Mini Project 4

Name: Tejas Ravi Rao (txr171830)

Section – 1

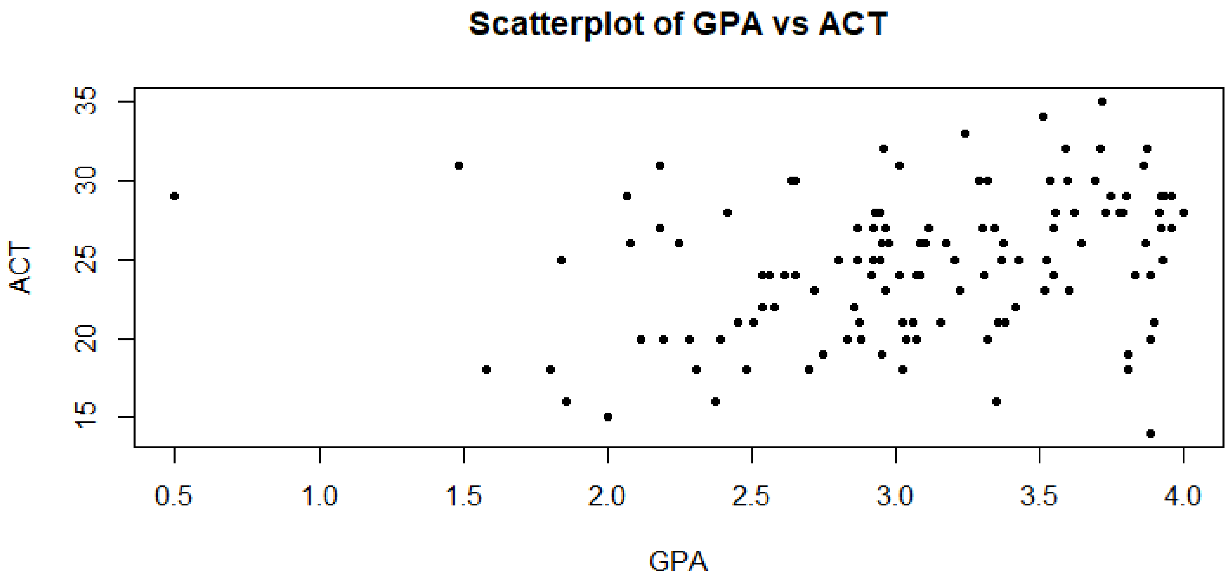
Q(1) Initially, read the following given “gpa.csv” file

```
# read "gpa.csv" file
gpacsv = read.csv("D:/Users/rao29/Documents/Sem 5/Stats for Data Science/MP4/gpa.csv", header = TRUE, sep = ',')
```

As required, obtain Scatterplot for GPA vs ACT.

```
# plot scatter plot of gpa against act
plot(gpacsv$gpa, gpacsv$act, main = "Scatterplot of GPA vs ACT", xlab = "GPA", ylab = "ACT", pch=20)
```

The following plot was obtained.



Also, the point estimate of population correlation between GPA and ACT is obtained as follows

```
# point estimate of population correlation(rho) between GPA and ACT
rhop = cor(gpacsv$gpa,gpacsv$act)
print(paste("Point estimate of correlation between GPA and ACT = ", rhop))
```

```
[1] "Point estimate of correlation between GPA and ACT =  0.269481803266264"
```

The population correlation (rho) obtained is 0.269481803266264.

- From the scatterplot we see that the GPA and ACT have a positive and linear correlation.
- However, this may not be a strong correlation as the plot seems very scattered.
- The value of correlation coefficient tells us about the strength of the linear relationship.
- The point estimate of correlation between GPA and ACT is 0.269481803266264. This value is greater than 0 which shows a positive association.

