**STATISTICAL METHODS FOR DATA SCIENCE**

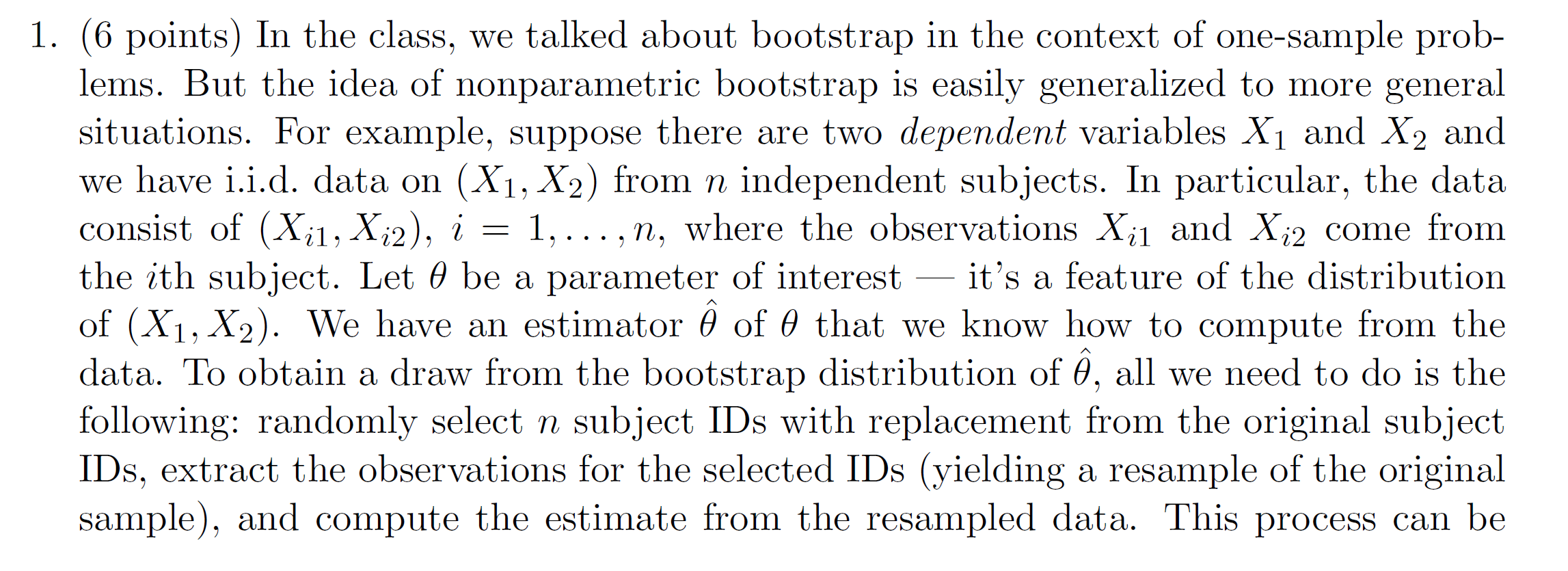
**CS 6313-001 FALL 2019**

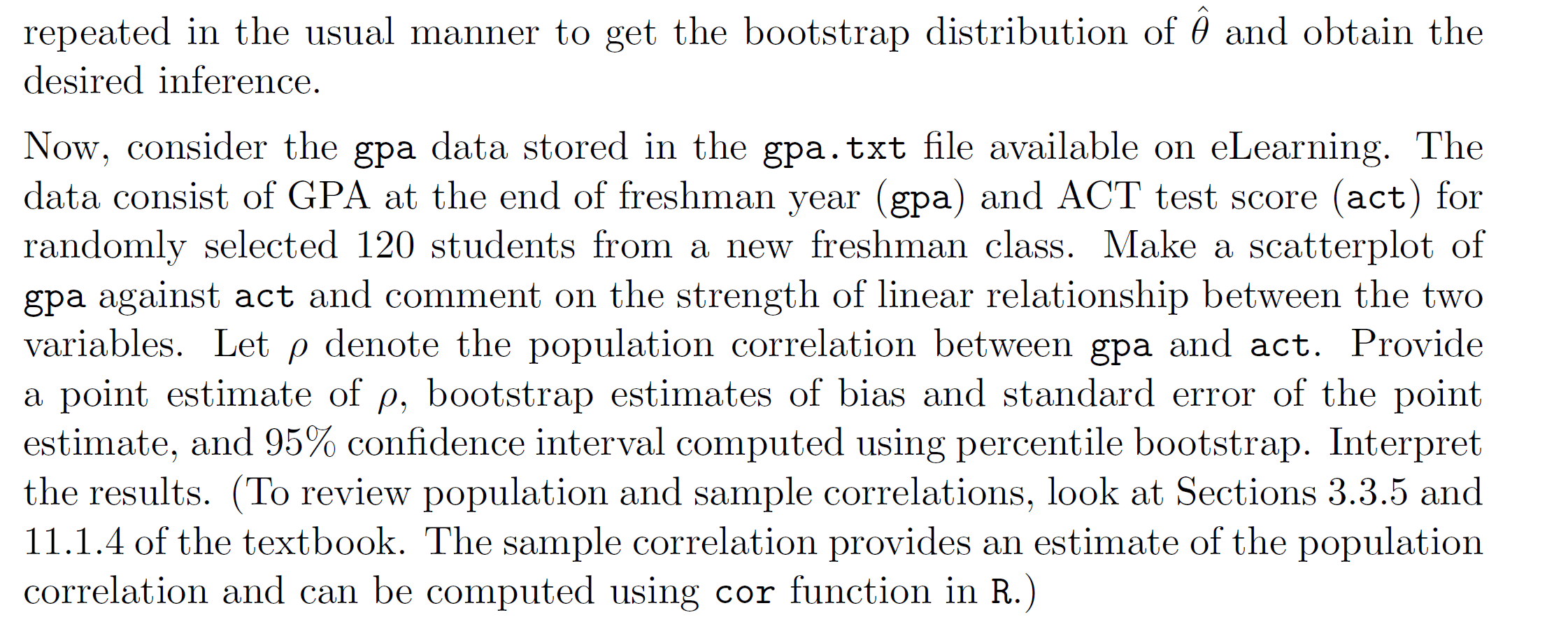
**Mini Project #4**

**Participants :**

**Reetish Chand Guntakal Patil (RXG190006)**

**Question 1:**





Reading the gpa data

total\_data <- read.csv(“gpa.csv”)

|  |
| --- |
| > cor(total\_data$gpa, total\_data$act)  [1] 0.2694818  > corcoeff.nonpar <- function(x,y,indices){  + result <- corr(x[indices],y[indices])  + return(result)  + }  > corcoeff.nonpar.boot <- boot(total\_data, corcoeff.nonpar,  R = 999,sim = "ordinary",stype = "i") |
|  |
| |  | | --- | |  | |

> boot(total\_data, corcoeff.nonpar,R = 999,sim = "ordinary",stype = "i")

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

boot(data = total\_data, statistic = corcoeff.nonpar,

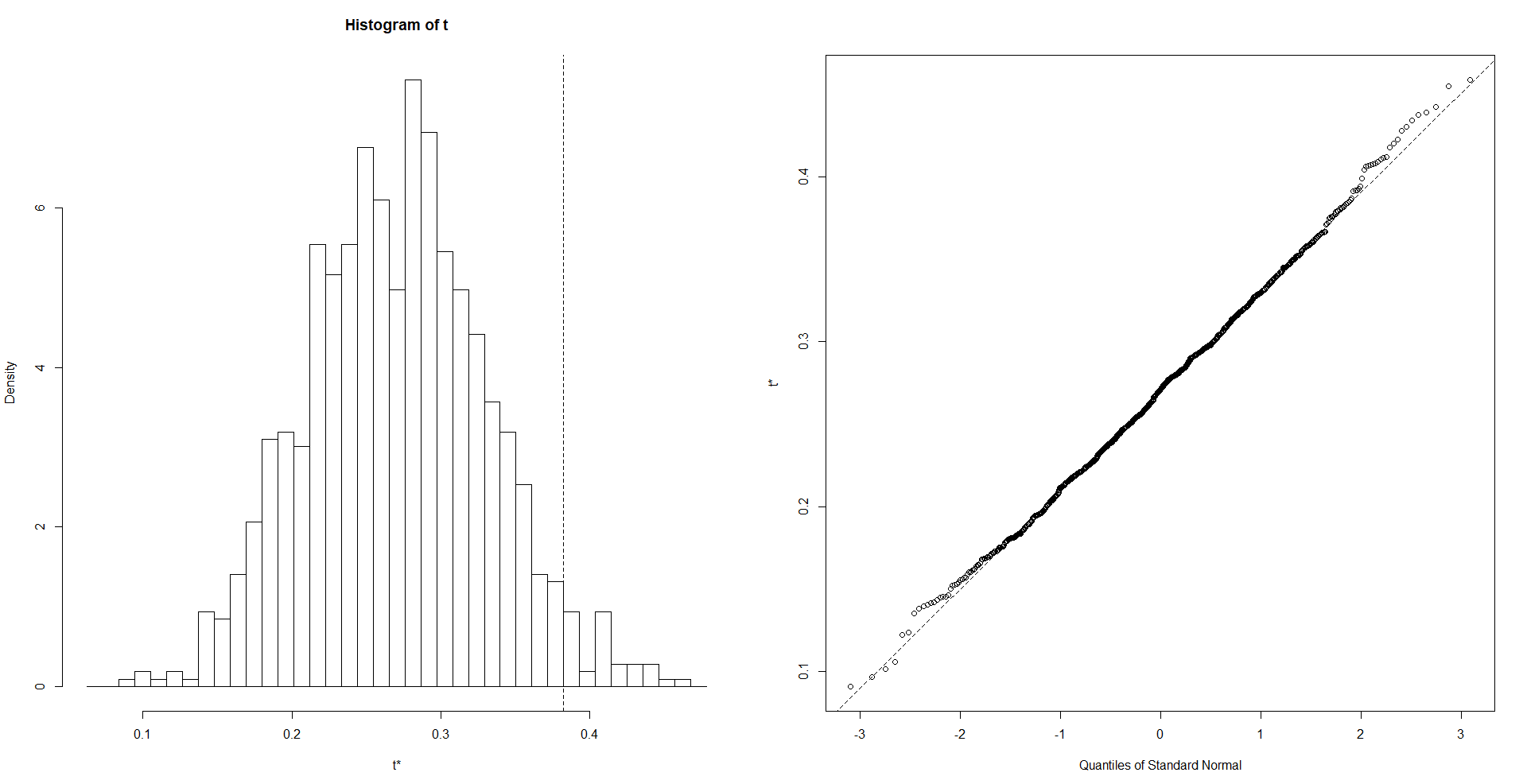
R = 999, sim = "ordinary", stype = "i")

