Mini Project # 4

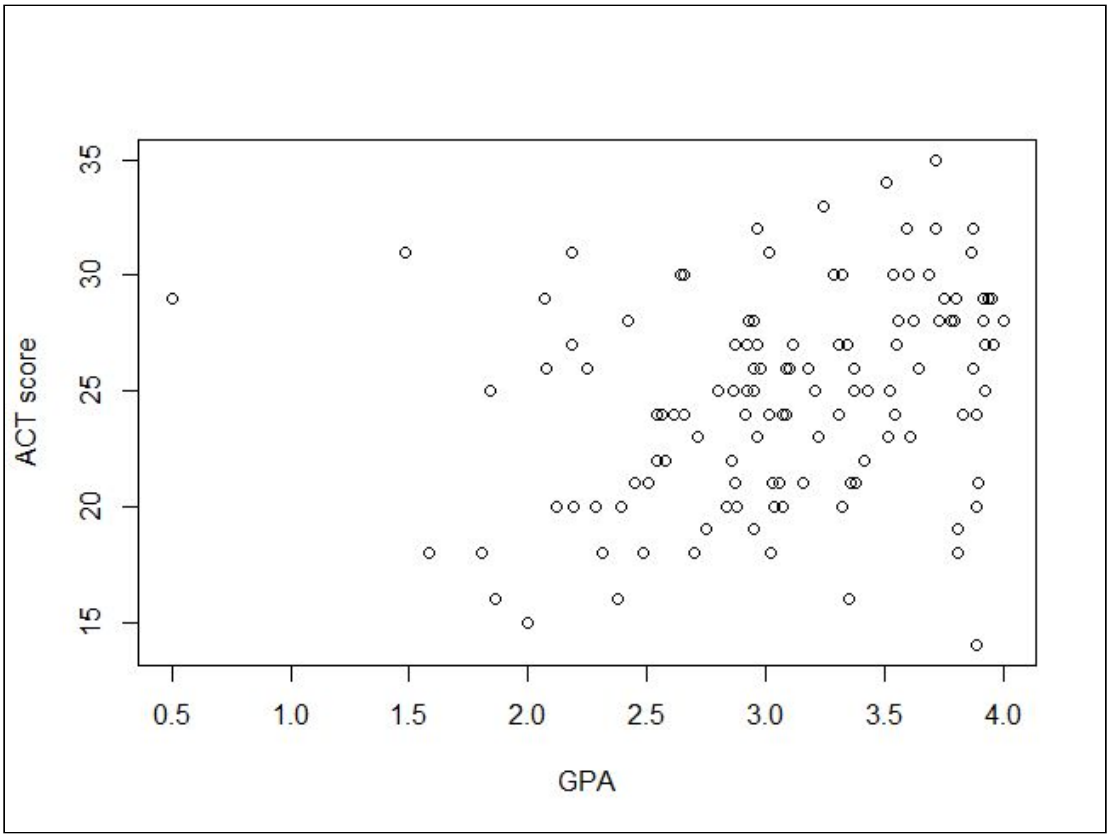
Qingyu Lan, Lakshmi Priyanka Selvaraj

Contribution of each group member :

Both members worked on the questions together.

Section 1. Answers to the specific questions asked

1. Question 1



As we can observe from this scatterplot that there is no clear relationship between the GPA and the ACT scores. So we use cor() function to obtain the correlation between the two values.
#we will compute the correlation to show the strength

```
initial_cor <- cor(gpa$gpa, gpa$act)
#0.2694818
```