A function to determine the correlation between GPA and ACT values is defined. This function is used in the following non-parametric bootstrap function.

```
# define function to find correlation between GPA and ACT
corr.npar = function(x,indices){
  result = cor(gpacsv$gpa[indices], gpacsv$act[indices])
  return(result)
}
```

Bootstrap estimates of bias and standard error of the point estimate are determined. The above defined function is used.

```
# calculate point estimate, bias and standard error values
(corr.npar.boot = boot(data = gpacsv, corr.npar, R = 999, sim = "ordinary", stype = "i"))
```

The following output was obtained.

```
ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = gpacsv, statistic = corr.npar, R = 999, sim = "ordinary",
    stype = "i")

Bootstrap Statistics :
     original       bias      std. error
t1* 0.2694818  0.00225264   0.1070387
```

95% Confidence Interval for correlation between GPA and ACT is computed using percentile Bootstrap.

```
# calculate 95% confidence interval using percentile bootstrap
boot.ci(corr.npar.boot,conf = 0.95, type = "perc")
```

The following 95% Confidence Interval was obtained

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

CALL :
boot.ci(boot.out = corr.npar.boot, conf = 0.95, type = "perc")

Intervals :
Level     Percentile
95%   ( 0.0631,  0.4784 )
Calculations and Intervals on Original Scale
```

- The 95% Confidence Interval Obtained was [0.0631, 0.4784].
- As observed above, the bias value from bootstrap is 0.00225264. This bias value is small, which shows that bootstrap estimate is close to the actual correlation value.
- Standard Error obtained from bootstrap is 0.1070387. As a result, this indicates that the observations are closer to actual values.
- From the above statistics, we may confirm from our findings that the above Confidence Interval (using the Percentile Bootstrap method) contains the correlation point estimate.

Q2(a) Initially read the following given “VOLTAGE.csv” file.

```
# read "VOLTAGE.csv" file
data = read.csv("D:/Users/rao29/Documents/Sem 5/Stats for Data Science/MP4/VOLTAGE.csv",
header = TRUE, sep = ",")
```

Obtain and Separate the remote and local location values.

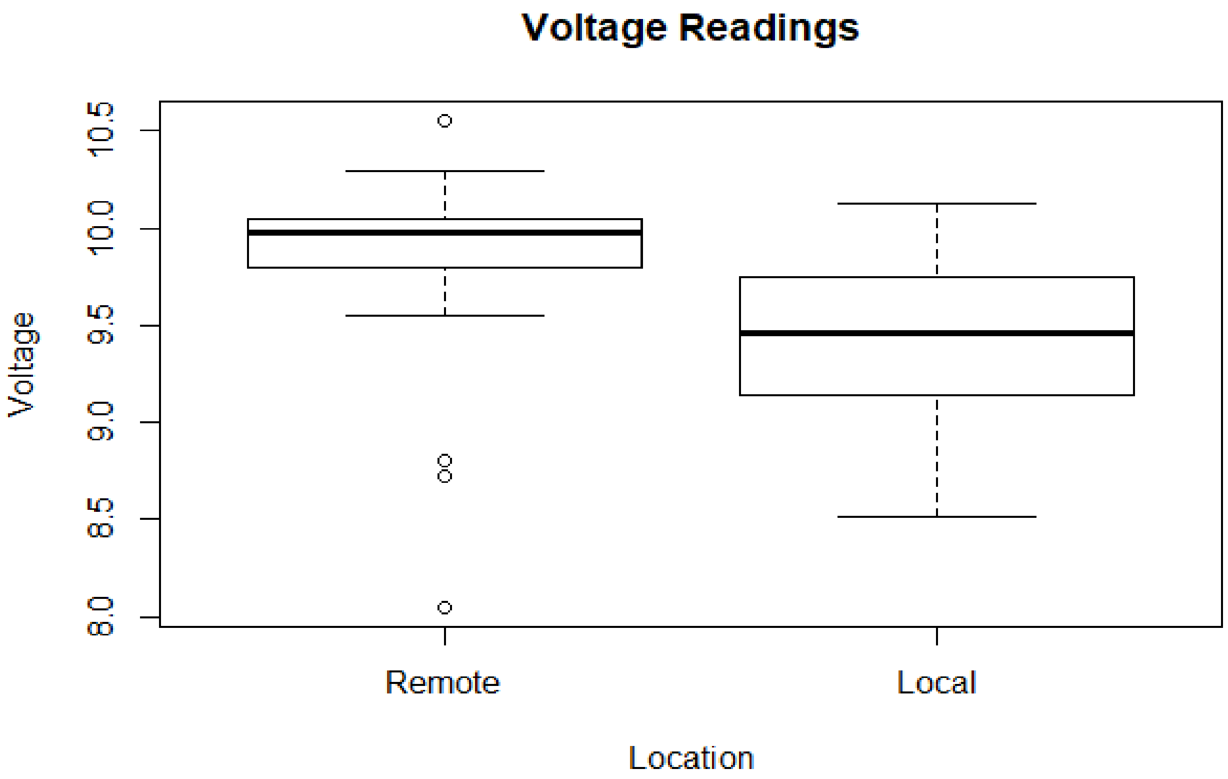
```
# obtain remote and local values separately
remoteVal = subset(data, location == "0")
localVal = subset(data, location == "1")
```

To perform exploratory data analysis, initially side by side boxplots are obtained. These boxplots compare Voltage readings obtained at Remote Locations with Local Locations.

# Question 2(a)

# plot side by side boxplots of voltage distributions at both locations
boxplot(remoteVal$voltage, localVal$voltage, main = "Voltage Readings", names = c("Remote", "Local"), xlab = "Location", ylab = "Voltage")

The following plots were obtained.



Additionally, Summary statistics along with IQR value and Standard deviation values were obtained for Voltage readings at both categories of locations.

# get summary statistics for voltage at remote location

summary(remoteVal$voltage)
IQR(remoteVal$voltage)
sd(remoteVal$voltage)

The following values were obtained.

