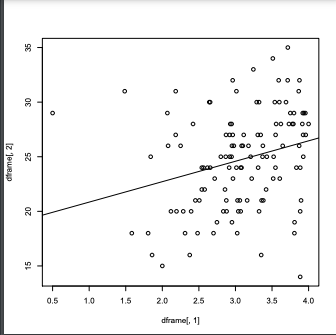
Mini Project #4

Sudarshana Jagadeeshi

Contribution of each group member: I completed the project in full

**Section 1**

**1.**



We will put GPA on the x-axis and ACT on the y-axis. The scatterplot shown has only a weak linear correlation. When we try to plot a line of best fit, it is mildly horizontal and doesn’t hug a good chunk of the data.

A point estimate of p is given by cor’ing the columns together. We get 0.2694818, which is reasonable considering that ACT scores seem to increase as GPA does, but not always.

Bias and SE are given by our R command.

Bias: 0.0007688025, SE: 0.1021135.

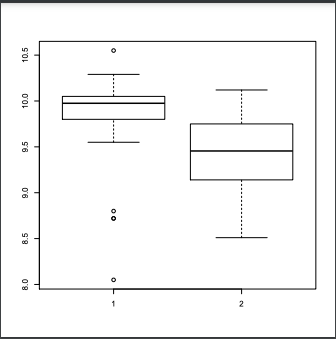
The percentile bootstrap given by our command is:

( 0.0643, 0.4720 )

This interval affirms that a weak correlation is to be expected. A strong correlation would be around .75, but this is several standard errors above our top bound of .47. A lower bound value of 0.06 is about half an SE above 0, which is no correlation. Certainly, this correlation is not strong.

**2.**

a)We make a boxplot, with the remote boxplot on the left.



We note that the local data has a much wider range, and even a different mean. The values considered typical in the remote distribution are above the 75% percentile in the local distribution. However, they are not considered outliers. By inspection, it seems likely the two distributions are different.

b)We take a confidence interval for the difference. The result is:

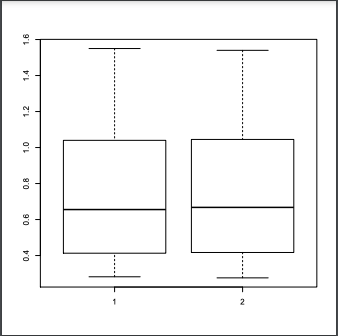
(0.1228135, 0.6398532)

The lower bound is one clear SD away from 0, and we can conclude that the two distributions are not the same. In addition, if we run a z-test in R, it will yield a p-value of 0.005, which is certainly significant for a 95% confidence interval.

c) It aligns with what we guessed earlier. While the two distributions do overlap, they only do so slightly enough that the null hypothesis should be rejected. One distribution has been shown to yield consistently higher observations over time.

**3.**

We start by exploring the data, similar to the previous section. We want to see if we can determine anything visually.



In this boxplot, both distributions seem almost identical. Their means, quantiles, max, and min are all extremely close to each other. For example, the means are 0.7605625 and 0.759875. The sd’s are 0.4054073 and 0.4041135. As long as these distributions come from roughly the same family, there is a strong chance that the CI of their difference of means will contain 0.

Then, we calculate the SE and set up a confidence interval. We want to do this to see if the interval contains the value 0, which can help confirm or deny our hypothesis.

The result is the following:

(-0.2350746, 0.2364496). This is a great result that confirms our initial suspicion. The value 0 is right in the middle of this interval, which means there is no skew difference between the two distributions. This is partly due to our similar means and SD’s between experimental and theoretical

**Section 2**

**Question 1**

**library(boot)**

**dframe <- read.csv('gpa.csv', header = TRUE) #first row is NOT data**

**plot(x=dframe[,1],y=dframe[,2])**

**abline(lm(dframe[,2] ~ dframe[,1])) #y depends on x, best fit line**

**print(cor(dframe[,1],dframe[,2])) #initial point estimate**

**bootcor <- function(data, idx){ #compute the parameter**

**return(cor(dframe[idx,1],dframe[idx,2]))**

**}**

**#from the drame data, bootcor is computed.**

**resample <- boot(data= dframe, statistic= bootcor, R=1000,sim = 'ordinary')**

**print(boot.ci(resample)) #prints all CI's**

**Question 2**

**dframe2 <- read.csv('VOLTAGE.csv', header = TRUE) #first row is NOT data**

**loc0 <- subset(dframe2, location == '0')[,2]**

**loc1 <- subset(dframe2, location == '1')[,2]**

**#SE Formula**

**SE <- (sd(loc0)\*\*2) /length(loc0) + (sd(loc1)\*\*2) / length(loc1)**

**SE <- SE\*\*0.5**

**print(SE)**

**#1.96 is the value needed for a 95% CI**

**boxplot(loc0,loc1)**

**print( c( mean(loc0)-mean(loc1) - 1.96\* SE, mean(loc0)-mean(loc1) + 1.96\* SE ))**

**Question 3**

**dframe3 <- read.csv('VAPOR.csv', header = TRUE) #first row is NOT data**

**theo <- dframe3[,2] #theoretical**

**expr <- dframe3[,3] #experimental**

**#SE formula for diff of mean**

**SE2 <- (sd(theo)\*\*2) /length(expr) + (sd(theo)\*\*2) / length(expr)**

**SE2 <- SE2\*\*0.5**

**print(SE2)**

**boxplot(theo,expr) #view distributions**

**t\_value<- qt(0.95, 15) #we use t this time as n<30**

**print( c( mean(theo)-mean(expr) - t\_value\*SE2, mean(theo)-mean(expr) + t\_value\* SE2 ))**