STATS 10 Assignment 5

Please submit both parts of the assignment in one single PDF file. You can use any PDF editor software to merge the two parts into one file. Please make sure that the questions are in the correct order and clearly labeled, and that the answers are legible and easy to read.
To submit your assignment, upload the PDF file under the designated assignment page on the course website before the deadline specified. **Email or hard copy submissions are not accepted.**

Part I

Include both the R commands and their corresponding outputs, results, or answers for all exercise questions in Part I.

Exercise 1

The Sweetums candy factory in Pawnee, Indiana, is under investigation for violating EPA regulations. Factory workers have improperly disposed of arsenic and sulfur waste from the candy-making process, and the contamination has reached the local water supply! We have data for arsenic and sulfur levels from the water in all houses within a 2-mile radius of the factory. Download the "pawnee.csv" file from the course website, then read it into RStudio with the following line:

```
pawnee <- read.csv("yourpath/pawnee.csv", header = TRUE)
```

Some important variables include:
• Arsenic: arsenic levels for each home in ppm
• Sulfur: sulfur levels for each home in ppm
• New_hlth_issue: Indicates "Y" if someone living at the home has experienced a major health issue after the date of contamination, "N" if no new health complications.

a. Use the head() function to print out the first few rows of this data. Then, use the dim() function to print out the number of rows and columns of this data frame.

b. Set the seed to 1337 and take a simple random sample of size 30 from the entire pawnee data frame. Save the random sample as a separate R object, and print the first few lines to make sure you saved it correctly.

c. Report the proportion of households experiencing a major health issue from your sample. Also report the population proportion of all households which experienced a new major health issue.

d. Generate confidence intervals for our sample proportion using the sample results. Produce 90%, 95%, and 99% confidence intervals for the true population proportion. Consult your lecture materials if you are unsure how to do this. You can use R and/or a calculator for this question, but please include code or calculations to show your work.

Exercise 2 – Hypothesis testing with one proportion.

We will be working with a modified Flint dataset, which can be found on the course website. Please download the file and read it into R. You may recall that lead levels were considered dangerous if the result was greater than or equal to 15PPB. We are interested in determining if the proportion of dangerous lead levels in Flint is greater than 10%. Assume the Flint data is a random sample used to address this research question.

a. We will conduct a hypothesis test for this research question. What are the null and alternative hypotheses? Is this a one-sided or a two-sided test?

b. Calculate the sample proportion and sample standard deviation of the sample proportion of dangerous lead levels.

c. Now, calculate the SE of sample proportions, and the z-value for this test. Consult the above instructions and/or the lecture materials for guidance.

d. Using the z-statistic in (c), calculate the p-value associated with this test. You may use R's pnorm() function or a normal table, but please show all work.

e. Using a significance level of 0.05, do you reject the null hypothesis?

f. If greater than 10% of households in Flint contain dangerous lead levels, the EPA requires remediation action to be taken. Based on your results, what should you tell the EPA?

Part II

You may choose to type or write your answers electronically or scan your handwritten solutions. Please ensure that you show all steps and explanations to receive full credit, unless otherwise instructed.

Exercise 1

Research done in 2013 found that only 48% of all the site users reported getting their news about world events on this site. A 2018 poll of 3625 randomly selected users of a social media site found that 1830 get most of their news about world events on the site.

a. Does this sample give evidence that the proportion of site users who get their world news on this site has changed since 2013? Carry out a hypothesis test and use a 0.05 significance level.

b. After conducting the hypothesis test, a further question one might ask is what proportion of all of the site users get most of their news about world events on the site in 2018. Use the sample data to construct a 95% confidence interval for the population proportion. How does your confidence interval support your hypothesis test conclusion?

Exercise 2

According to the Brookings Institution, 50% of eligible 18- to 29-year-old voters voted in the 2016 election. Suppose we were interested in whether the proportion of voters in this age group who voted in the 2018 election was higher. Describe the two types of errors one might make in conducting this hypothesis test.