Sampling Bias: American Indian Address

Your Name: _____

October 27, 2022

Learning outcomes

- Explain why a sampling method is unbiased or biased.
- Identify various biased sampling methods.
- 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 the effect of sample size on sampling variability.

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
- Types of samping bias
- Generalization

Part One: 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. 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

5. 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.

6. 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):

7. Calculate the mean word length in your selected sample. Is this value a parameter or a statistic?

Observation	W
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

8. Report your mean word length to Dr. Theobold to create a visualization of the distribution of results generated by the class. Draw a picture of the plot here. Be sure to include a descriptive x-axis label!

9. Based on the plot of sample mean word lengths in question 8, what is your best guess for the average word length of the population of all 359 words in the speech?

10. 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?

11. Where does the value of 3.95 fall in the plot above?

12. If your samples were truly representative, what proportion of sample means would you expect to be below 3.95?

13. What proportion of students' computed sample means were lower than the true mean of 3.95 letters?

14. Based on your answers to questions 11 and 12, would you say the sampling method used by the class is biased or unbiased? Justify your answer.

15. If the sampling method is biased, what type of bias is present? What is the direction of the bias, i.e., does the method tend to overestimate or underestimate the population mean word length?

16. 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?

Part Two: Random selection

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?"

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

2. Calculate the mean word length in your selected sample in question 1. Is this value a parameter or a statistic?

3. Report your mean word length to Dr. Theobold, who will guide you 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.

4. Where does the value 3.95, the true mean word length, fall in the distribution created in question 4?

5. How does the plot generated in question 4 compare to the plot generated in question 8 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 Canvas 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 Dr. Theobold 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".

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

- Click "Draw Sample(s)".

6. The plot labeled "Statistic" 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 7?

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 8 in Part One 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 Sample(s)".

The plot labeled "Statistic" displays the 1,000 randomly generated sample 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 of sample estimates? Explain your answer

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

Take-home messages

1. Random selection is an unbiased method of selection.
2. When we use a biased method of selection, we will over or underestimate the parameter.
3. 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 our true parameter.
4. If the sampling method is biased, inferences made about the population based on a sample estimate will not be valid.
5. Random selection eliminates selection bias. Random selection will not eliminate response or non-response bias however.
6. The larger the sample size, the more similar (less variable) the statistics will be from different samples.
7. 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.