STATS 10 ASSIGNMENT 5

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.

Code:

```
#exercise 1
pawnee <- read.csv("pawnee.csv", header = TRUE)
#1a
#Use the head() function to print out the first few rows of this data.
head(pawnee)
#Then, use the dim()function to print out the number of rows and columns
#of this data frame
dim(pawnee)
```

Output:

```
> #1a
> #Use the head() function to print out the first few rows of this data.
> head(pawnee)
  ID Latitude Longitude Arsenic Sulfur New_hlth_issue
1  1 41.09414 -85.60974       0      0              N
2  2 41.09054 -85.70344       0    130              N
3  3 41.08601 -85.71996       4    170              N
4  4 41.08100 -85.75415       0      0              Y
5  5 41.07435 -85.70043       0      0              N
6  6 41.07399 -85.71788       0      0              N
> #Then, use the dim()function to print out the number of rows and columns
of this data frame
> dim(pawnee)
[1] 541   6
```

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.

Code:

```
#1b
set.seed(1337)
n<- dim(pawnee)[1]
idx<- sample(n,30, replace= FALSE)
subset<- pawnee[idx,]
head(subset)
```

Output:

```
> set.seed(1337)
> n<- dim(pawnee)[1]
> idx<- sample(n,30, replace= FALSE)
> subset<- pawnee[idx,]
> head(subset)
     ID Latitude Longitude Arsenic Sulfur New_hlth_issue
147 147 41.03971 -85.72783       2    100              N
49   49 41.06113 -85.65553       0      0              Y
210 210 41.03178 -85.64253       0      0              N
356 356 41.01178 -85.66516       0      0              N
425 425 41.00096 -85.72899       0      0              N
239 239 41.02772 -85.72901       0      0              N
```

Values	
idx	int [1:30] 147 49 210 356 425 239 126 35…
n	541L

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.

Code:

```
#1c
#proportion of households experiencing a major health issue
p_hat<- mean(subset$New_hlth_issue =="Y")
p_hat
#report the population proportion of all households which experienced a new
#major health issue
p<-mean(pawnee$New_hlth_issue=="Y")
p
```

Output:

```
> #1c
> #proportion of households experiencing a major health issue
> p_hat<- mean(subset$New_hlth_issue =="Y")
> p_hat
[1] 0.2
> #report the population proportion of all households which experienced a
new
> #major health issue
```

```
> p<-mean(pawnee$New_hlth_issue=="Y")
> p
[1] 0.2920518
```

Hence the proportion of households experiencing a major health issue = 0.2 and the population proportion of all households which experienced a new major health issue = 0.292.

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.

Code:

```
##1d
#Generate confidence intervals for our sample proportion using the sample results.

##confidence interval formula: [p_hat-(z*)*SE, p_hat+(z*)*SE]
###z*is the critical value which depends on the confidence level
###SE:standard error
SE<- sqrt(p_hat * (1-p_hat)/30)
###to find the z* using qnorm()

## this is for 90% confidence interval
z_star_90<- qnorm(0.95)
c(p_hat-z_star_90*SE, p_hat + z_star_90*SE)

## this is for 95% confidence interval
z_star_95<- qnorm(0.975)
c(p_hat-z_star_95*SE, p_hat + z_star_95*SE)

## this is for 99% confidence interval
z_star_99<- qnorm(0.995)
c(p_hat-z_star_99*SE, p_hat + z_star_99*SE)
```

Output:

```
##1d
> #Generate confidence intervals for our sample proportion using the sampl
e results.
>
> ##confidence interval formula: [p_hat-(z*)*SE, p_hat+(z*)*SE]
> ###z*is the critical value which depends on the confidence level
> ###SE:standard error
> SE<- sqrt(p_hat * (1-p_hat)/30)
> ###to find the z* using qnorm()
>
> ## this is for 90% confidence interval
> z_star_90<- qnorm(0.95)
> c(p_hat-z_star_90*SE, p_hat + z_star_90*SE)
[1] 0.07987688 0.32012312
>
> ## this is for 95% confidence interval
> z_star_95<- qnorm(0.975)
> c(p_hat-z_star_95*SE, p_hat + z_star_95*SE)
[1] 0.05686447 0.34313553
>
> ## this is for 99% confidence interval
> z_star_99<- qnorm(0.995)
```

```
> c(p_hat-z_star_99*SE, p_hat + z_star_99*SE)
[1] 0.01188802 0.38811198
```

Hence,

The lower bound for the 90% confidence interval is 0.08 and the upper bound is 0.320. It can be observed that the population parameter lies in this interval.