Bootstrap Statistics :

original bias std. error

t1\* 0.3824455 -0.1122878 0.06036216

> plot(corcoeff.nonpar.boot)



> sort(corcoeff.nonpar.boot$t)[c(25,975)]

[1] 0.1568071 0.3920582

> boot.ci(boot.out = corcoeff.nonpar.boot, type = 'perc')

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 999 bootstrap replicates

CALL :

boot.ci(boot.out = corcoeff.nonpar.boot, type = ‘perc’)

Intervals :

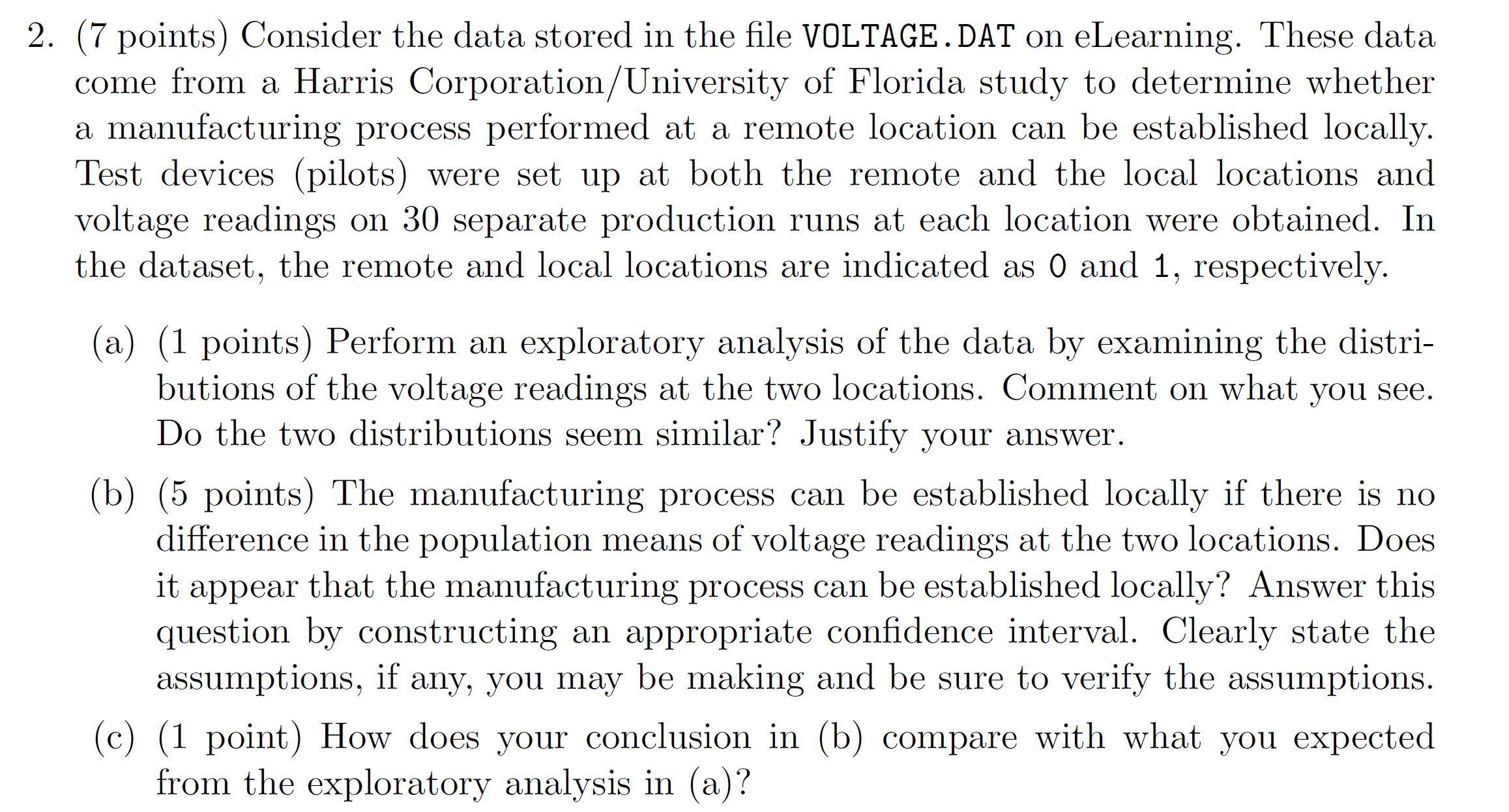
Level Percentile

95% ( 0.1568, 0.3921 )

Calculations and Intervals on Original Scale