```
 Min.   1st Qu.   Median   Mean    3rd Qu.    Max.
 8.050   9.800    9.975    9.804   10.050    10.550
```

IQR:

```
[1] 0.25
```

Standard Deviation:

```
[1] 0.5409155
```

Similarly,

```
# get summary statistics for voltage at local location

summary(localVal$voltage)
IQR(localVal$voltage)
sd(localVal$voltage)
```

The following values were obtained.

```
  Min.   1st Qu.   Median   Mean    3rd Qu.   Max.
  8.510   9.152    9.455    9.422   9.738     10.120
```

IQR:

```
[1] 0.585
```

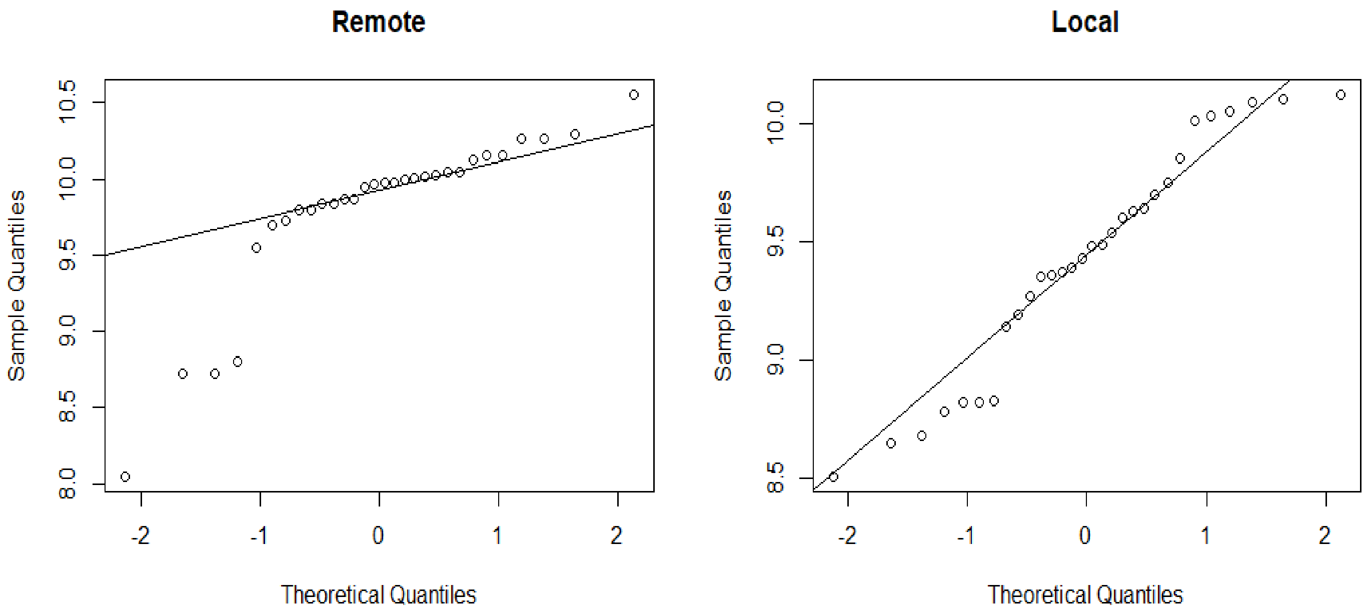
Standard Deviation:

