MA256 Lesson 5 - Generalization (2.1, 2.2)

Warm up

Match the following symbols with their definitions: $\pi \quad \mu \quad \hat{p} \quad \bar{x} \quad s \quad \sigma$

1. Which symbol represents the *population's* parameter for a categorical variable?
2. Which represents the *sample's* average for a quantitative variable?
3. Which represents the *population's* standard deviation for a quantitative variable?
4. Which represents the *sample's* proportion for a categorical variable?
5. Which represents the *sample's* standard deviation for a quantitative variable?
6. What does the last remaining symbol represent?

Definitions

- **generalization**
- **population**
- **sample**
- **convenience sample**
- **biased sampling method**
- **simple random sample**

Questions

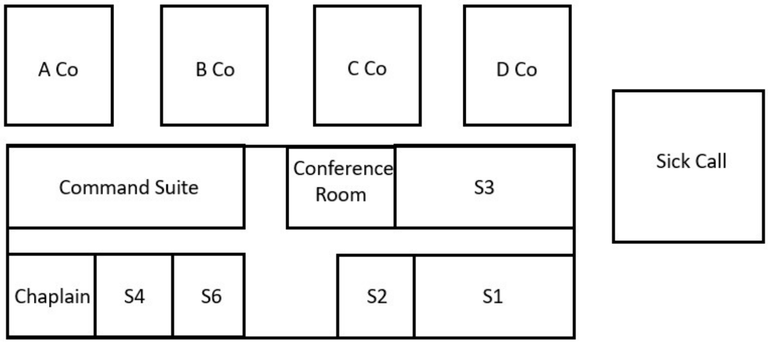
Q1) Does a larger sample size fix sampling bias?

Q2) What are some Non-sampling concerns which might also introduce bias into our data?

Q3) How do you calculate the standard deviation for a sample proportion (from a large population)? The standard deviation of a sample mean (for a large sample)?

Scenario:

You've been assigned to the Battalion S3 shop as the AS3. The BC comes in one day and asks you to give him a snapshot of the health of the soldiers of the Battalion. Here is a schematic of your Battalion area:



You decide to sample the 20 Soldiers in your immediate vicinity and discover none of them are sick.

Q4) What is the population?

Q5) What is the sample?

Q6) What kind of sample did you just take?

Q7) Is this type of sample biased or unbiased?

Realizing that perhaps your sample is too small for the entire battalion, you walk over to the S1 office and ask 10 more of those Soldiers to get a more representative sample. None of them report being sick.

Q8) Will asking the Soldiers in the S1 shop (10 more not sick) fix the bias?

Luckily, the CSM stopped you before you went in with the BC. After his blood pressure came back down he told you that he knows for a fact that some Soldiers went to sick call this morning, so you decide to call over to the aid station to expand your sample. You discover that 10 Soldiers from your unit are currently there. Given the 20 healthy S3 Soldiers, 10 healthy S1 Soldiers, and these 10 sick soldiers, you determine your sample proportion is 25% sick.

Q9) Is this new proportion representative of the population?

This time the XO stops you. He suggests that you randomize your sample and try again. You use R to pick 20 random numbers that identify Soldiers on the recall roster. By texting each one to ask if they are sick or not, you learn 2 out of the 20 went to sick call.

Q10) What is your new unbiased and representative proportion and what symbol would you use to represent it?

Q11) Knowing this won't exactly match the population proportion, how do we get a closer estimate?

Gettysburg Address

(Note: Reference Exploration 2.1 (p. 126) and 2.2 (p. 138))

1. Select a representative set of 10 words from the Gettysburg Address passage by circling them.

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 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.

2. 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 one letter "e".

	Word	**Length of Word (number of letters)**	**Contains *e*? (Y or N)**
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

3. Identify the observational units and the variables you have recorded on those observational units. (Keep in mind that observational units do not have to be people!)

4. Is the variable "length of word" quantitative or categorical?

5. Calculate the average length of the 10 words in your sample and write it below. Is this number a parameter or a statistic?

6. The average length of the 268 words in the entire speech equals 4.295 letters. Is this number a parameter or a statistic?

7. Calculate the proportion of words in your sample that contain at least one "e" and write it below. Is this number a parameter or a statistic?

8. The proportion of all words in the entire speech that contain at least one letter "e" is $\frac{125}{268} \approx 0.466$. Is this number a parameter or a statistic?

9. Compare the sample statistics with the population parameters.

a) How many and what proportion of cadets obtained a sample average word length larger than the population parameter?

b) How many and what proportion of cadets obtained a sample proportion of "e" words larger than the population parameter?

10. Do our results indicate that our sampling method is biased? Explain.

11. Do you think that closing our eyes and pointing blindly at a page with a pencil 10 times would result in a biased sample? Explain.

12. Do you think that selecting 20 words instead of 10 would remove any bias in the previous two methods? Explain.

13. What is another technique for selecting 10 words from the population that might be a better representation of the population with regards to word length and "e" words?

The following code will read in the Gettysburg address, remove commas and periods, calculate the length of each word, and indicate whether or not a word contains the letter “e”. The final result is a dataframe called `GW`.

```
library(tidyverse)

gw <- scan('http://www.isi-stats.com/isi/data/chap3/GettysburgAddress.txt', character(), quote ="")

gw <- gw %>% gsub(",", "", x=. ) %>% gsub("\\.", "", x=.)
GW <- data.frame(words = gw) %>%
  mutate(wordlen = str_length(words),
         cont.e = grepl("e", words))
```

14. In R, use sample(1:268, 10) to randomly sample 10 numbers between 1 and 268 (without replacement).

```
# set.seed(256)
#
#
```

15. Calculate the average length of the 10 words in your new sample and write it below. How does this compare to your previous sample?

```
#
```

16. Calculate the proportion of words in your new sample that contain at least one “e” and write it below. How does this compare to your previous sample?

```
#
```

17. Compare the sample statistics with the population parameters.

a) How many and what proportion of cadets obtained a sample average word length larger than the population parameter?

b) How many and what proportion of cadets obtained a sample proportion of “e” words larger than 0.466?

18. Did our new simple random sample show evidence that it is a biased sampling method?

19. Do we meet the validity conditions for estimating strength of evidence? What are the implications?