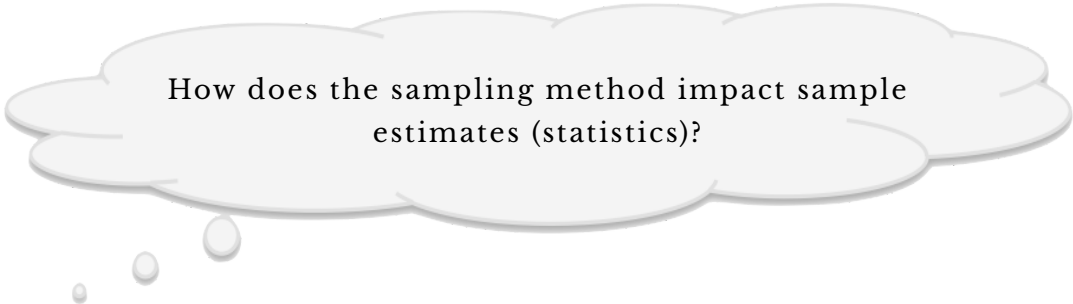


Gettysburg Address



In statistical inference, generalization refers to the process of using sample data to draw conclusions about the larger population from which the sample was drawn. Statisticians are typically concerned with making inferences about some population parameter using a sample statistic (*Remember:* Population summary measures are called **parameters**. Sample estimates of parameters are referred to as **statistics**.) Whether that sample statistic is a statistically good estimate of the population parameter depends on whether the sampling method used is biased. In this activity you will begin by exploring the following question:



How does the sampling method impact sample estimates (statistics)?

To help answer this research question, you are going to compare two different sampling methods using the population of 268 words in the passage on the following page. The passage is, of course, Lincoln's *Gettysburg Address*, given November 19, 1863 on the battlefield near Gettysburg, PA.

Four score and seven years ago, our fathers brought forth upon this continent a new nation: conceived in liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war.

We have come to dedicate a portion of that field as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we cannot dedicate, we cannot consecrate, we cannot hallow this ground. The brave men, living and dead, who struggled here have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember, what we say here, but it can never forget what they did here.

It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us, that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion, that we here highly resolve that these dead shall not have died in vain, that this nation, under God, shall have a new birth of freedom, and that government of the people, by the people, for the people, shall not perish from the earth.

The goal in many studies is to provide information about some characteristic of a population. For example, you may want to say something about the percentage of Americans who would support a particular piece of legislation. Or, you may want to provide information about the average amount of time University of Minnesota students take to graduate. One potential solution to obtain such information would be to collect the necessary data from every member of the target population.

In many studies, however, it may not be feasible given time and money constraints to collect data from each member of the population. In these cases it is only possible to consider data collected for a smaller subset, or **sample** from that population. In these cases, the characteristic of the population would be estimated from the sample data and inferences would be drawn about the population. The key is then to carefully select the sample so that the results estimated from the sample are representative of the characteristic in the larger population.

The **population** is the entire collection of who or what (e.g., the observational units) that you would like to draw inferences about. A **sample** is a subset of observational units from the population.

Circle a sample of ten words in the text of the Gettysburg Address (the population) such that the sample you select is representative (i.e., has the same characteristics) of the population.

12. Describe how the ten words in your sample are representative of the 268 words in the population.

13. Record the ten sampled words and their lengths:

Word	Length

14. Determine the average (mean) word length for your sample. This sample average (a statistic) is an estimate of the average word length in the population.

Add your sample estimate to the case table on the instructor's computer.

When the sampling method produces characteristics of the sample that systematically differ from those characteristics of the population, you say that the **sampling method is biased**. To try to eliminate potential biases, it is better to take a random sample. This should create a representative sample, no matter what variable is focused on. Humans are not very good “random samplers”, so it is important to use other techniques to do the sampling for us.

Simple Random Sampling

A **simple random sample** (SRS) is a specific type of random sample. It gives every observational unit in the population the same chance of being selected. In fact, it gives every sample of size n the same chance of being selected. In this example you want every possible subset of ten words that could be sampled to have the same probability of being selected.

The first step in drawing a simple random sample is to obtain a **sampling frame** or a list of each member of the population. Then, you can use software to randomly select a sample from the sampling frame. We have already prepared a sampling frame of the words in the Gettysburg Address for you and saved it in a Tinkerplots™ file.

Use TinkerPlots™ to Draw a SRS

- Open the file *gettysburg.tp*.
- Draw a simple random sample of ten words from the sampler.

19. Record the ten randomly sampled words and their lengths:

Word	Length

Use TinkerPlots™ to automatically compute the length of each word in your sample. To do this,

- Create a new attribute in the case table called *wordLength*.
- Right-click the attribute name *wordLength* and select the **Formula Editor**.
- Select **stringLength()** from the **Text** functions, and add the sampled words attribute between the parentheses.

20. Use TinkerPlots™ to plot and compute the mean word length for your ten randomly sampled words. Record the mean below.

21. Use **Collect** to carry out 500 trials of this simulation in which you randomly sample ten words and compute their mean length. Sketch the plot of these means. Make sure to label the axis appropriately.