```
[1] 0.4788757
```

Finally, plot QQ-plots for both voltage readings to get an idea of their distribution.

```
# obtain normal QQ plots
par(mfrow = c(1,2))
qqnorm(remoteVal$voltage, main = "Remote")
qqline(remoteVal$voltage)
qqnorm(localVal$voltage, main = "Local")
qqline(localVal$voltage)
```

The following plots are shown in the next page.



- No, both the distributions do not seem to be similar. From the boxplot, we find that the Remote Location data have outliers which would deviate the plot from being normal. This can be observed in the respective QQ plot.
- Whereas, the Local Location data has a distribution very much closer to that of Normal as observed in the respective QQ plot.
- Based on values of mean, median, Q1 and Q3, the Remote Locations have higher Voltage reading values than Local Locations.
- Both the distributions have different variability as they have different Inter Quartile Ranges which are 0.25 for Remote and 0.585 for Local location respectively.

Q2(b) As shown above, we assume the two distributions to be normal. Hence, a 95% confidence interval will be constructed.

The mean and variance values are calculated as follows. These will be used in difference of means to construct the appropriate confidence intervals.

```
# Question 2(b)
# construct appropriate confidence interval with assumption
# consider 95% CI, alpha = 0.05

alpha = 1-0.95

# mean of voltages at remote
(remoteVal.mean = mean(remoteVal$voltage))
```

```
[1] 9.803667

# mean of voltages at local
(localVal.mean = mean(localVal$voltage))

[1] 9.422333

# variance of voltages at remote
(remoteVal.var = var(remoteVal$voltage))

[1] 0.2925895

# variance of voltages at local
(localVal.var = var(localVal$voltage))

[1] 0.229322
```

Notice the unequal variances of the voltage reading at the two locations. Note the values of n and m below.

```
# number of remote locations
(n = nrow(remoteVal))

[1] 30

# number of local locations
(m = nrow(localVal))

[1] 30
```

With the above given values, let us construct the appropriate 95% Confidence Interval for the difference in the two population means.

```
# CI for difference in means
diff_means = remoteVal.mean - localVal.mean + c(-1,1)*qnorm(1-(alpha/2))*sqrt((remoteVal.var/n) +
(localVal.var/m))
diff_means