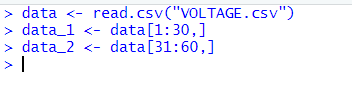
So, the 95% percentile bootstrap confidence interval comes out to be **[ 0.1568, 0.3921 ]** and the correlation coefficient is calculated as **0.2694818**. Therefore, the correlation coefficient is a really good estimate of the given GPA vs ACT data.

**Question 2 :**



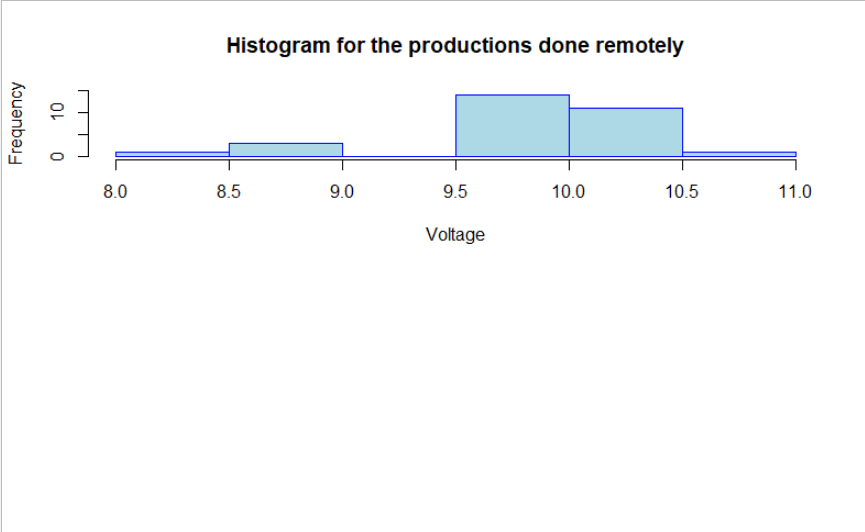
a.

Reading the data:



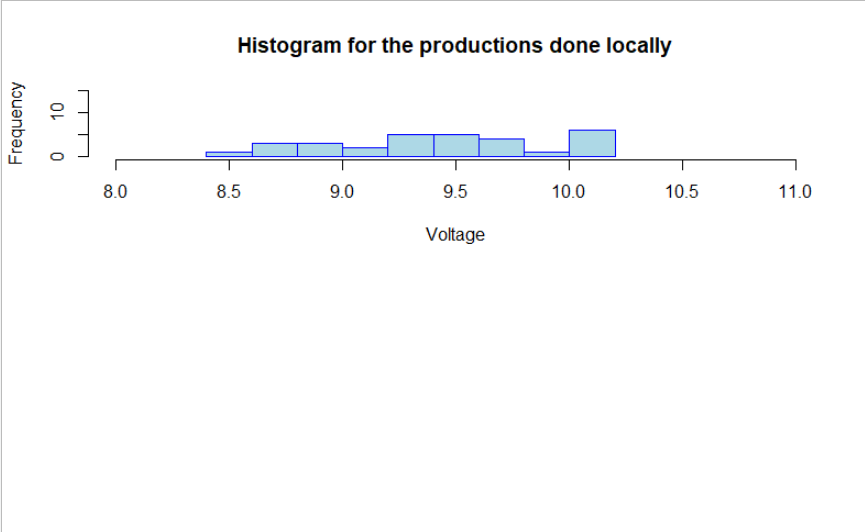
Histogram for the productions done remotely:





Histogram for the productions done locally:

