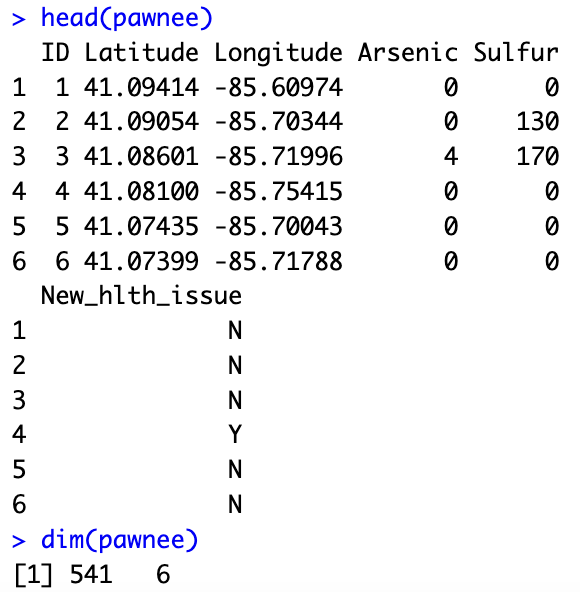
Preyasi Gaur  
705704939

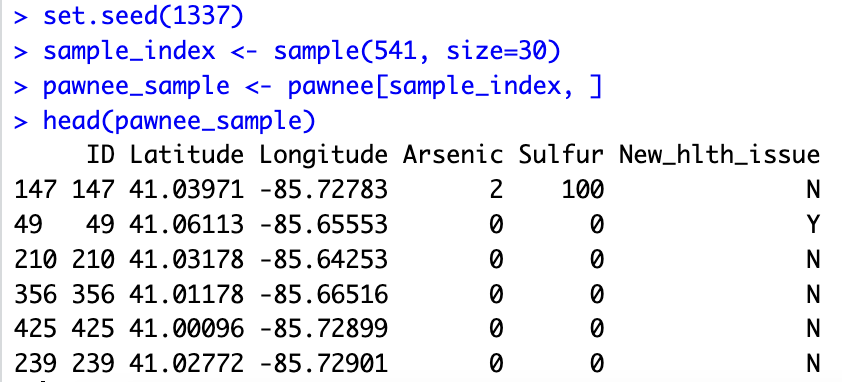
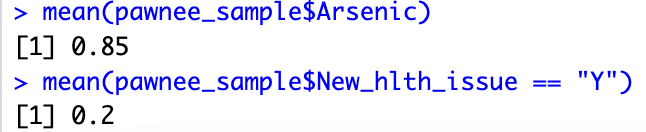
Discussion: 4A

Stats 10

Lab 5

1. a. Question:  
   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.  
     
   Input code:   
   pawnee <- read.csv("/Users/preyasigaur/Desktop/pawnee.csv")  
   head(pawnee)  
   dim(pawnee)  
   Output:   
   

b. Question: 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.  
  
Input code:   
> set.seed(1337)  
> sample\_index <- sample(541, size=30)  
> pawnee\_sample <- pawnee[sample\_index, ]  
> head(pawnee\_sample)

Output:  
  
  
c. Question:  
Report the mean arsenic level from the sample you took in b. Also report the proportion of households experiencing a major health issue from your sample.  
  
Input Code:  
mean(pawnee\_sample$Arsenic)  
mean(pawnee\_sample$New\_hlth\_issue == "Y")  
  
Output:  
  
  
d. Question:   
What symbol from the lecture would we use for the mean arsenic level in the sample? What symbol would we use for the proportion of health issues in the sample?  
  
Answer:  
(x-bar) Mean Arsenic Level: x̄  
(p-hat) Proportion of Health Issues: p̂  
  
e. Question:   
Now, let’s 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 slides 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  
  
Input Code:   
> p\_hat = mean(pawnee\_sample$New\_hlth\_issue == "Y")  
> se = sqrt(p\_hat \* (1-p\_hat)/30)

> #90% confidence interval

> z\_90 = qnorm(0.95)

> p\_hat + c(-1,1) \* z\_90 \* se  
 > #95% confidence interval

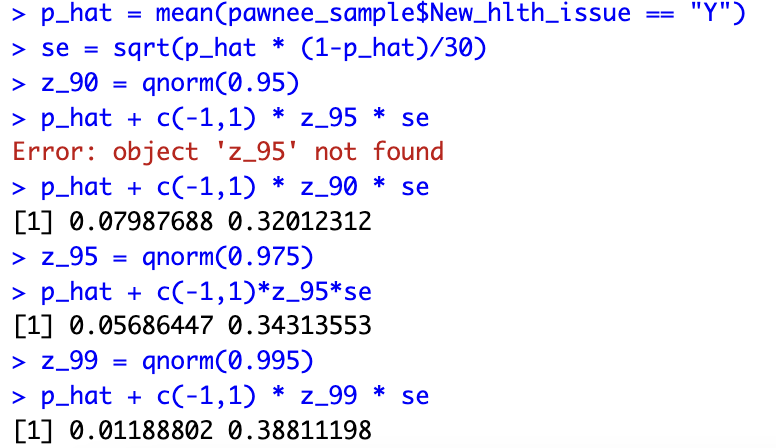
> z\_95 = qnorm(0.975)

> p\_hat + c(-1,1)\*z\_95\*se

> #99% confidence interval

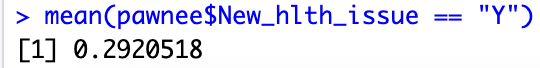
> z\_99 = qnorm(0.995)

> p\_hat + c(-1,1) \* z\_99 \* se

Output:  
   
  
f. Question:  
What would be the bounds of a 100% confidence interval for the population proportion?