[1] 0.1228182 0.6398484
```

The Confidence Interval obtained is [0.1228182, 0.6398484]. The Confidence Interval is found to be part on the right of 0.

There is difference in the mean of voltages recorded at remote locations and local locations. The mean value of voltages recorded at remote location is higher than the mean of voltages recorder at local locations. The difference of the means lies in [0.1228182, 0.6398484].

The previously shown QQ plot shows that the two distributions are closer to normal. It is also observed that n and m values are large enough. The sample variances are unequal.

Therefore, the manufacturing process cannot be established locally.

Q2(c) From 2(a) the following mean values were obtained from the summary statistics.

The mean voltage at remote location is 9.803667.
The mean voltage at local location is 9.422333.

The difference in the mean voltages of remote and local locations is 0.381334. This value lies in the confidence interval obtained from 2(b) which is [0.1228182, 0.6398484].

From this, it can be observed that 0 does not lie in the Confidence Interval and the obtained Confidence Interval lies to the right of 0. The mean voltage at remote location is greater than the mean voltage at local location.

Therefore from 2(a) and 2(b) we may conclude that the manufacturing process cannot be established locally.

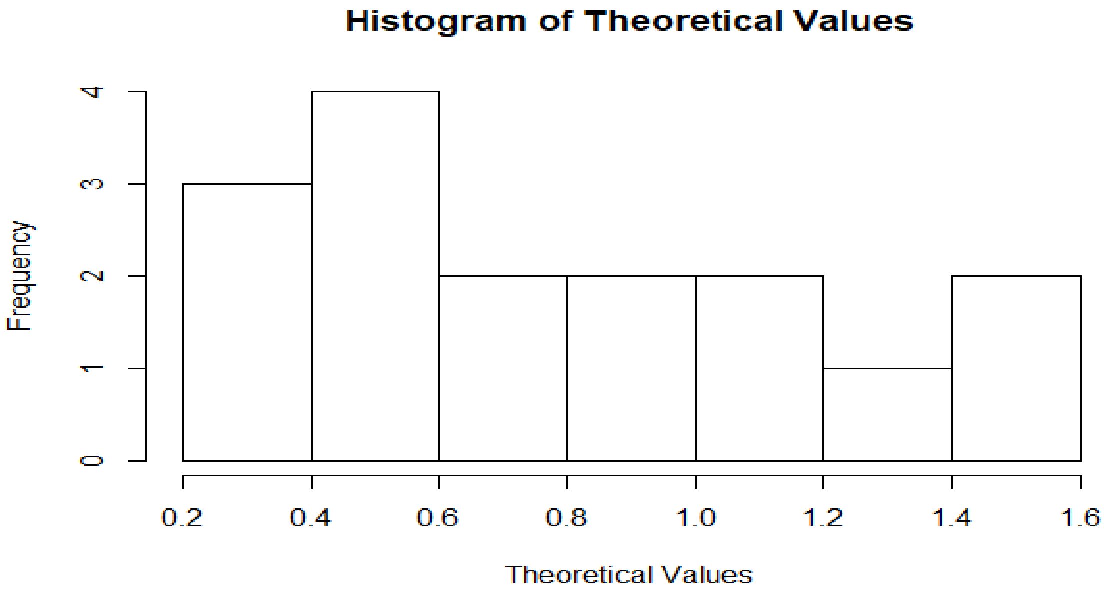
Q(3) Initially read the “VAPOR.csv” file.

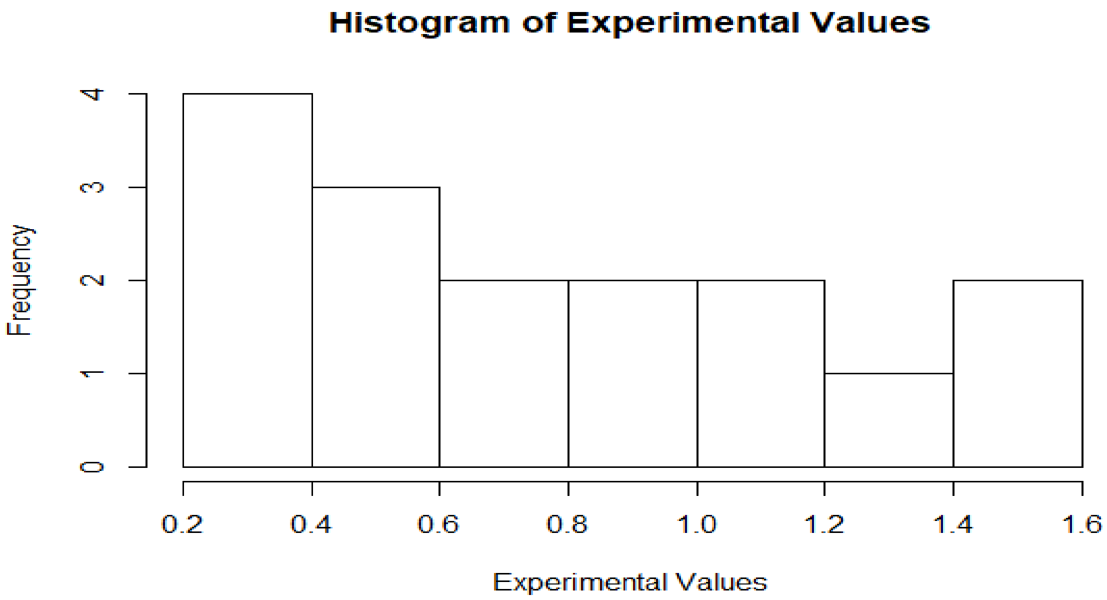
```
# read "VAPOR.csv" file
vapdata = read.csv("D:/Users/rao29/Documents/Sem 5/Stats for Data Science/MP4/VAPOR.csv", header = TRUE, sep = ",")
```

The size of the sample N = 17 is very small. The histogram and QQ plots of theoretical as well as experimental values must be first observed to gain an understanding of the distributions.

