Collecting Data: Sampling from a Population Part II

Math 107, Fall 2016

Overview

A sample is only useful to us if the data we collect on our sample are similar to the results we would find in the entire population. In this sense, we say the sample is **representative** of the population. The key to obtaining a representative sample is using some type of random mechanism to select the observational units from the population, rather than relying on convenience samples or any type of human judgement. Instead of choosing arbitrary words using your own judgement, in this part of the exploration you will take a **simple random sample** (a sample in which every sample of size n is equally likely to be chosen) of words and evaluate your results.

The first step in obtaining a simple random sample is to construct a **sampling frame**—a complete list of every member of the population where each member can be assigned an ID number. The file gettysburg.csv provides such a sampling frame, a complete list of the words in the Gettysburg Address.

Once the sampling frame is established, then we need to randomly pick n rows of the sampling frame in order to choose a simple random sample of size n. All of this can be done in R

1. Load the data set into R. Here this is the `gettysburg.csv` file.

```
gettysburg <- read.csv("data/gettysburg.csv", as.is = TRUE)
```

2. Load the `dplyr` package.

```
library(dplyr)
```

3. Use the `sample_n` function to draw a simple random sample of size n rows from a data set. Note that this will change every time you run the code since it will obtain a different simple random sample.

```
sample_n(gettysburg, size = 10)
```

```
##     word
## 80   for
## 63  that
## 251  and
## 175 work
## 130   it
## 87  that
## 50    so
## 64   war
## 173  the
## 33   are
```

On Your Own

1. Randomly draw a random sample of size $n = 10$ from the `gettysburg` data set. Record the words selected, the length of the word (i.e., the number of letters), and whether the word contains the letter "e" (yes or no). If you do not have a computer with you, ask a classmate if you can pair up with them.

2. Calculate the average length of the words in your sample.

3. Calculate the proportion of e-words in your random sample.

4. Fill out the google form: https://goo.gl/forms/BeTwHT6yhxss2Tsx1. This will create allow us to create a data set for the entire class.

As a class/group

5. We will create a histogram of the word lengths as a class. Comment on how this distribution compares to the one based on non-random sampling.

6. 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 the population average (4.29 letters)?

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

c. What do these answers reveal about whether simple random sampling is an unbiased sampling method? Explain.