The lower bound for the 95% confidence interval is 0.057 and the upper bound is 0.343. It can be observed that the population parameter lies in this interval.

The lower bound for the 99% confidence interval is 0.012 and the upper bound is 0.388. It can be observed that the population parameter lies in this interval.

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 arandom sample used to address this research question.

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

Null Hypothesis (H0): The proportion of dangerous lead levels in Flint is less than or equal to 10%
Alternative Hypothesis (Ha): The proportion of dangerous lead levels in Flint is greater than 10%
This is a one-sided test because we're only interested in the case where the proportion is greater than 10%.

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

Code:

```
#2b
#the sample proportion and sample standard deviation of the sample
#proportion of dangerous lead levels.
flint<- read.csv("flint.csv")
n<- dim(flint)[1]
p_null<-0.1
p_hat<- mean(flint$Pb>=15)
p_hat
sample_sd<-sqrt(p_hat * (1-p_hat)/n)
sample_sd
```

Output:

```
> #2b
> #the sample proportion and sample standard deviation of the sample
> #proportion of dangerous lead levels.
> flint<- read.csv("flint.csv")
> n<- dim(flint)[1]
> p_null<-0.1
> p_hat<- mean(flint$Pb>=15)
> p_hat
[1] 0.04436229
> sample_sd<-sqrt(p_hat * (1-p_hat)/n)
> sample_sd
[1] 0.008852277
```

Hence, the sample proportion of dangerous lead levels = 0.044 and the sample standard deviation of the sample proportion of dangerous lead levels = 0.009.

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.

Code:

```
#2c
#calculate the SE of sample proportions, and the z-value for this test
SE <- sqrt(p_null*(1-p_null)/n)
SE
z_star <- (p_hat- p_null)/ SE
z_star
```

Output:

```
> #2c
> #calculate the SE of sample proportions, and the z-value for this test
> SE <- sqrt(p_null*(1-p_null)/n)
> SE
[1] 0.01289801
> z_star <- (p_hat- p_null)/ SE
> z_star
[1] -4.313667
```

Hence, the SE of sample proportions= 0.013, and the z-value for this test= -4.314.

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.

Code:

```
#2d
#we want to find the probability that we observe a sample statistic
#that has a z-score >=z_star

p_val<-1-pnorm(z_star)
p_val
```

Output:

```
> #2d
> #we want to find the probability that we observe a sample statistic
> #that has a z-score >=z_star
>
> p_val<-1-pnorm(z_star)
> p_val
[1] 0.999992
```

Hence, the p-value associated with this test =0.999992.

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

We can reject the null hypothesis only when the calculated p-value is less than the significance level. However, in this case, the p-value=0.999992 as calculated in part 2d is much larger than the significance level of 0.05. Hence, we do not have enough evidence to 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?

Since we failed to reject the null hypothesis we have no evidence to suggest that more than 10% of homes in Flint have a dangerous lead level. Hence, we would tell the EPA that they are not required to take remediation action.

__/__/____

Assignment 5
Part II

2) a) $n = 3625$ $\quad p_0 = 0.48$

sample proportion $= \frac{1830}{3625} = 0.504828$ $\quad \alpha = 0.5$

test statistic $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 - (1 - p_0)}{n}}}$

$$Z = \frac{0.504828 - 0.48}{\sqrt{\frac{0.48(1-0.48)}{3625}}} = \frac{0.024828}{0.008298}$$

$= 2.992082$

$= 2.99$

P value associated with $= 2.99 = 0.0014$

As it is a two sided test pvalue $= 2 \times 0.0014$

$= 0.0028$

As the pvalue calculate is much less than one significance level 0.05 we can say that this sample gives no evidence that the proportion of site users who got their world news on this site has changed since 2013

(b) Confidence level = $\hat{p} \pm$ margin of error

margin of error = $z^* SE \sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n}}$

$= z^* SE$

z critical value associated with 95% confidence interval = 1.96

$MOE = 1.96 \times \sqrt{\frac{0.48 \times (1-0.48)}{3625}}$

$= 0.016264$

Confidence level $= 0.48 \pm 0.016264$

$= (0.463736, 0.496264)$

$= (0.464, 0.496)$

The confidence interval agrees with the result of the hypothesis test because the p-value is present in the 95% confidence interval.

2) Type I error occurs when a true null hypothesis is rejected. In this case the error will occur if they ~~wrongly~~ decided that the proportion of voters in this age group who voted in the 2018 election was higher when it is actually equal to 0.5.

Type 2 error occurs when we fail to reject a false null hypothesis. In this case the error will occur if we say said that the proportion of voters in this age group who voted in the 2018 election was equal to 0.5 when actually $p > 0.5$.