```
# histogram for theoretical and experimental values
hist(vapdata$theoretical, main = "Histogram of Theoretical Values", xlab = "Theoretical Values")
hist(vapdata$experimental, main = "Histogram of Experimental Values", xlab = "Experimental Values")
```

The following plots were obtained.



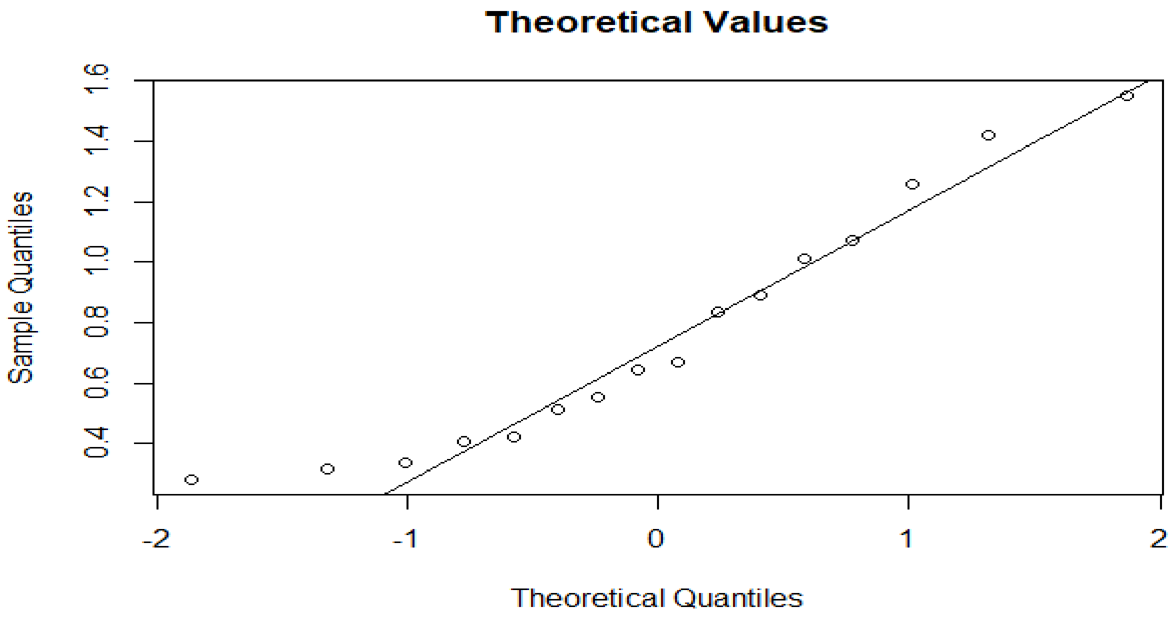


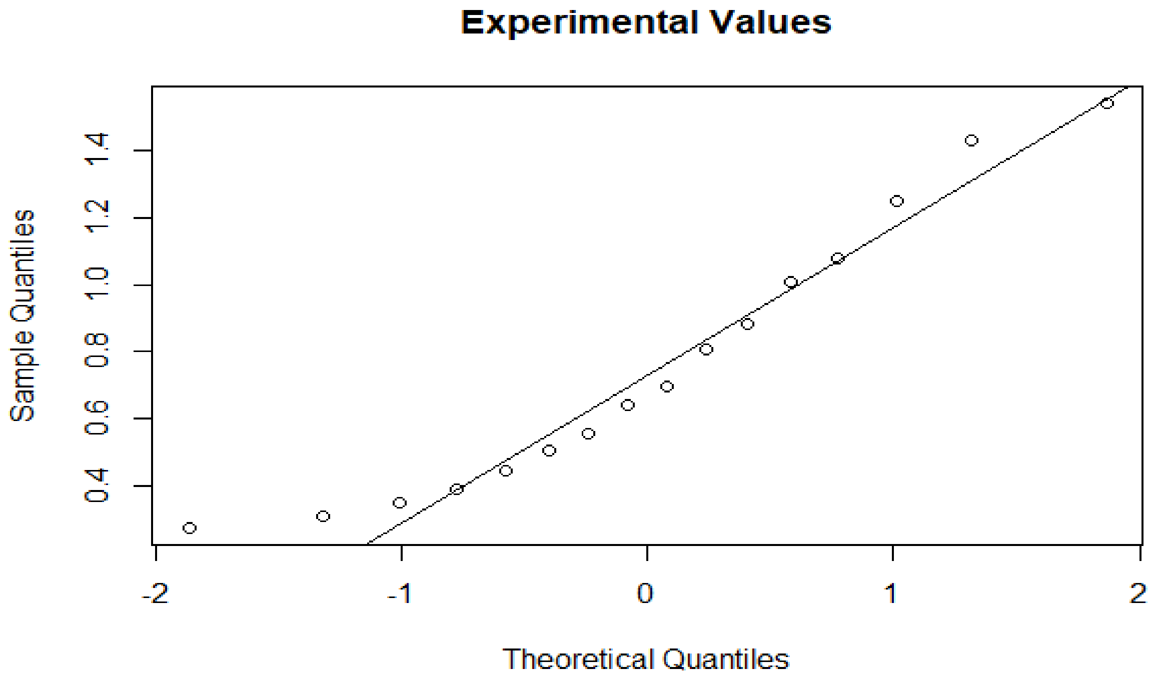
The two distributions seem more likely Uniform. However, because the sample size is small such a concrete assumption cannot be made accurately.