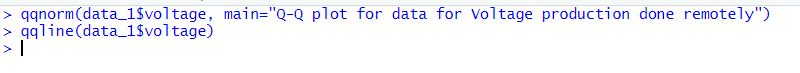


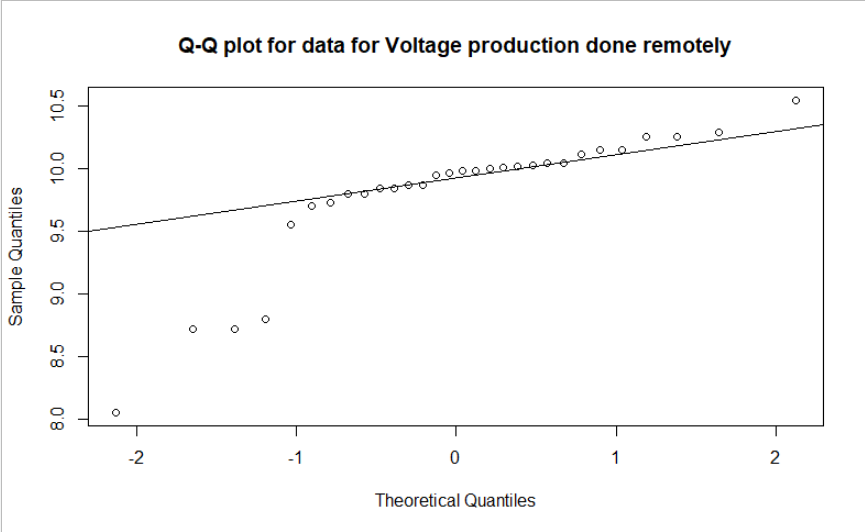


Observation:

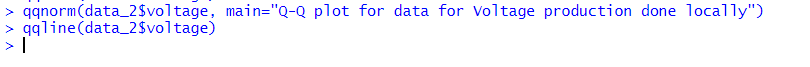
The distribution of the voltage produced remotely and locally varies a lot.

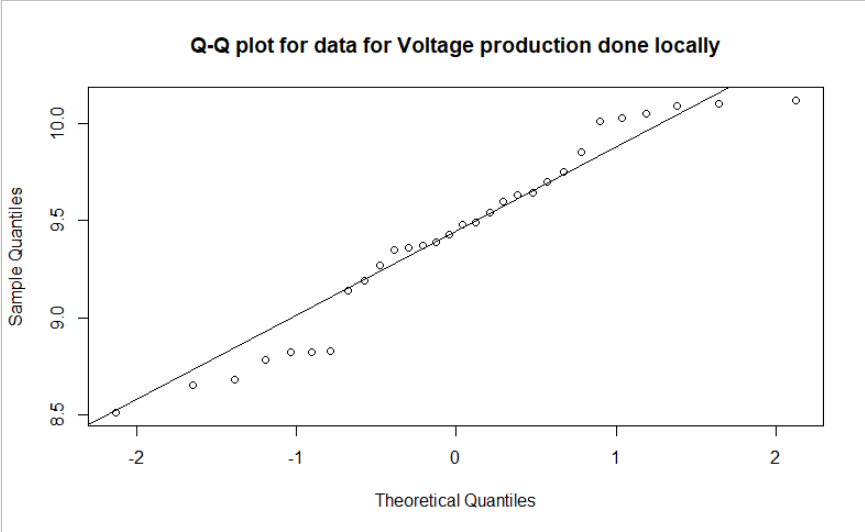
Q-Q plot for data for Voltage production done remotely:





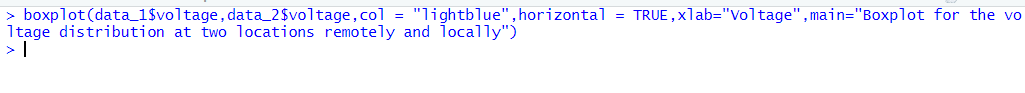
Q-Q plot for data for Voltage production done locally:

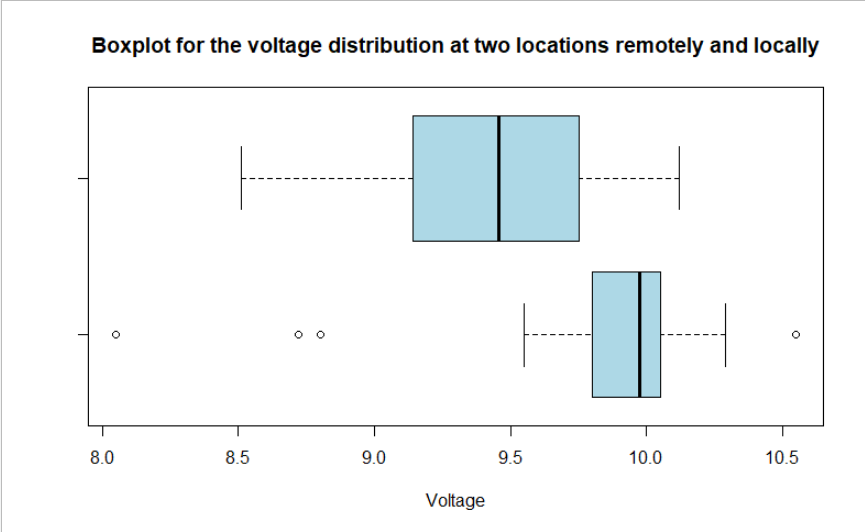




Observation:

Both distributions are approximately normal





Obseravtion:

The variance for both the data is not the same. So, the variance can’t be assumed to be equal for both populations.