Answer:

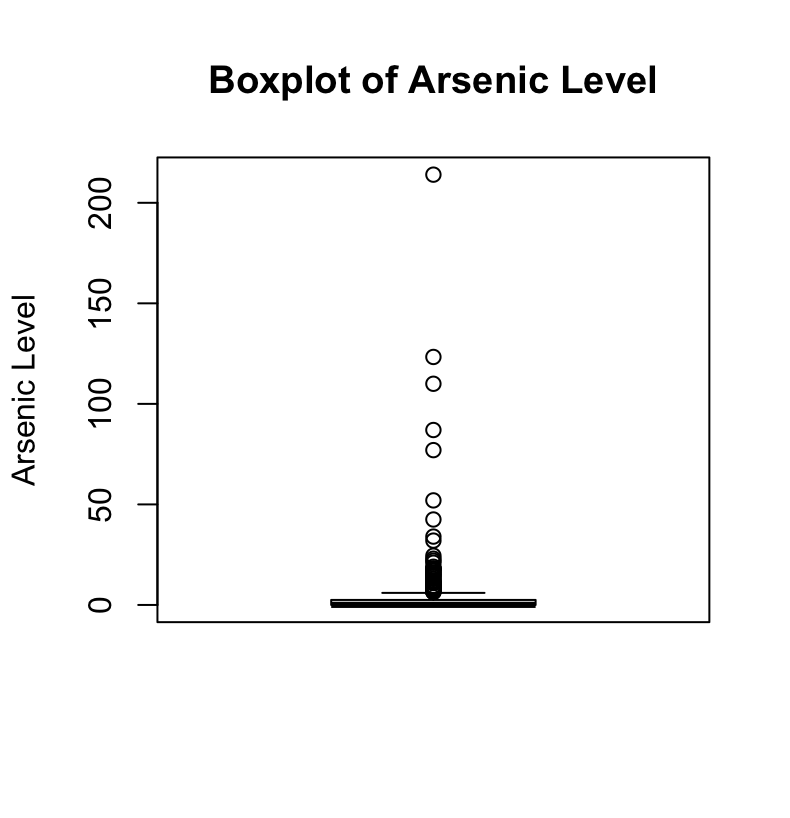
Since α = 0, the population proportion with a 100% confidence interval would have the bound [0,1], both inclusive   
  
g. Question:

Report the proportion of all households which experienced a new major health issue  
  
Input Code:   
> mean(pawnee$New\_hlth\_issue == "Y")  
  
Output:  


h. Question:  
Create a plot that visualizes the distribution of arsenic levels for the houses in Pawnee.

Hint: we can consider the arsenic levels continuous data

Input Code:  
> boxplot(pawnee$Arsenic, ylab = "Arsenic Level", main = "Boxplot of Arsenic Level")

Output:  


1. a. Question:  
   We will conduct a hypothesis test for this research question. Using symbols from lecture, what are the null and alternative hypotheses? Is this a one-sided or a two-

sided test?  
  
Input Code:

> flint <- read.csv("/Users/preyasigaur/Desktop/flint\_2015.csv", header = TRUE)

Answer:

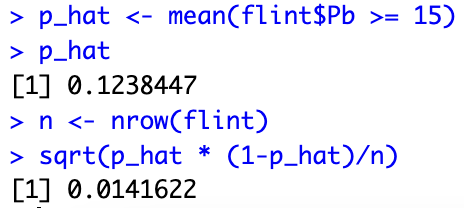
Since, we have a greater than/less than sign, the hypothesis is one-sided

b. Question:   
Calculate the sample proportion and sample standard deviation of the sample proportion of dangerous lead levels.  
  
Input Code:   
> p\_hat <- mean(flint$Pb >= 15)

> p\_hat

> n <- nrow(flint)

> sqrt(p\_hat \* (1-p\_hat)/n)

Output:  


c. Question:

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

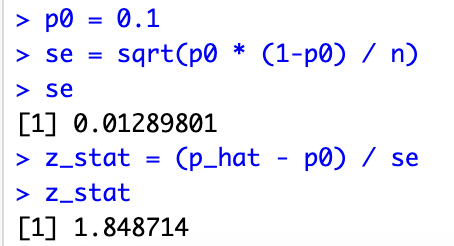
Input Code:   
> p0 = 0.1

> se = sqrt(p0 \* (1-p0) / n)

> se

> z\_stat = (p\_hat - p0) / se

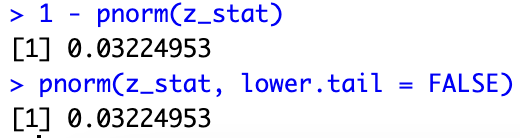
> z\_stat

Output:   


d. Question:   
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.