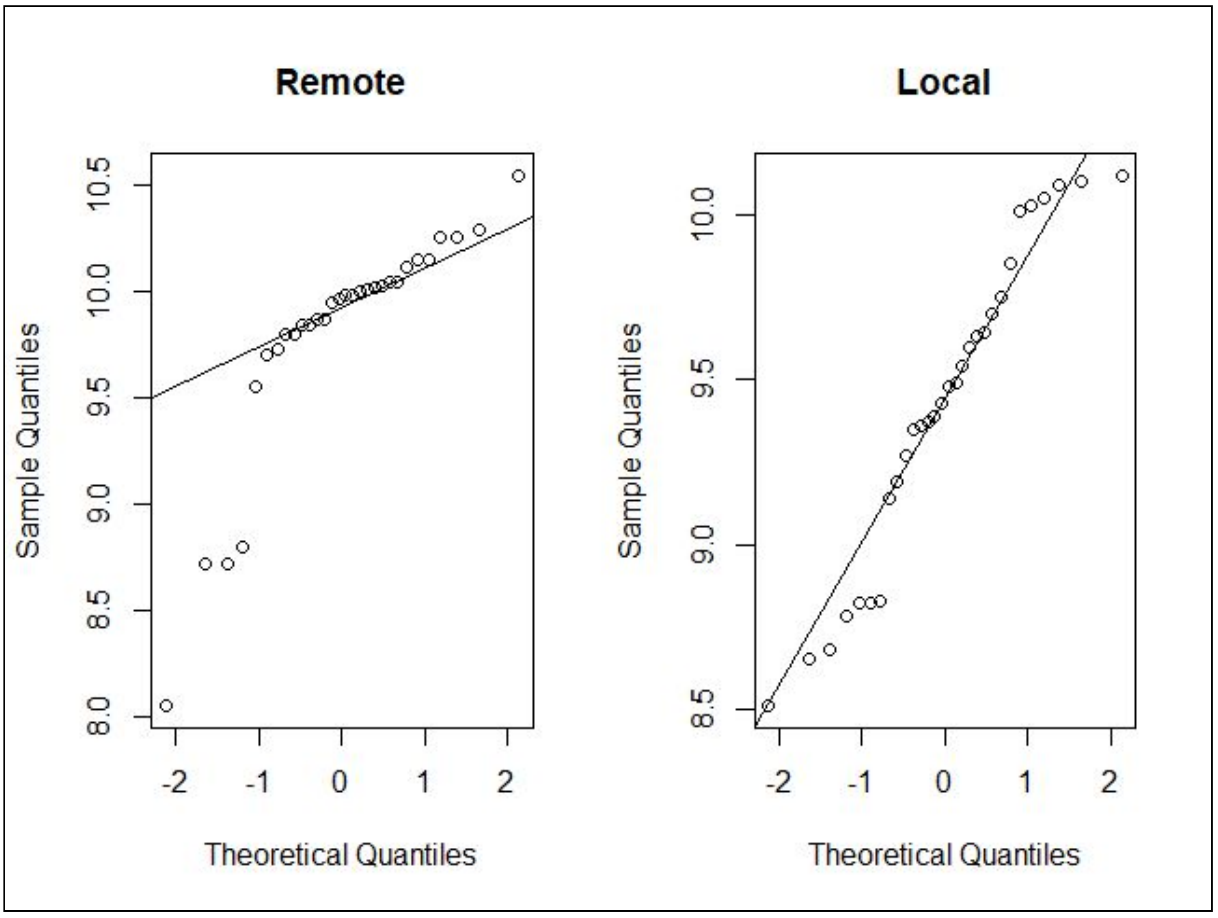
The correlation value is 0.269, which is closer to 0 and on the weaker side of correlation. If the magnitude of correlation is closer to 1, we say the two values have strong positive or negative correlation.

