

STAT 215: SAMPLING

The following is the full text of Abraham Lincoln's Gettysburg Address:

“Four score and seven years ago our fathers brought forth, on 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 battle-field 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 can not dedicate—we can not consecrate—we can not 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 here 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.”

Instructions.

- (1) Circle ten words that you think are representative of the speech (it is up to you what “representative” means).

For this activity we are considering this passage a **population** of words, and the 10 words you selected are considered a **sample** from this population. In most studies, we do not have access to the entire population and can only consider results for a sample from that population. The goal is to learn something about a very large population (e.g., all American adults, all American registered voters) by studying a sample. The key is in carefully selecting the sample so that the results in the sample are representative of the larger population (i.e., has the same characteristics).

- (7) The actual average of all 268 lengths is 4.29 letters. Would you say that the sampling method you used is biased? If so, in which direction? Explain how you can tell this from the class-wide dotplot of sample means.
- (8) Why do you think this sampling method turned out to be as biased as it did?
- (9) Consider a different sampling method: closing your eyes and point to the page ten times in order to select the words for your sample (don't actually do this, just think about doing it). Would this sampling method be unbiased? Explain.
- (10) Would using this same sampling method but with a larger sample size (say, 20 words) eliminate the sampling bias? If not, can you think of an alternative method that would?

A **simple random sample** (SRS) gives every case 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 the Gettysburg address example we want every *set of ten words* to be equally likely to be the sample selected.

Although it is easy to state what a simple random sample is, it is by no means simple to implement. The first step is to obtain a **sampling frame** where each member of the population can be assigned a number. Here we just need to number the words in the above passage. This sampling frame is attached.

- (11) Go to `random.org` (on a phone or tablet) and generate five random numbers between 1 and 268. Record the ID numbers that you selected, the corresponding words, and the lengths of the words:

	1	2	3	4	5
ID					
Word					
Length					

- (12) Determine the average length in your sample of five words.
- (13) Again combine your sample mean with those of your classmates to produce a dotplot.
- (14) Do the sample averages from the random samples tend to over- or underestimate the population average, or are they roughly split evenly on both sides?