Input Code:   
> 1 - pnorm(z\_stat)

> pnorm(z\_stat, lower.tail = FALSE)

Output Code:  


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

Yes, we can reject the null hypothesis as the p-value is smaller than the significance level (0.05)

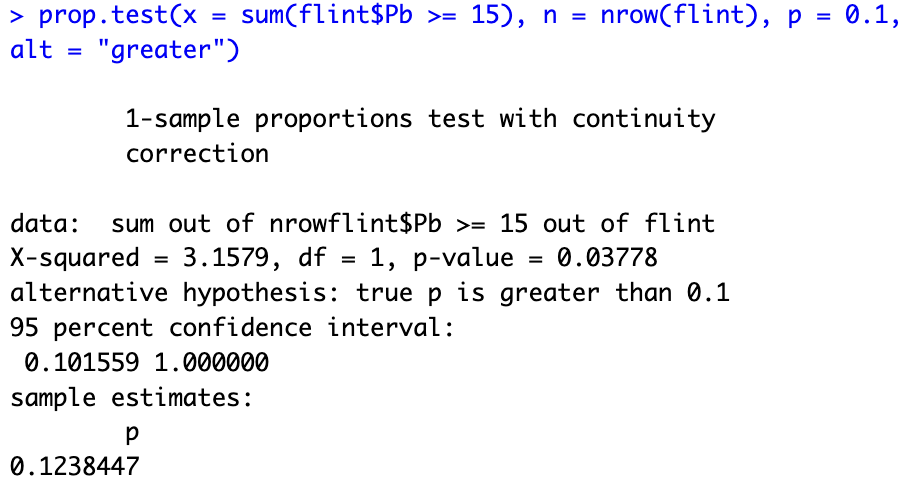
f. Question:  
If greater than 10% of households in Flint contain dangerous lead levels, the EPA requires remediation action to be taken. Based on your results, what should you tell the EPA?  
  
Answer:  
The EPA should take remediation action as soon as possible as greater than 10% of the households in Flint contain dangerous lead levels. As we reject the null hypothesis, we are in favor of the alternate hypothesis.

g. Question:  
Another way to run this test is to use the prop.test() function using the mosaic package. You will need to know your sample size, and the number of “successes'' in the sample. Use this function to conduct the same hypothesis test in (a) - (d) and obtain a p-value from the test. Using the same significance level of 0.05, do your results change? An example of the prop.test() function is shown in the two lines below:

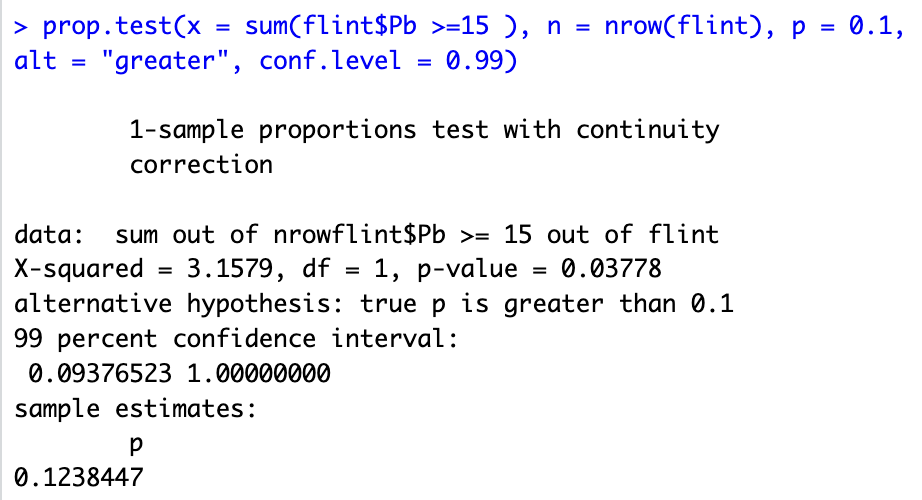
## We flipped 100 coins and 60 were heads. Is the long-run proportion of heads greater than 0.5? prop.test(x = 60, n = 100, p = 0.5, alt = "greater")

Input Code:   
> library(mosaic)

> prop.test(x = sum(flint$Pb >= 15), n = nrow(flint), p = 0.1, alt = "greater")

Output:   


h. Question:  
Notice that the prop.test() output produced a confidence interval. Try using the help screen under the mosaic package for prop.test() to find the argument for the confidence interval. Modify the argument and re-run the code in (g) to produce a 99% confidence interval instead of a 95% interval.  
  
Input Code:  
> prop.test(x = sum(flint$Pb >=15 ), n = nrow(flint), p = 0.1, alt = "greater", conf.level = 0.99)

Output:   


The 99% confidence level interval is (0.09376523, 1.00000000)