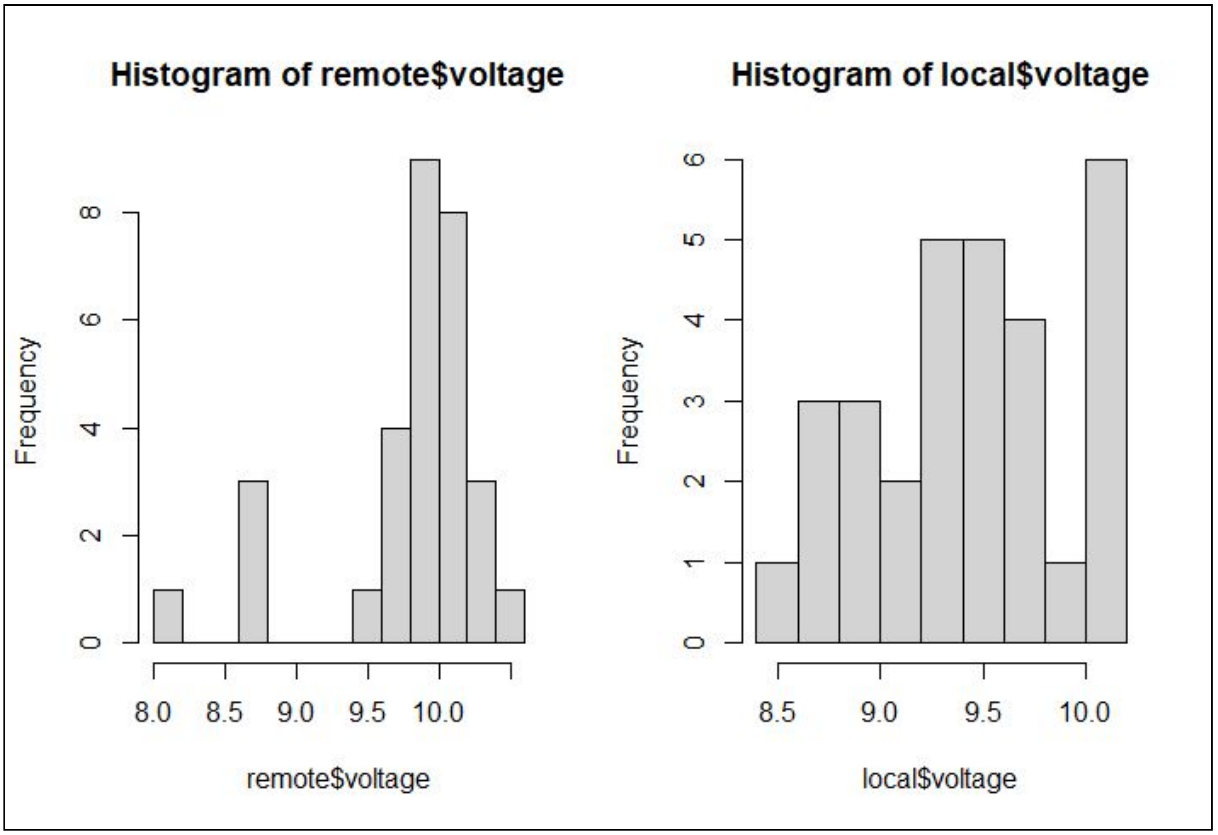
Bootstrap Sampling:

