Collecting Data: Sampling from a Population

Math 107, Fall 2016

The authorship of literary works is often a topic for debate. For example, researchers have tried to determine whether some of the works attributed to Shakespeare were actually written by Bacon or Marlow. The field of "literary computing" examines ways of numerically analyzing authors' works, looking at variables such as sentence length and rate of occurrence of specific words. In this exploration we will conduct a very basic text analysis.

1. The entire text of Lincoln's Gettysburg Address is given below. Select a representative set of 10 words from the Gettysburg Address by circling them with your pen or pencil.

   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.

2. A population consists all units of interest while a sample consists of all cases on which we have collected data. In this situation, what is the population and what is the sample?

3. Record each word from your sample, and then indicate the length of the word (number of letters) and whether or not the word contains at least on letter *e*.

	Word	Length (no. of letters)	Contains *e*? (Y or N)
1.			
2.			
3.			
4.			
5.			
6.			
7.			
8.			
9.			
10.			

The table you filled in above is your dataset.

4. Identify the cases and the variables you have recorded on those cases.

5. Is the variable “length of word” quantitative or categorical?

6. Calculate the average length of the 10 words in your sample.

7. A *parameter* is a number that describes some aspect of a population. A *statistic* is a number that is computed from the data in a sample. Is the average you found in part 6 a parameter or a statistic?

8. The average length of the 268 words in the entire speech is 4.29 letters. Is this number a parameter or a statistic?

9. Calculate the proportion of words in your sample that contain at least one *e*. Is this number a parameter or a statistic?

10. The proportion of all words that contain at least one *e* is $125/268 \approx 0.47$. Is this number a parameter or a statistic?

11. Do you think the words you selected are representative of the 268 words in this passage? Suggest a method for deciding whether you have a representative sample. (Hint: Whereas any one sample may not produce statistics that exactly equal the population parameters, what would we like to be true in general?)

12. Combine your results with your classmates' by recording your average word count on the google form: https://goo.gl/forms/XSuZcwLk1orthhiW2. We will create a histogram of the word lengths as a class.

13. Comment on the shape and location of the distribution of average word count.

14. Let's compare your sample statistics to the population parameters.

    (a) How many, and what proportion, of students in your class obtained a sample average word length larger than 4.29 letters?

    (b) How many, and what proportion, of students in your class obtained a sample proportion of *e*-words larger than 0.47?

A sampling method is biased if, when using the sampling method, statistics from different samples *consistently* overestimate or underestimate the population parameter of interest.

15. What do your answers to the previous question tell you about whether the sampling method used (i.e., asking students to quickly pick 10 representative words) is biased or unbiased? If biased, what is the direction of the bias (i.e., the tendency to overestimate or underestimate)?

16. Explain why we might have expected this sampling method to be biased.

17. Do you think asking each student to pick 20 words instead of 10 words would have helped with this issue? Explain.