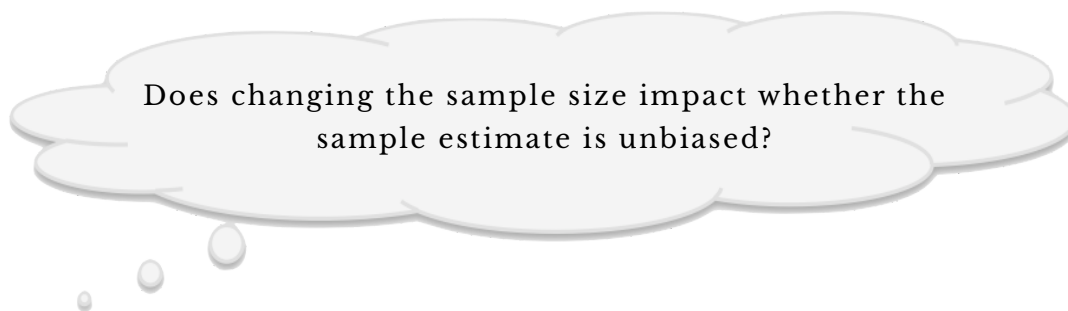
22. If the **sampling method is unbiased** the sample statistics should be centered at the population average word length of 4.3. Does simple random sampling produce an unbiased estimate of the population average? Explain.

Sample Size

Even when an unbiased sampling method, such as simple random sampling, is used to select a sample, you do not expect the estimate from each individual

sample drawn to match the population average exactly. You should see, however, that the estimates are just as likely to over- or underestimate the population parameter. Because of this predictability to the variation in the possible sample estimates, inferences drawn about the population are said to be valid.

On the other hand, if the sampling method is biased, any inferences made about the population based on a sample estimate may not be valid. In such cases the estimate of the parameter is more likely to be too large or too small compared to the parameter. It is therefore very important to determine how a sample was selected before believing inferences drawn from sample results.



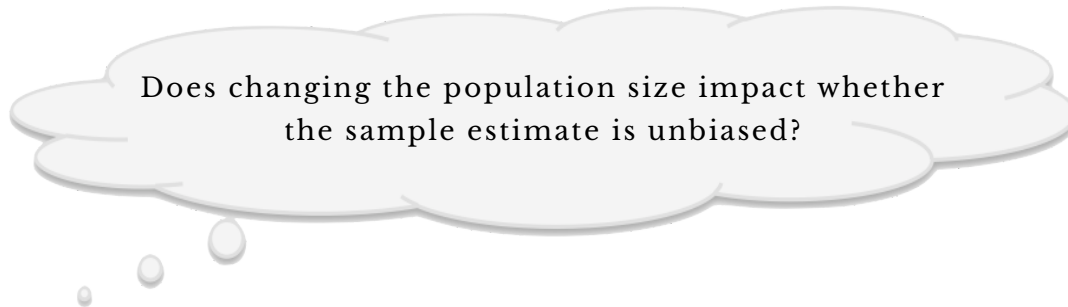
- Change the sample size from 10 to 25.
- Use TinkerPlots™ to draw 500 random samples of 25 words, and collect the average word length for each sample.

23. Sketch the plot of the sample estimates based on the 500 samples drawn. Make sure to label the axis appropriately.

24. Record the average value for the estimate of the average word length.
25. Does the sampling method still appear to be unbiased? Explain.
26. Compare and contrast the distribution of sample estimates for $n = 10$ and the distribution of sample estimates for $n = 25$. How are they the same? How are they different?
27. Using the evidence from your simulations, answer the research question:
Does changing the sample size impact whether the sample estimates are unbiased?

Population Size

It is clear that changing the size of the sample does not affect whether or not an unbiased estimate is produced. Now we examine another question:



To examine this we will now sample from a population that is quadruple the size of the original population (size = 1072) while keeping the population characteristics the same (e.g., mean word length is still 4.3 letters).

- Open the file *gettysburg-larger-population.tp*.
- Draw a simple random sample of ten words from the sampler.
- Compute the word length for each randomly sampled word.
- Plot and compute the mean word length for the ten randomly sampled words.
- Collect the mean word length for 500 random samples.

28. Sketch the plot of the sample estimates based on the 500 samples drawn. Make sure to label the axis appropriately.

29. Record the average value for the estimate of the average word length.
30. Does the sampling method still appear to be unbiased? Explain.
31. Compare and contrast the distribution of sample estimates for $n = 10$ now that you are sampling from a larger population to the distribution of sample estimates for $n = 10$ from before. How are they the same? How are they different?
32. Use the evidence collected from the simulation to answer the research question: Does changing the size of the population impact whether the sample estimates are unbiased?

A rather counterintuitive, but very crucial, fact is that when determining whether or not a sample estimate produced is unbiased **the size of the population does not matter!** Even more counterintuitive might be that the precision of the sample estimate is unaffected by the size of the population! (You will learn about the precision of a sample estimate in Unit 5.) This is why organizations like Gallup can state poll results about the entire country based on samples of just 1,000–2,000 respondents as long as those respondents are randomly selected.

In summary, it is important to note some caveats about random sampling:

- One still gets the occasional “unlucky” sample whose results are not close to the population even with large sample sizes.
- Second, the sample size means little if the sampling method is biased. As an example, in 1936 the *Literary Digest* magazine had a huge sample of 2.4 million people, yet their predictions for the Presidential election did not come close to the truth about the population.
- The size of the population does not affect the bias of the estimate, even if a small sample size is used.