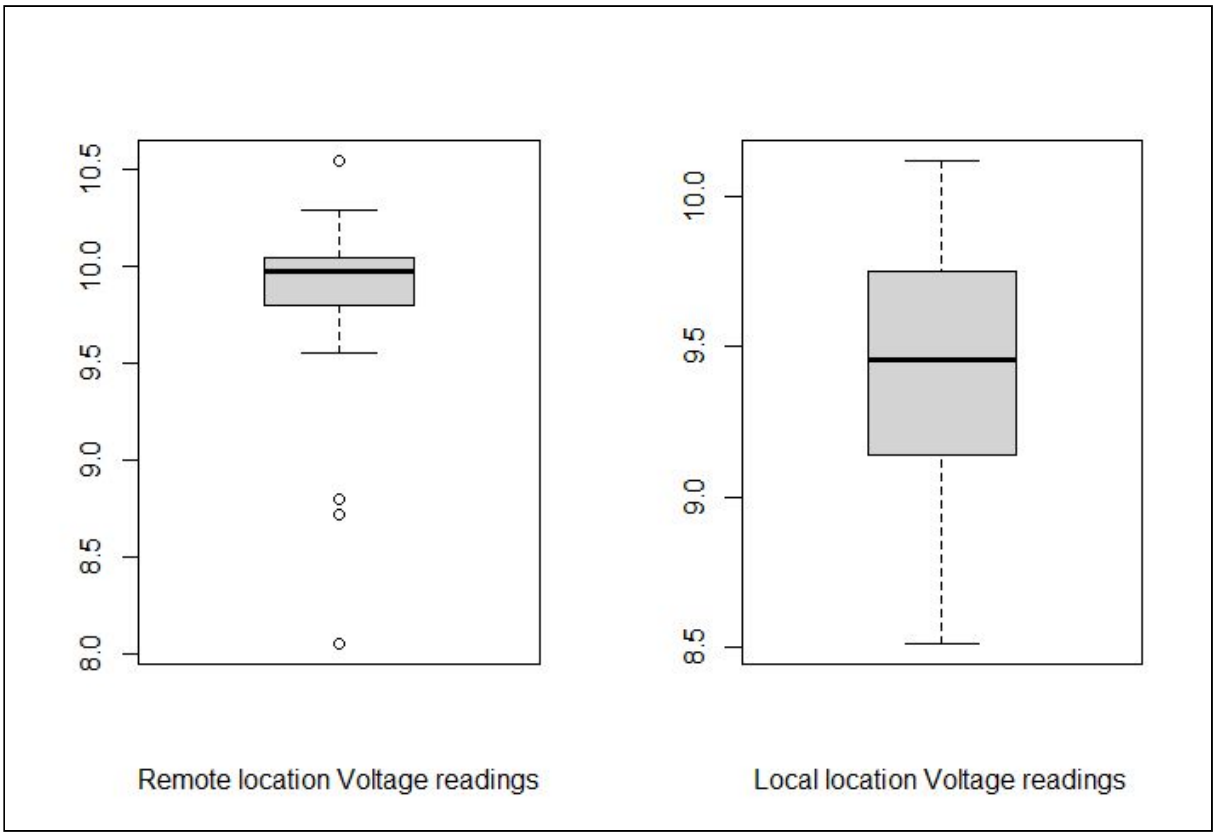
```
#resample_boot contains 1000 resampling correlation values
mean_resample <- mean(resample.nboot)
Mean_resample
0.2699309

#bias of estimate
bias <- mean_resample - initial_cor
Bias
0.0004490583

#Standard deviation of resampling
sd(resample.nboot)
0.1055667

#95% confidence interval using percentile bootstrap

sort(resample.nboot)[ceiling(c(nboot*0.025,nboot*0.975))]

#0.06460706 0.48080984
```

The bootstrap sampling correlation CI shows that there is a weak correlation between the GPA and ACT scores. The lower CI is almost 0, which suggests a weak or almost no correlation between the two values.

2. Question 2


Remote
Local
Sample Quantiles
Theoretical Quantiles
8.0
8.5
9.0
9.5
10.0
10.5
-2
-1
0
1
2


Histogram of remote$voltage
Histogram of local$voltage
Frequency
remote$voltage
local$voltage
0
2
4
6
8
1
3
5
8.0
8.5
9.0
9.5
10.0



a. The remote location reading is concentrated and highly precise, with approximately 90% of the values close to median value 10. The values of voltage readings from local locations are widespread and not very precise. Also, in remote location readings, the values are heavily left skewed as suggested from the bigger left box, while the local location values are fairly symmetric. Therefore, the distributions are different.

b. Null hypothesis:

The difference between mean of remote and mean of local = 0

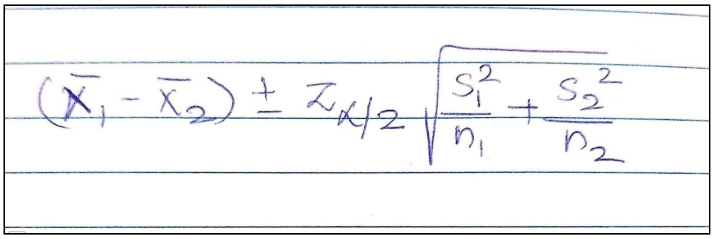
Alt hypothesis:

The difference between mean of remote and mean of local is not 0

Estimate Population mean difference using sample mean difference & building a CI.

Step 1: Is the population normal? No, the population is not normal.

