R Homework 13

Fisher Ankney

December 4th, 2018

Statistics 5193

Question 1a

Consider Student Data - Class size and snapchat. Get a 95% students-T based confidence interval for the mean high school class size and interpret it.

```
library(readxl)
StudentData <- read_excel("/Users/fisher/Documents/data_science/r_stat_5193/data/StudentData.xlsx")

t.test(StudentData$HSClass)
```

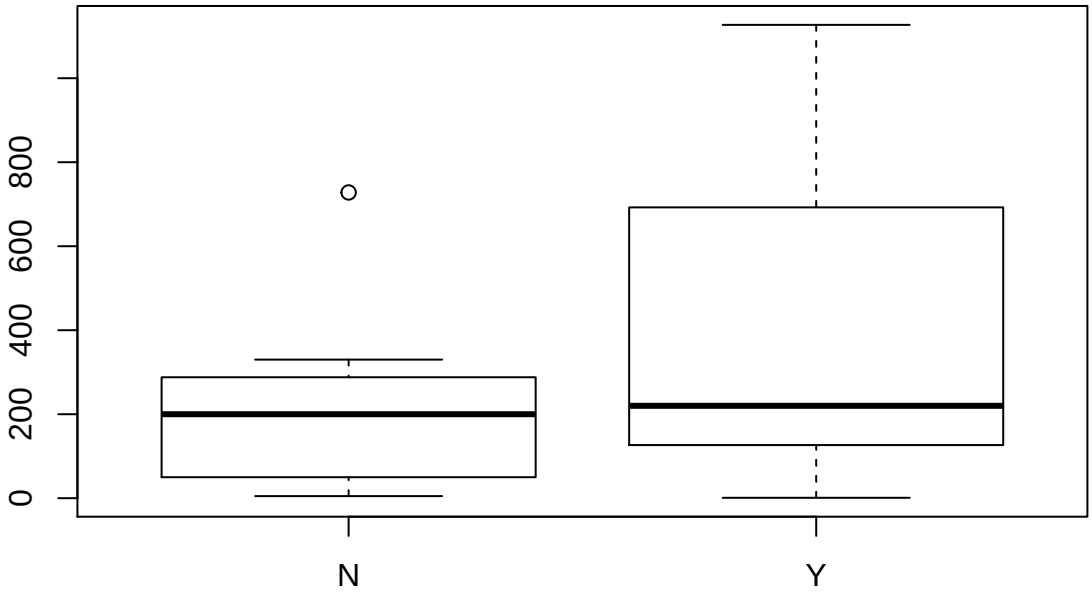
```
##
##  One Sample t-test
##
## data:  StudentData$HSClass
## t = 6.2402, df = 34, p-value = 4.208e-07
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  216.7292 426.0708
## sample estimates:
## mean of x
##     321.4
```

We can be 95% confident that the mean high school class size is between 216.73 and 426.07 students.

Question 1b

Get side by side boxplots for high school class among Snapchat users and non- Snapchat users.

```
boxplot(StudentData$HSClass~StudentData$Snapchat)
```



Question 1c

Get a p-value for testing the null hypothesis that the mean high school class size is smaller among students who do not have a Snapchat account and interpret it. Assume unequal variances and that data are normally distributed.

```
hs_no_snap <- StudentData[StudentData$Snapchat == 'N',]$HSClass
hs_yes_snap <- StudentData[StudentData$Snapchat == 'Y',]$HSClass

t.test(hs_no_snap, hs_yes_snap, var.equal = F, alternative = 'less')
```

```
##
##  Welch Two Sample t-test
##
## data:  hs_no_snap and hs_yes_snap
## t = -1.704, df = 29.876, p-value = 0.04938
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.5962173
## sample estimates:
## mean of x mean of y
##  214.0909  370.5833
```

Assuming a confidence 95%, there is sufficient evidence to suggest that the mean high school size is smaller among students who do not have snapchat. The p-value is 0.049, this barely signifying significance.

Question 1d

Get a 90% confidence interval for the difference in mean high school class size among Snapchat vs. non-Snapchat users and interpret it. Assume unequal variances and that data are normally distributed.

```
t.test(StudentData$HSClass ~ StudentData$Snapchat, var.equal = F, conf.level = .9)
```

