Sampling Bias: American Indian Address

Your Name: Dr. Theobold's Key

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 sampling 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.

   Words chosen highlighted in yellow.

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):

Word Selected (by you)	Length (in characters)
1. my	2
2. skilled	7
3. life	4
4. go	2
5. Navajo	6
6. go	2
7. Santa	5

Word Selected (by you)	Length (in characters)
8. white	5
9. is	2
10. young	5

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

   Mean = 4, this is a **statistic** as it comes from my **sample**

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!

[Visualization created in class]

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 358 words in the speech?

   The center of the distribution was about 7.5, so that would be my best guess for the average word length of all of the words in the speech.

10. The true mean word length of the population of all 358 words in the speech is 3.95 letters. Is this value a parameter or a statistic?

    This value is a **parameter** because it refers to the **population** of all the words in the speech.

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

    This value falls in the far-right tail of the distribution. Based on the distribution of means, this value would be considered unusual or an "outlier."

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

    If the samples were representative, then about half (50%) should be below 3.95 and about 50% should be above 3.95.

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

In our class, only 1 of the means was lower than 3.95 (out of 35), which is only 2.85%.

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.

    Yes! There are definitely not 50% of the means below 3.95.

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?

    Our samples seem to be overestimating the mean word length, since the center of our distribution was closer to 7.5 and not 3.95.

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?

    No! When we have biased samples, we **should not** make statements about the population based on our sample. With a sample that is really far away from the truth, we could mislead people to think that **all** the words in the speech were very long.

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 358 (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 358 words in the speech.

- Set "Choose Minimum" to 1 and "Choose Maximum" to 358 to represent the 358 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):

Computer words: 252, 143, 230, 36, 182, 349, 49, 146, 118, 276

Word Selected (by the computer)	Length (in characters)
1. brotherly	9
2. from	4
3. to	2
4. I	1
5. hope	4
6. able	4
7. on	2
8. had	3
9. brother	7
10. until	5

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

   mean = 4.1, this is another statistic since it was calculated from a sample

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.

   [visualization created in class]

4. Where does the value 3.95, the true mean word length, fall in the distribution created in question 3?

The value of 3.95 falls closer to the center of the distribution.

5. How does the plot generated in question 3 compare to the plot generated in question 8 from Part One?

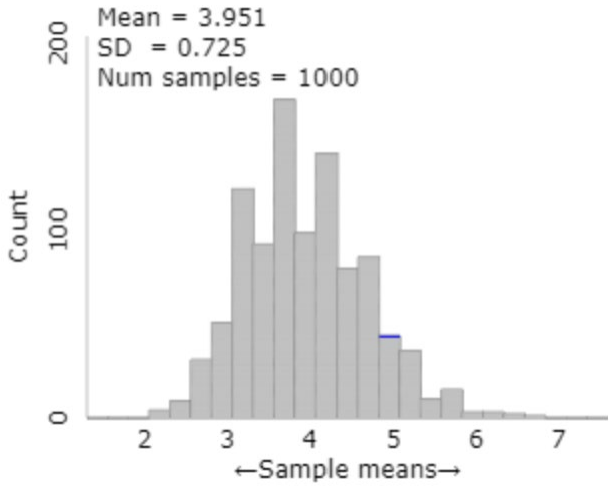
- Which features are similar?
  Both have ranges that go from about 2 to about 9, both are unimodal (have one peak)
- Which features differ?
  The centers of the two distributions differ, one is centered around 7.5 and the other is centered around 4.
- Why didn't everyone get the same sample mean?
  Because of randomness! The computer randomly selected words (from 1 to 358), which should make it really unlikely that two people would get exactly the same numbers chosen. There are many samples that result in the same mean, but not every sample will have the same mean because we are selecting the words randomly.

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 358. 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".
- Click "Draw Sample(s)".

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.

Statistics:



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

   The center is about 3.951 (or the mean).

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".

   Looking at the distribution, about 50% of the samples are below the true value of 3.95 and about 50% are above. That means we are overestimating about as often as we are underestimating, and we do not have bias in our samples.

9. Is random selection an unbiased method of selection? Explain your answer. Be sure to reference your plot from question 6!

   Yes! The distribution of these samples is centered at the true value. So, the majority of the time, we will get a sample that is close to the true value of the parameter (population mean).

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.

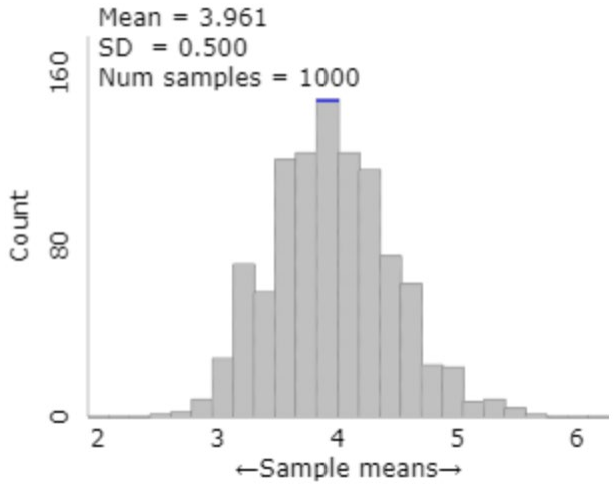
    As a statistician, I know the answer to this question. So, instead let me summarize what I've seen students hypothesize. I sometimes see people think that the sample will become even more biased, if we are selecting more words that are too long.

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, click "Reset" (next to "Draw Samples").
- Change the Sample size to 20.
- Click "Draw Sample(s)".

The plot labeled "Statistics:" 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.

Statistics:



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

- Which features are similar?
  The centers are similar! They are both centered around 3.95. Both distributions have one peak (unimodal) and are symmetrical (bell shaped).

- Which features differ?
  The range of values on the distribution! Before we had samples below 2 and above 7, but now the means are between about 2.5 and 6.

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?

    The second plot has a smaller standard deviation (0.500 compared to 0.725). This means in the second plot, the samples are closer to the population mean of 3.95.

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.

No! Both distributions were centered at the population parameter (3.95).

- Does changing the sample size impact the variability of sample estimates? Explain your answer

  Yes! When increasing the sample size from 10 to 20 we saw the spread of the distribution (the standard deviation) decrease.

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

    Our biased sample was not centered at the correct value, but when using the computer to randomly select words our distribution was centered at the correct value. So, random samples ensure our statistics are close to the population parameter.