Step 2: Is the sample size large enough, n>=30? Yes, n=30. So by CLT, the population can be considered to be normal and we can use sample variance and standard normal distribution values. We use the formula



We obtain the 95% confidence interval

0.1228182 0.6398484

According to the condition given, there should be no difference between the population means of the voltages of two locations to establish the manufacturing process locally. But even with the widest CI of 99% confidence, the difference seems to be more than 0, between the two sample means. Based on the CI, we have to accept the alternative hypothesis. Therefore, the manufacturing process cannot be installed locally.

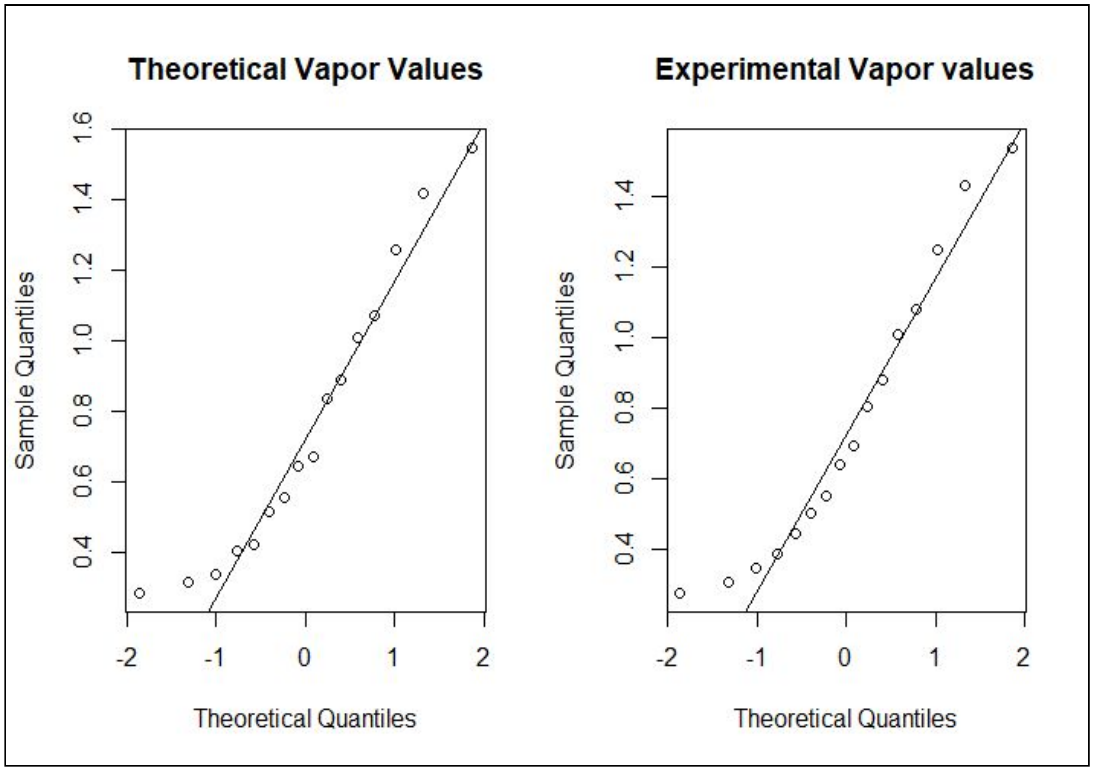
c. The 90,95,99% CI doesn't contain 0 and the difference in the population mean shows that the remote location voltages are on the higher side when compared to the local voltages