Finding the difference between the mean of voltage distribution remotely and mean of voltage distribution locally:

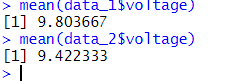


Observation:

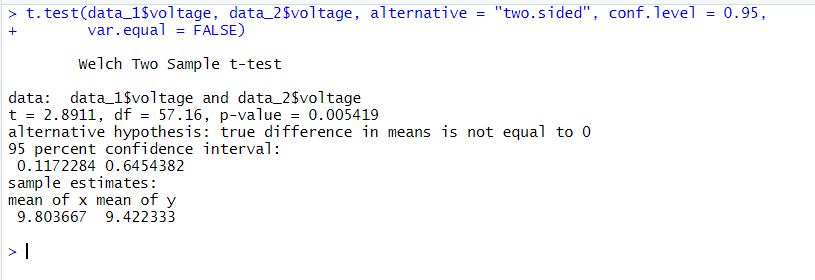
From the boxplot we can infer that the median for the locally voltage distribution is greater than the median for remotely voltage distribution and the voltages for the local location are have less varied range than the production done remotely. Also, the difference in the mean of the two population comes out to be positive.

b.

Finding mean value of voltage distribution remotely and mean of voltage distribution locally:



Performing the T-test:



Observation:

The confidence interval comes out to be positive

95% CI [0.1172284, 0.6454382]

This means the following

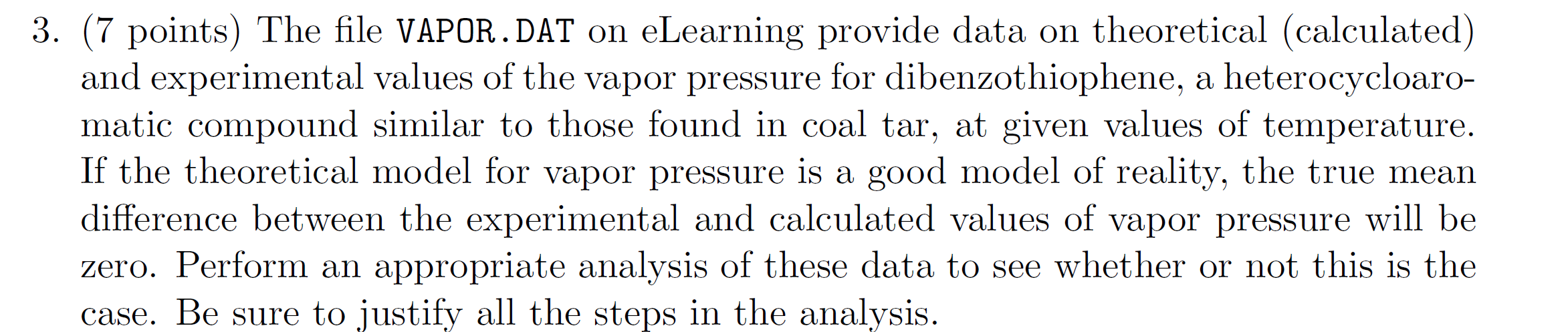
**mean(voltage\_production\_remotely) - mean(voltage\_production\_locally)** > 0 => **mean(voltage\_production\_remotely)** > **mean(voltage\_production\_locally)**

So, we can say that the Null Hypothesis is rejected and a manufacturing process performed at a remote location can’t be established locally.

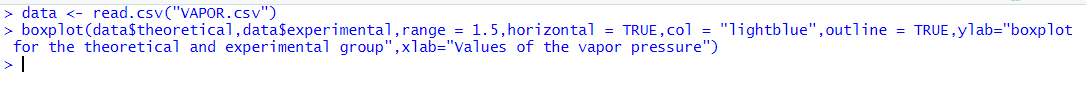
c.

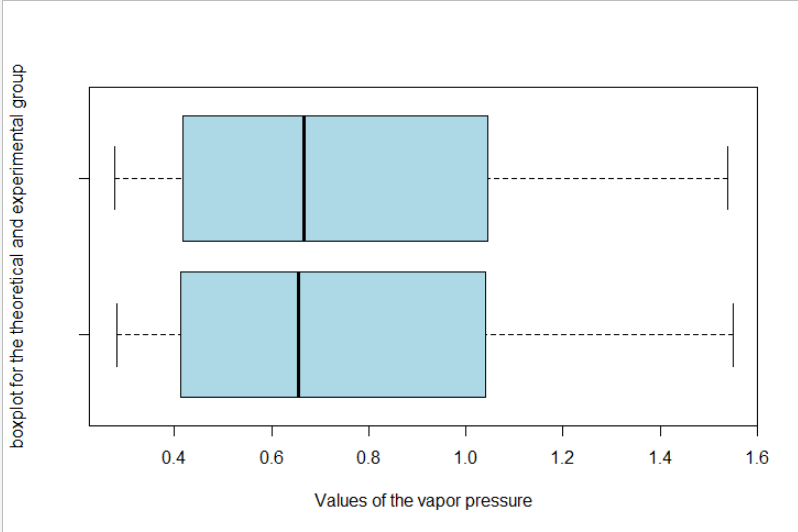
From the deterministic tests (mean,boxplot and histogram) it was clear that the voltages produced at both the locations were not similar(the mean for both the population varied a lot) but we needed an inferential test (Hypothesis Test) to validate the results.

