**Lab Assignment 1 SOLUTIONS**

***20 points***

**Part I: Basics of R**

1. ***(1 point)***  Were you able to install R and RStudio onto your personal device (desktop, laptop)?

***Everyone gets 1 point for this question!!***

*If you answered no to the above question, you may still be able to use R some other way. But, it’ll be much easier for you if you can install R and RStudio onto your own computer. Talk with one of the TAs or the instructor before your next assignment is due for help with installing R and RStudio onto your computer.*

2. ***(2 points)*** State where each of the following window panes are located in RStudio (upper left, lower left, upper right, or lower right and a brief explanation of what each window pane is used for and/or what appears in the window pane.

a. Console window pane ***lower left – commands are executed here***

b. Plot window pane ***lower right – plots will be displayed here***

c. Script window pane ***upper left – place to write commands***

d. Environment window pane ***upper right – shows list of variables created***

***½ point for each part (1/4 point for correct location and ¼ point for brief description – description doesn’t have to be too detailed as long as student has idea)***

**Part II: Self-sampling versus random sampling (from the Lab 1 Notes)**

The entire text of Lincoln’s Gettysburg Address is given below. Your task is to select a *sample of 10 words* to estimate the average length of words in this speech.

***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 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 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.”

3. ***(1 point)*** Select 10 words from the Gettysburg Address on the previous page. Write the words and their lengths (number of letters) for each of the 10 words you selected. Remember, the idea is that you want to estimate the average length of words for the entire Gettysburg Address based on your sample of 10 words.

***½ point for a list of 10 words chosen and ½ point for list of lengths of the 10 words.***

4. ***(1 point)*** In the Script window pane in RStudio, write the script to record and store the length of your 10 words into a vector. Copy and paste that script for your answer to Question 4. (*See page 4 in Lab 1 Notes for an example of writing a script.)*

***Should be something like this: mysample <- c(1,2,3,4,5,6,7,8,9,10)***

* ***If student just has c(1,2,3,4,5,6,7,8,9,10), that’s fine***

5. In the next line in the Script window pane in RStudio, write the script to calculate the average of the lengths of the 10 words in your sample. (That is, write the script to calculate the average of the vector you created in Question 4). (*Hint: see the middle of page 4 in the Lab 1 Notes.)*

a. ***(1 point)*** Copy and paste the script for your answer to Question 5a.

***mean(mysample) or mean(c(1,2,3,4,5,6,7,8,9,10))***

b. ***(1 point)*** Run both lines of the script you created in Questions 4 and 5a. The average (or “mean”) will be displayed in the Console window. What is the average length of words for your sample of 10 words? (Do not round.)

***answers will vary. Reasonable answers are somewhere between 3 and 6, although it could be anywhere from 2 (if student chose all two-letter words) to close to 11 (which is the maximum length of words)***

c. ***(1 point)*** Some students in the class may calculate the same average as yours, but most will have a different average. Explain why the averages from all of the samples are not the same.

***Everyone took a different sample of 10 words and, therefore, will have a different mean.***

The goal of the rest of this assignment is to determine if “self-selecting” a sample is “representative” of a population. For this problem, the population is all words in the Gettysburg Address. The mean length of *all* words in the Gettysburg Address is 4.24 letters.

6. ***(1 point)*** Calculate the difference between your sample mean (i.e. the mean of the 10 words you “self-selected” from Question 5b) and the population mean (i.e. the mean length of *all* words in the Gettysburg Address). Report that difference for your answer to Question 6. (Your answer could be positive or negative depending on how you subtracted. Either is acceptable for this answer.)

* *Note: you can use RStudio to do this calculation – try it to see if you get the same answer as when you do the calculation manually or from a calculator*

***Answers will vary. It should be their answer to question 5b – 4.24 (or 4.24 – 5b answer). No work needs to be shown.***

One way to evaluate whether a sample is representative of the population is to compare the sample mean and population mean – if the sample is representative of the population, the sample mean should be equal to or very close to the population mean.

7. ***(2 points)*** Based solely on the difference between your sample mean and the population mean, do you feel that self-selecting provides a representative sample of the population? Briefly explain why or why not. (Again, you are basing the answer to this question on the difference you calculated in Question 6.)

***Answers can vary. Look for reasonable support, which should be seeing how close their answer to question 6 is to 0. Some students may have a sample mean from their self-selected words close to the actual population mean of 4.24 and, therefore, may have reasonable support to say that self-selecting is representative of the population. Some will have sample means quite a bit different than 4.24, which could be support saying that self-selecting is not representative of the population.***

***Support must include comparing their difference to 0, but students may add additional support. For example, they may say that even though their sample mean was close to 4.24, which may indicate that the self-selection sample is representative of the population, they don’t feel that it really is because other self-selected samples may give a sample mean far from 4.24.***

***If student did not have any support involving their answer to question 6, the maximum credit for this problem is 1.5 points.***

8. ***(1 point)*** Use R to find the average of the variable ***mean*** using the following code in R (type this code in the Script Window Pane and run the script. The mean will appear in the Console Window Pane.) Note that you did this on page 7 in the Lab 1 Notes.

* + **mean(gettyself$mean)**

Report the mean of all the sample means. Round to **four** decimal places.

***5.5203 letters (do not take off if unit is not included)***

9. ***(1 point)*** Calculate and report the difference between the mean of all the sample means (answer to Question 8) and the population mean. Do not round your answer.

***1.28132 is the unrounded answer using full code (mean(gettyself$mean) - mean(gettyaddress$length).***

***Using the unrounded answer to question 8 – 4.24 = 1.2803 letters.***

10. ***(2 points)*** Let’s return to the question posed in Question 7: do you feel that self-selecting provides a representative sample of the population? This time, support why or why not with your answer to **Question 9**. (Note: your answer and reasoning may or may not be the same as your answer and reasoning for Question 7, depending on what the mean of your self-selected sample was.)

***We might have accepted a number of different answers in question 7, but not so much with this answer. The difference of 1.28 letters between the mean of all sample means and the population mean should be considered a “large” difference, which would indicate that generally, self-selecting does not provide a good representation of the population. Again, support for this question must be based on the difference calculated in #9.***

***If student did not use the difference in #9 as support for their answer, the maximum points that can be earned is 1 point on this problem. Students may add additional support.***

***If a student uses their answer to question 9 in support of saying self-selecting is representative of a population, maximum credit should be 1.5 points.***

11. ***(1 point)*** What is the mean of the 10,000 random sample means from your simulation?

***Answers may vary slightly, but should be fairly close to what I got with my simulation: 4.24969 letters.***

12. ***(1 point)*** Calculate and report the difference between the mean of your random sample means (answer to Question 11) and the population mean. Do not round.

***Again, answers may vary slightly, but should be in the neighborhood of 4.24969 – 4.24 = 0.969 letters (i.e. their difference should be close to 0).***

13. ***(3 points)*** Which type of sampling (self-selection or random) provides a better representation of a population? Why? (Your support should be based solely on the answers to the questions above. That is, think about how to use your answers to Questions 9 and 12 to support the answer to this question.)

***While answers may vary, support must be with using the differences calculated above. The most reasonable answer is to say that random sampling is more representative of the population because the tendency is for the mean from a random sample to be closer to the population mean than the mean from a self-selected sample. Additional support can be given, but the comparison of differences in some manner must be included in their support.***

***If students support answer without using differences, maximum points that can be given is 1.5 points.***