the subject, there are more potential **confounding variables** in the study. A confounding variable is a variable that is *both*

1. associated with the explanatory variable, *and*
2. associated with the response variable.

When both these conditions are met, if we observe an association between the explanatory variable and the response variable in the data, we cannot be sure if this association is due to the explanatory variable or the confounding variable—the explanatory and confounding variables are “confounded.”

**Random assignment** means that subjects in a study have an equally likely chance of receiving any of the available treatments.

10. You will now investigate how randomly assigning subjects impacts a study’s scope of inference.
  - Navigate to the “Randomizing Subjects” applet under the “Other Applets” heading at: <http://www.rossmanchance.com/ISIapplets.html>. This applet lists the sex and height of each of the 24 subjects. Click “Show Graphs” to see a bar chart showing the sex of each subject. Currently, the applet is showing the strategy outlined in question 7.
  - Click “Randomize”.

In this random assignment, what proportion of males are in group 1 (the placebo group)?

What proportion of males are in group 2 (the drug group)?

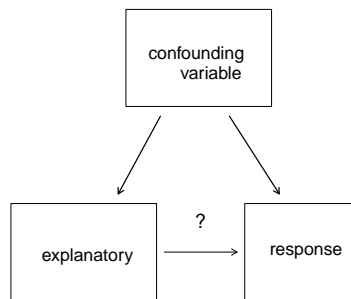
What is the difference in proportion of males between the two groups (placebo - drug)?

11. Notice the difference in the two proportions is shown as a dot in the plot at the bottom of the web page. Un-check the box for Animate above “Randomize” and click “Randomize” again. Did you get the same difference in proportion of males between the placebo and drug groups?
12. Change “Replications” to 998 (for 1000 total). Click “Randomize” again. Sketch the plot of the distribution of difference in proportions from each of the 1000 random assignments here. Be sure to include a descriptive  $x$ -axis label.



13. Does random assignment *always* balance the placebo and drug groups based on the sex of the participants? Does random assignment *tend* to make the placebo and drug groups *roughly* the same with respect to the distribution of sex? Use your plot from question 12 to justify your answers.
14. Change the drop-down menu below Group 2 from “sex” to “height”. The applet now calculates the average height in the placebo and drug groups for each of the 1000 random assignments. The dot plot displays the distribution of the difference in mean heights (placebo - drug) for each random assignment. Based on this dot plot, is height distributed equally, on average, between the two groups? Explain how you know.

The diagram below summarizes these ideas about confounding variables and random assignment. When a confounding variable is present (such as sex or height), and an association is found in a study, it is impossible to discern what caused the change in the response variable. Is the change the result of the explanatory variable or the confounding variable? However, if all confounding variables are *balanced* across the treatment groups, then only the explanatory variable differs between the groups and thus *must have caused* the change seen in the response variable.





15. What is the purpose of random assignment of the subjects in a study to the explanatory variable groups?
16. Suppose in this study on atrial fibrillation, the scientists did randomly assign groups and found that the drug group has a higher proportion of subjects whose heart rates stabilized than the placebo group. Can the scientists conclude the new drug *caused* the increased chance of stabilization? Explain your answer.

17. Both the sampling method (which we covered last week) and the study design will help to determine the *scope of inference* for a study: To *whom* can we generalize, and can we conclude *causation or only association*? Use the table below to determine the scope of inference of this trial study described in question 16.

*Scope of Inference:* If evidence of an association is found in our sample, what can be concluded?

	Study Type		
Selection of cases	Randomized experiment	Observational study	
Random sample (and no other sampling bias)	Causal relationship, and can generalize results to population.	Cannot conclude causal relationship, <b>but</b> can generalize results to population.	→ Inferences to population can be made
No random sample (or other sampling bias)	Causal relationship, <b>but</b> cannot generalize results to a population.	Cannot conclude causal relationship, <b>and</b> cannot generalize results to a population.	→ Can only generalize to those similar to the sample due to potential sampling bias

  
 Can draw cause-and-  
effect conclusions

  
 Can only discuss association  
due to potential confounding  
variables

18. Use the table to determine the scope of inference for the study in question 1.
19. Use the table to determine the scope of inference for the study in question 2.

#### 2.4.4 Take-home messages

1. The study design (observational study vs, experiment) determines if we can draw causal inferences or not. If an association is detected, a randomized experiment allows us to conclude that there is a causal (cause-and-effect) relationship between the explanatory and response variable. Observational studies have potential confounding variables within the study that prevent us from inferring a causal relationship between the variables studied.
2. Confounding variables are variables not included in the study that are related to both the explanatory and the response variables. When there are potential confounding variables in the study we cannot draw causal inferences.
3. Random assignment balances confounding variables across treatment groups. This eliminates any possible confounding variables by breaking the connections between the explanatory variable and the potential confounding variables.
4. Observational studies will always carry the possibility of confounding variables. Randomized experiments, which use random assignment, will have no confounding variables.

### **2.4.5 Additional notes**

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.