Question 3:



Reading the data and plotting a boxplot:

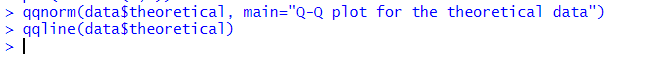


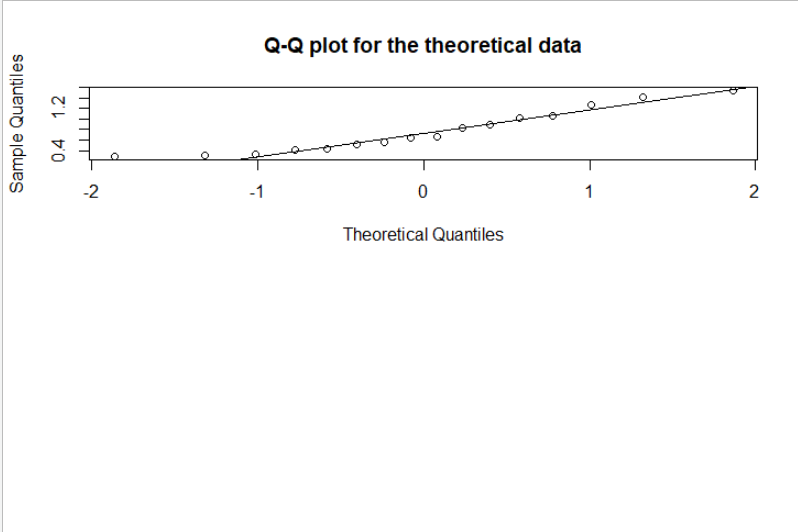


Observation:

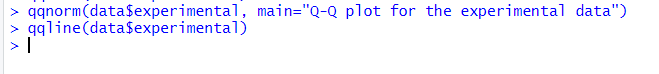
Both the experimental and the theoretical distributions are approximately the same except for a slight difference in the median of the population

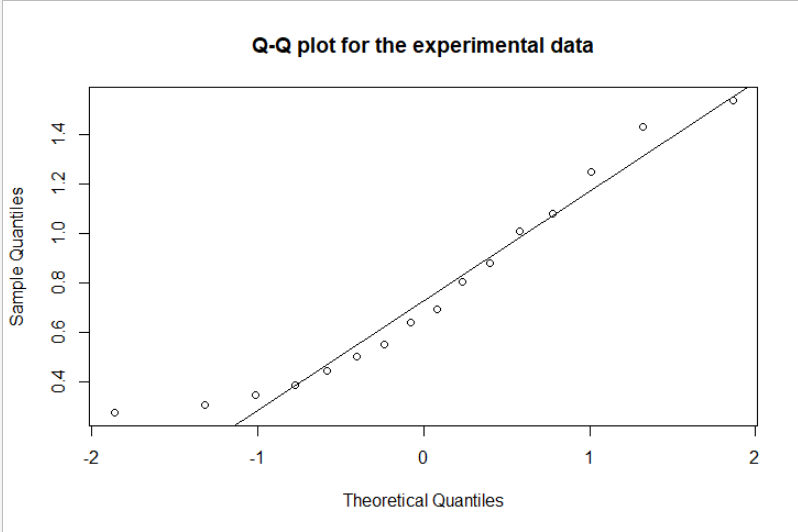
QQplot for theoretical data:





QQplot for experimental data:



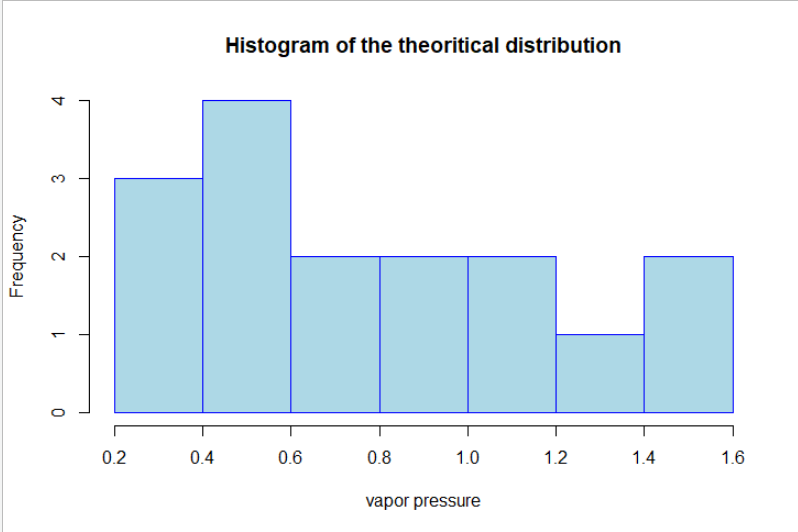


Observation:

Both the distributions are approximately normal.