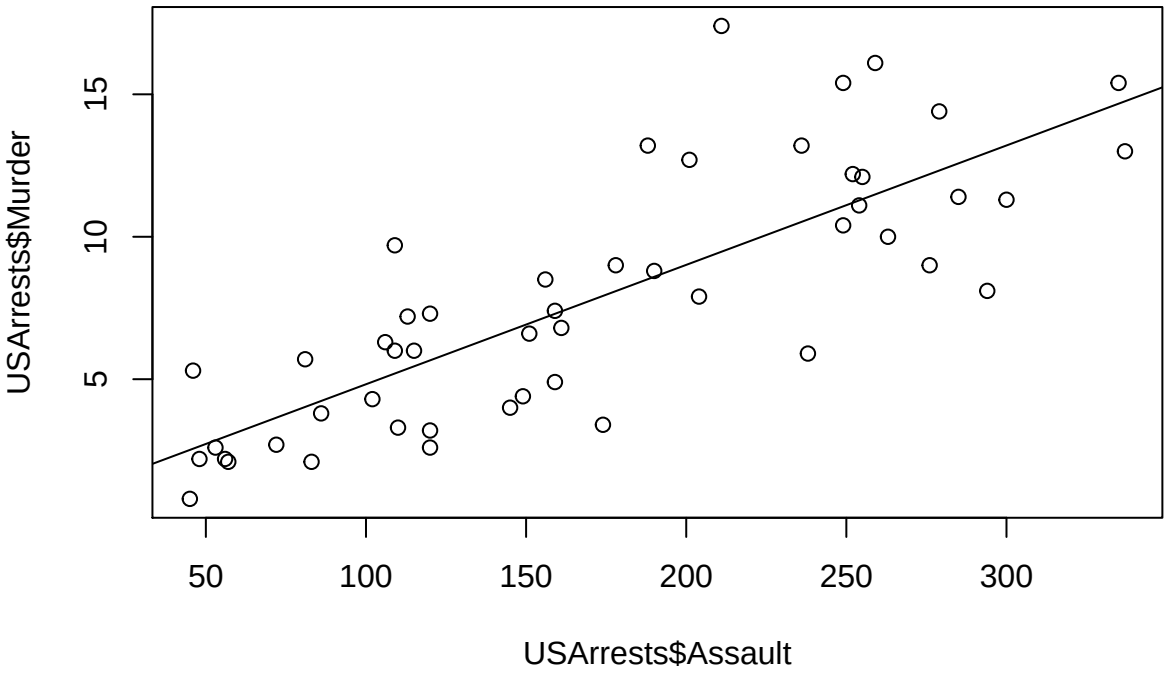
```
##
##  Welch Two Sample t-test
##
## data:  StudentData$HSClass by StudentData$Snapchat
## t = -1.704, df = 29.876, p-value = 0.09877
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -312.3886312   -0.5962173
## sample estimates:
## mean in group N mean in group Y
##        214.0909        370.5833
```

We're 90% confident that the difference in mean high school class size is between -312.39 and -0.59 students. Meaning that students who do have snapchat, have an mean high school class size that is between 0.59 and 312.39 students larger than their counterparts that do not have snapchat.

Question 2a

Consider the USArrests data - Construct a plot of Assault vs. Murder with Assault on the x axis and draw the least squares regression line on the plot.

```
plot(USArrests$Murder ~ USArrests$Assault)
lines(abline(lm(USArrests$Murder ~ USArrests$Assault)))
```



Question 2b

Are assault and murder rates independent? Use a p-value to justify your answer?

```
cor.test(USArrests$Murder, USArrests$Assault)
```

```
##
##  Pearson's product-moment correlation
##
## data:  USArrests$Murder and USArrests$Assault
## t = 9.2981, df = 48, p-value = 2.596e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6739512 0.8831110
## sample estimates:
##       cor
## 0.8018733
```

Assault and murder rates are not independent, they have a correlation coefficient of 0.8 (high positive correlation), and a pvalue of < 0.005.

Question 2c

Read the help file for predict.lm. Get a 95% prediction interval for the murder rate of a state that has an assault rate of 151 (Oklahoma's) using the predict function.

```
?predict.lm

assault_murder <- lm(Murder ~ Assault, data = USArrests)

new_data <- data.frame(Assault = 151)

predict.lm(assault_murder, new_data, interval='prediction')
```

```
##        fit      lwr      upr
## 1 6.959886 1.617596 12.30218
```

Question 2d

Add lines representing the 95% prediction intervals. That is, get lower limits and upper limits for the 95% prediction intervals as a function of x and add them to the plot.

```
pred_int <- predict.lm(assault_murder, newdata = data.frame(Assault = 50:400), interval = 'prediction')
new_x <- seq(50, 400, by = 1)

# original plot
plot(USArrests$Murder ~ USArrests$Assault)
lines(abline(lm(USArrests$Murder ~ USArrests$Assault)))

# prediction interval
lines(new_x, pred_int[,2], col = 'orange', lty = 2)
lines(new_x, pred_int[,3], col = 'orange', lty = 2)
```