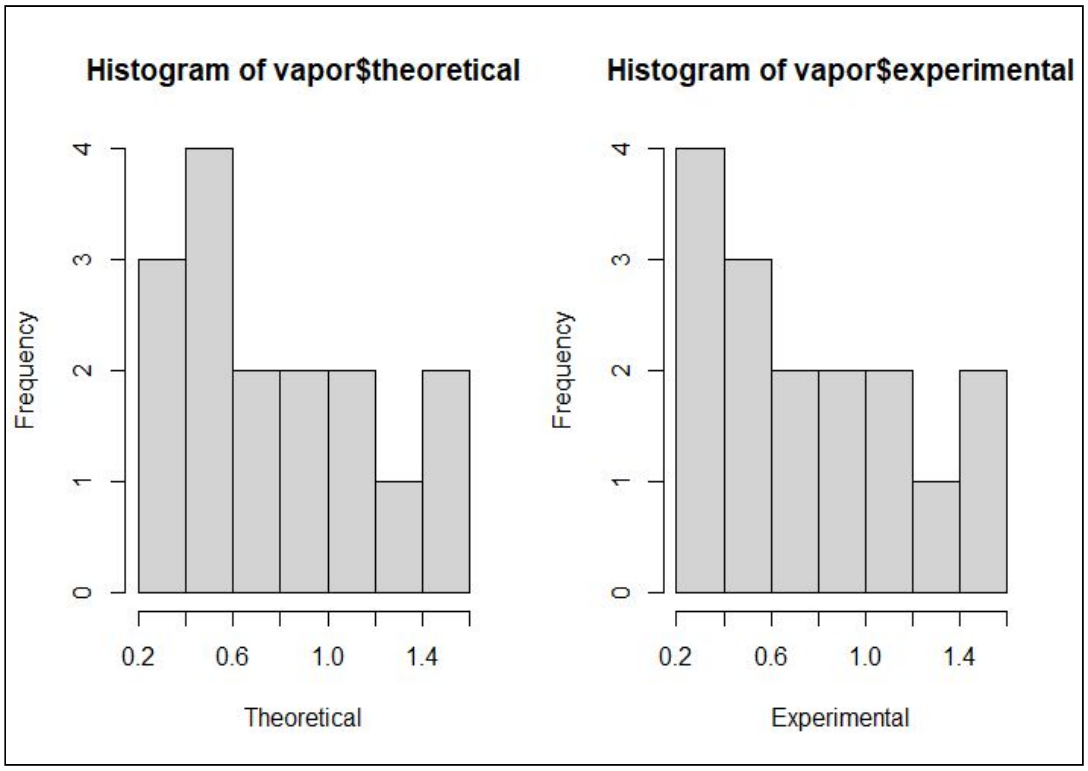
CI_90
[1] 0.1643806 0.5982860

CI_99
[1] 0.04158692 0.72107975

From part A we can tell that voltage reading is on average higher at remote locations than those at local locations. Since the higher voltage is required to power the manufacturing process at the remote location, based on the result of parts A and B we can conclude that we cannot have the manufacturing process locally.

3. Question 3





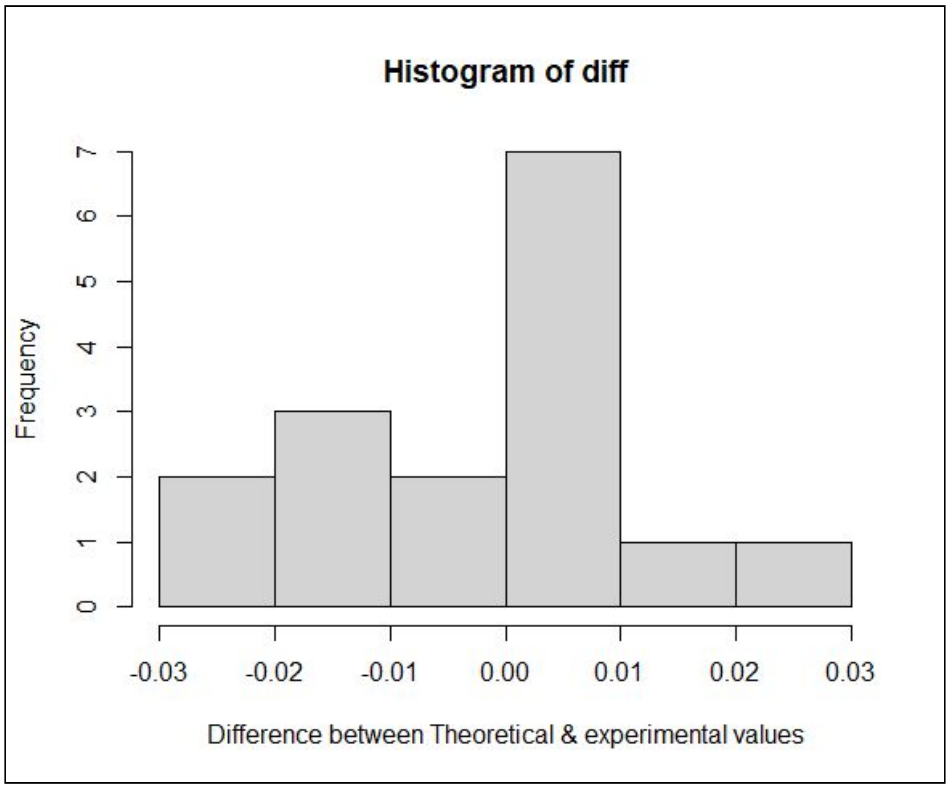
As we observe from the above plot, both have really similar distributions and histograms. This is a paired sample and so, we can convert it into a one sample problem, by using difference between the paired values as the parameter.