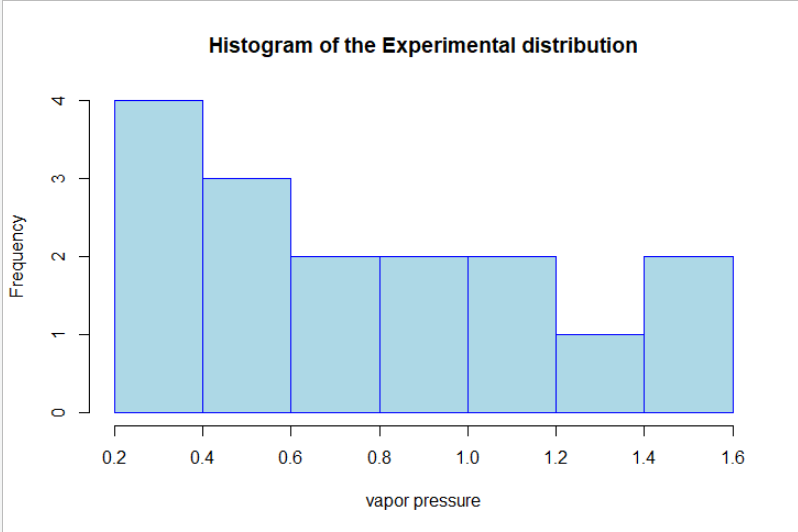
Histogram for theoretical data:





Histogram for experimental data:

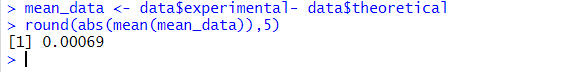




Observation:

The Histogram tells us except for the frequency change in first two breaks, rest all the frequency distribution tends to be same which tells us the mean difference might be 0.

The true mean difference between the experimental and calculated values of vapor pressure approximately is 0.

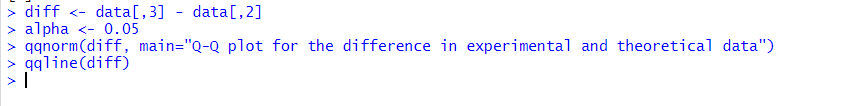


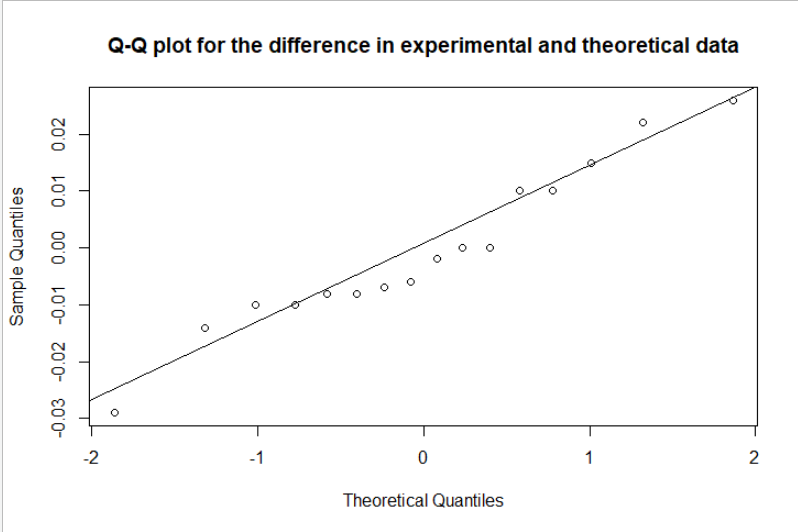
Let’s do the Hypothesis testing to validate the results:

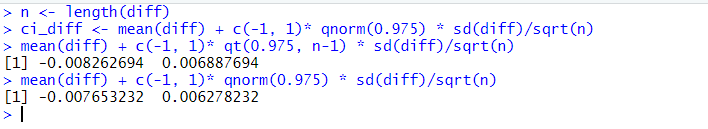
**Null Hypothesis** : The true mean difference between the experimental and calculated values of vapor pressure will be equal to zero.

**Alternate Hypothesis** : The true mean difference between the experimental and calculated values of vapor pressure will **not** be equal to zero.

We will calculate the confidence Interval using paired sample C







Observation:

The confidence is **[-0.007653232 , 0.006278232]** which tells that the difference may be equal to zero at many points, so we accept the Null Hypothesis that the theoretical model for vapor pressure is a good estimate of the real world model. We cannot accept the alternative hypothesis because of lack of evidence which supports it.