```
# QQ plot for theoretical values
qqnorm(vapdata$theoretical,main = "Theoretical Values")
qqline(vapdata$theoretical)

# QQ plot for experimental values
qqnorm(vapdata$experimental, main = "Experimental Values")
qqline(vapdata$experimental)
```

The following QQ plots were obtained.





On observing both the distributions, both seem more closer to normal distribution. However, we cannot conclude this due to smaller sample size.

Finally, it would be better to plot side by side boxplots and get summary statistics to determine the distribution.

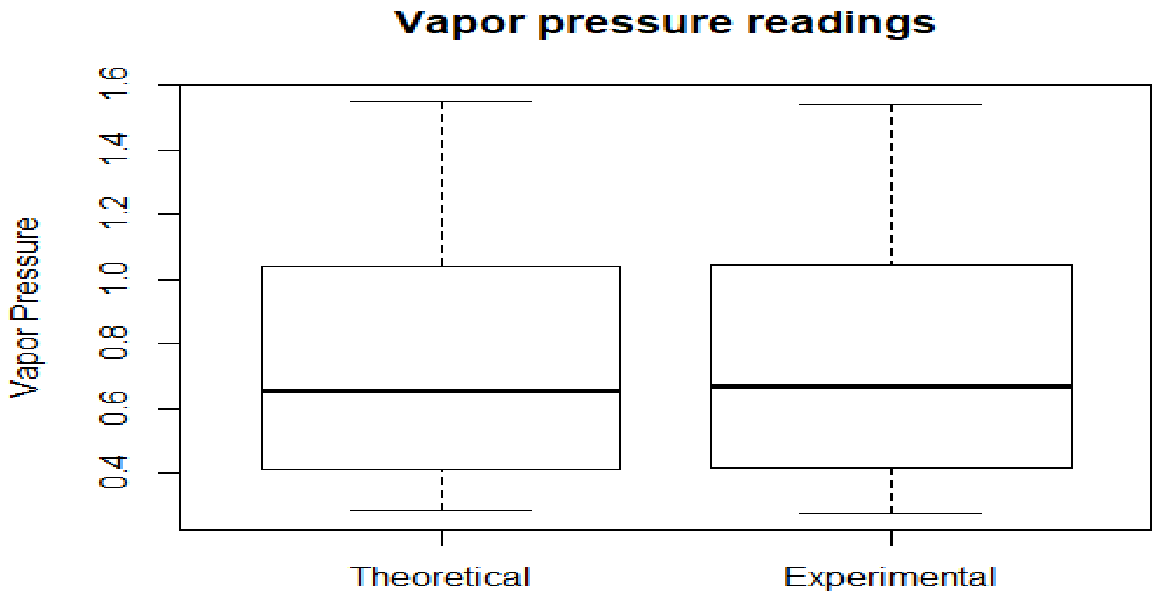
```
# side by side boxplots for theoretical and experimental values
boxplot(vapdata$theoretical,vapdata$experimental, main ="Vapor pressure readings", names = c("Theoretical", "Experimental"), ylab = "Vapor Pressure")
```