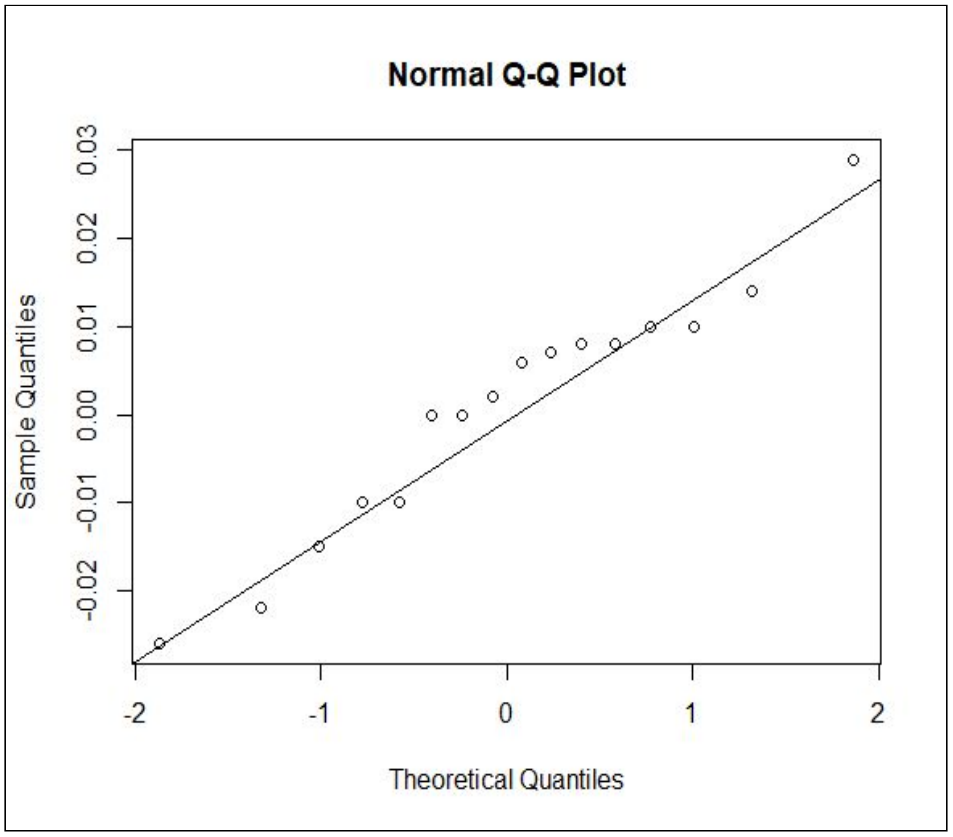
Difference = Theoretical - Experimental

Now using this new sample dataset of difference values, we construct plots.

We observe from the below plots that the distribution approaches normal behavior. Since we don't know the population SD, we use the SD of the new sample and Student's T distribution comes in handy for calculating CI for the above assumptions.
We use the formula

$$\bar{X} \pm t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}}$$





Null Hypothesis: True mean difference between mean of Theoretical and Experimental is 0
Alt Hypothesis: True mean difference between mean of Theoretical and Experimental not 0

The 95% confidence interval was calculated using both manual steps and the t.test() method in R. The values are as follows:

-0.006887694 0.008262694

We can observe the confidence interval contains 0 and both the upper and lower limits are very close to 0 value, therefore we can accept our null hypothesis. There is very minimal difference between the theoretical and the experimental values from the random sample, and this goes to prove that the theoretical model for vapor pressure will serve as a good model of reality for the population as well.

