D08 Biased and Unbiased Sampling

## Objective

The objective of this exploration is to understand the difference between a biased sample and an unbiased sample.

## Scenario #1: Gettysburg Address

List the words you selected from the Gettysburg Address:

consecrate, by, dedicate, met, nation, sense, great, score, which, who, unfinished

Based on your sample of the 10 words above, what would you estimate is the length of an average word in the Gettysburg Address? (Don’t do any calculations, just make a guess)

Answers will vary

Do you think that your sample is representative of all words in the Gettysburg Address?

Answers will vary

## Gettysburg - Your Sample

Using the words in your sample, let’s compute some numerical estimates.

my.sample <- c(10,2,9,3,6,5,5,5,5,3,10)  
favstats(~my.sample)

## min Q1 median Q3 max mean sd n missing  
## 2 4 5 7.5 10 5.727273 2.796101 11 0

containsE <- c(1,0,1,1,0,1,1,1,0,0,1)  
tally(~containsE, format = "percent")

## containsE  
## 0 1   
## 36.36364 63.63636

The typical word in my sample is 5.7 letters long. 63% of the words contain the letter E.

## Gettysburg - Random Sample

Now let’s take a **random** sample of the words in the Gettysburg Address and compute some numerical estimates.

GA.words <- read.csv("gettysburgWords.csv", header=T, sep = " ")  
  
# Let's select a random sample of 10 words  
set.seed(10)  
random.sample <- sample\_n(GA.words, 10)  
str(random.sample)

## 'data.frame': 10 obs. of 4 variables:  
## $ Place : int 137 72 211 143 24 13 110 231 155 92  
## $ Word : chr "add" "of" "devotion" "little" ...  
## $ Letter : int 3 2 8 6 11 9 8 4 5 2  
## $ containsE: logi FALSE FALSE TRUE TRUE FALSE TRUE ...

favstats(~Letter, data=random.sample)

## min Q1 median Q3 max mean sd n missing  
## 2 3.25 5.5 8 11 5.8 3.119829 10 0

tally (~containsE, data=random.sample, format = "percent" )

## containsE  
## TRUE FALSE   
## 50 50

The typical word in the random sample is 5.8 letters. 50% of them contain the letter E.

Identify each of the following terms for the above scenario :

What is the population of interest?  
> All the words in the Gettysburg Address

What is the sample? > 10 randomly selected words

What are the two parameters of interest? > Length of a word in the GA and whether the word contains the letter E

favstats(~Letter, data=GA.words)

## min Q1 median Q3 max mean sd n missing  
## 1 3 4 5 11 4.29368 2.119384 269 0

tally (~containsE, data=GA.words, format = "percent" )

## containsE  
## TRUE FALSE   
## 46.84015 53.15985

What are the two statistics? > Length of a word in my sample (5.8 letters) and whether the word contains the letter E (50%)

Explain how the sample you chose (your 10 words) may have been biased. > My sample may have been biased because I chose lots of words that were longer (like consecrated or dedicated) - they are easier to see.

## Scenario 2: English Language

OK, now let’s think about the more general case. How long do you think a typical word is in the **English Language**?

Answers will vary

Now we have a different population of **words in the English Language**. Let’s use the Gettysburg Address as a sample of size 268.

favstats(~Letter, data=GA.words)

## min Q1 median Q3 max mean sd n missing  
## 1 3 4 5 11 4.29368 2.119384 269 0

tally (~containsE, data=GA.words, format = "percent" )

## containsE  
## TRUE FALSE   
## 46.84015 53.15985

We estimate the length of a typical word **in the English Language** at 4.29 letters (with an average deviation of 2.12 letters). We estimate that 53% of the words have the letter E.

What is the population of interest (be sure to use the word “all”)?  
> All words in the English language

What is the sample? > The 268 words in the Gettysburg Address

What are the two parameters of interest? > Length of a word in the English Language and whether the word contains the letter E

What are the two statistics? > Length of a word in the GA and whether the word contains the letter E

Explain at least one way in which this sample might be biased.

Answers will vary