**DATA 557**

**Winter 2019**

**Homework Assignment 1**

**Instructions**

Submit your solutions **in pdf format** to the dropbox on the canvas page by **5:00PM, Wednesday January 16**. Include your R code with your solutions.

You may work together to help each other solve problems, but you should do all the work, create your own solutions, and hand in your own work without copying others’ work.

Data set: ‘iq.csv’ (see Week 1 folder on canvas)

The data come from a study to examine the effect of lead exposure on IQ in children. The sample was randomly selected from the children in a large community (the “population”) near a source of lead. IQ tests were administered to each child. The IQ scores were age-standardized using established normal values for the US population. Such age-standardized scores have a mean of 100 and a standard deviation of 15 in the US population.

1. Create a histogram and normal q-q plot of the IQ variable. Using these plots comment on how well the distribution is approximated by a normal distribution.

2. Calculate the sample mean and sample SD of IQ. Give two different possible explanations for why these values differ from the corresponding values for the US population.

3. Calculate a 95% confidence interval for the mean IQ score. For calculating the SE use the value 15 for the population SD of IQ. Is 100 in the confidence interval? Explain why it matters whether or not 100 is in the confidence interval.

4. Suppose that you were planning to repeat the study and that you wanted the 95% confidence interval for mean IQ to have width 30 or less. Assume that the distribution of IQ in the population is normal with mean 100 and standard deviation 15. What is the minimal sample size that would be needed for the new study?

5. Calculate the confidence interval for mean IQ using the sample SD instead of the population SD value of 15. In what way does the confidence interval differ from the one obtained previously? Explain the difference.

6. Perform a simulation study to estimate the coverage probability of the 95% confidence interval for mean IQ that uses the SD value of 15 to calculate the SE (as in Question 3). For your simulation, assume the distribution of IQ in the community has a normal distribution with mean 100 and standard deviation 15. Justify your choice of the number of replications (simulated samples). Do your results provide evidence that the coverage probability for this confidence interval is different than 0.95?

7. Perform a simulation study to estimate the coverage probability of the 95% confidence interval for mean IQ that uses the sample SD to calculate the SE (as in Question 5). (Note: save the sample means and their SEs for answering the following questions.) As previously, assume the distribution of IQ in the community has a normal distribution with mean 100 and standard deviation 15 to generate your simulated samples. How do the results compare with those of the previous simulation?

8. Display the distribution of the simulated sample means and describe how well it is approximated by a normal distribution. Is this what you expected? Why or why not?

9. Display the distribution of the estimated SEs. Compare the average value of the SEs to the SD of the sample means. Are they approximately equal? Explain why this matters.

10. Now assume that the distribution of IQ in the community is a normal distribution with mean 90 and SD 15. Perform a simulation study of the 95% confidence interval for mean IQ (the version from Question 5). Calculate the proportion of the confidence intervals that are entirely below the value 100 (i.e., the upper limit of the confidence interval is lower than 100). Give an interpretation of the result.

11. For the previous question, estimate the minimum sample size required in order that the proportion of 95% confidence intervals lying entirely below the value 100 is at least 0.9. (There is no exact answer.)

12. Provide an estimate and a 95% confidence interval for the proportion of children in the community with IQ score less than 100. What does the result tell you about differences between the distribution of IQ in the community and in the US population?

13. Perform a simulation study to assess the performance of the 95% confidence interval for the proportion of children in the community with IQ less than 100. For the simulation, assume that the true population proportion is equal to 0.5. Does the confidence interval have approximate 95% coverage? (Note: save the sample proportions and their estimated SEs for answering the following questions.)

14. Display the distribution of the simulated sample proportions and describe how well it is approximated by a normal distribution. Is this what you expected? Why or why not?

15. Display the distribution of the estimated SEs of the sample proportions. Compare the average value of the estimated SEs to the SD of the sample proportions. Are they approximately equal? Explain why this matters.

16. Using simulation, determine an approximate minimal sample size that yields a valid 95% confidence interval for the proportion of children with IQ less than 100 if the true population proportion is equal to 0.5. (Note: the key word here is “approximate”; there is no precise answer to this question).

17. Using the formula for the SE of a sample proportion, calculate the minimal sample size that would yield a 95% confidence interval for the sample proportion of width 0.1 or less. The confidence interval should have length 0.1 or less for **any** value of the population proportion.