Section 2: R code.

R code for question 1

```
gpa = read.csv("gpa.csv")

plot(gpa$gpa, gpa$act, xlab = "GPA", ylab = "ACT score")
initial_cor <- cor(gpa$gpa, gpa$act)

set.seed(123)
#creating paired resampling
nboot <- 1000
resample.nboot <- numeric(nboot) #creates a numeric vector of size nboot

for(h in 1:nboot){
  boot.index <- sample(1:nrow(gpa),replace=TRUE)
  boot.data <- gpa[boot.index,]
  resample.nboot[h] <-cor(boot.data$gpa, boot.data$act)
}

#boxplot(resample.nboot)

#head(resample.nboot)

mean_resample <- mean(resample.nboot)
mean_resample

#bias of estimate
bias <- mean_resample - initial_cor
bias

#Standard deviation of resampling
sd(resample.nboot)
```

```
#95% confidence interval using percentile bootstrap

sort(resample.nboot)[ceiling(c(nboot*0.025,nboot*0.975))]
```

R code for question 2

```
a)
voltage <- read.csv("VOLTAGE.csv")

head(voltage)

remote <- subset(voltage, location==0, select=c(location,voltage))
local <- subset(voltage, location==1, select=c(location,voltage))

plot(remote$location, remote$voltage)
qqnorm(remote$voltage)
qqline(remote$voltage)

plot(local$location, local$voltage)
qqnorm(local$voltage)
qqline(local$voltage)

par(mfrow=c(1,2))
hist(remote$voltage, breaks=10)
hist(local$voltage, breaks=10)

boxplot(local$voltage, xlab="Local location Voltage readings")
boxplot(remote$voltage, xlab="Remote location Voltage readings")

b)

x1_mean <- mean(remote$voltage)
x2_mean <- mean(local$voltage)

sd1 <- sd(remote$voltage)
sd2 <- sd(local$voltage)

n1 <- length(remote$voltage)
n2 <- length(local$voltage)

#CI
CI_95 <- (x1_mean-x2_mean)+c(-1,1)*qnorm(0.975)*(sqrt( (sd1^2)/n1 + (sd2^2)/n2 ))
CI_95

CI_90 <- (x1_mean-x2_mean)+c(-1,1)*qnorm(0.95)*(sqrt( (sd1^2)/n1 + (sd2^2)/n2 ))
CI_90
```

```
CI_99 <- (x1_mean-x2_mean)+c(-1,1)*qnorm(0.995)*(sqrt( (sd1^2)/n1 + (sd2^2)/n2 ))
CI_99
```

R code for question 3

```
vapor <- read.csv("vapor.csv")
head(vapor)

qqnorm(vapor$theoretical)
qqline(vapor$theoretical)

qqnorm(vapor$experimental)
qqline(vapor$experimental)

#the distributions show that they are very similar to each other

hist(vapor$theoretical, breaks = 8, xlab="Theoretical")
hist(vapor$experimental, breaks = 8, xlab= "Experimental")

v1_mean <- mean(vapor$theoretical)
v2_mean <- mean(vapor$experimental)
v1_mean
v2_mean

diff <- vapor$theoretical - vapor$experimental
hist(diff, breaks=6, xlab="Difference between values")
qqnorm(diff)
qqline(diff)

#Since sample almost normal, we use Difference(merge two sample to single sample)
#as the base for computing CIs

ci_inbuilt <- t.test(diff,alternative = "two.sided")

#or

x_bar <- mean(diff)
t_val <- qt(0.975,length(diff)-1)
s <- sd(diff)
sqrt_n <- sqrt(length(diff))

ci_manual <- x_bar+c(-1,1)*t_val*s/sqrt_n

#-0.006887694  0.008262694
#Values for both the manual & inbuilt method are same.
```