```
# Summary statistics of theoretical values
Summary(vapdata$theoretical)

Min.     1st Qu.  Median    Mean    3rd Qu.    Max.
0.2820  0.4175   0.6555     0.7606  1.0250     1.5500
```

```
# Summary statistics of experimental values
Summary(vapdata$experimental)

Min.     1st Qu.  Median    Mean    3rd Qu.    Max.
0.2760  0.4305   0.6675     0.7599  1.0275     1.5400
```



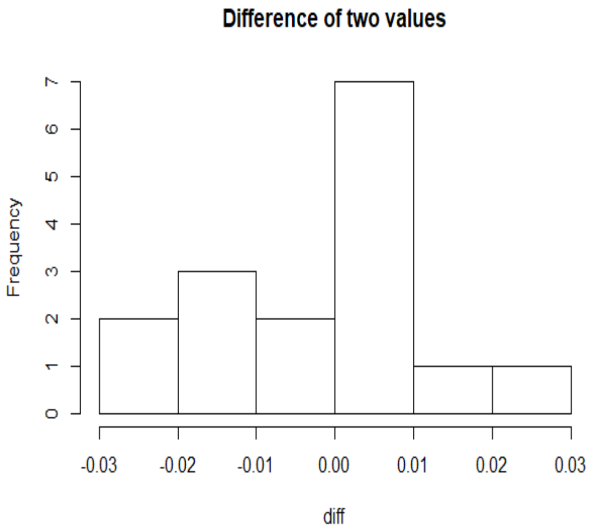
Observing the above values and the side by side boxplots, we find both the mean values are very close and there is very small difference. The variance of theoretical and experimental values is also nearly equal. Also, the medians differ by about 0.01. We can assume equal variances as both sample sizes are equal.

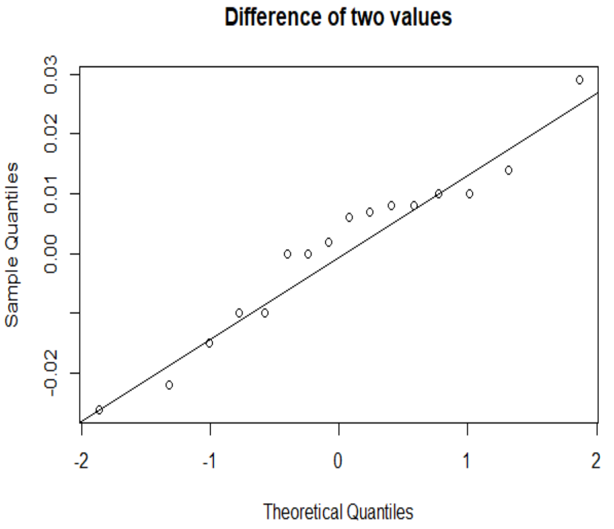
We do not observe any outliers in both plots and the data distributions seem similar. With N = 17 we cannot definitively conclude that the distributions are normal, hence nonparametric bootstrap will be used to determine the Confidence Interval for the given data. The difference of the two populations at each temperature will be used.

The histogram and QQ plot of the difference of theoretical and experimental values are plotted below.

```
# obtain difference between Theoretical values and Experimental values
diff = vapdata$theoretical - vapdata$experimental
```

```
# Histogram and QQ plot for difference
hist(diff, main = "Difference of two values")
qqnorm(diff, main = "Difference of two values")
qqline(diff)
```





The Confidence Interval is constructed using nonparametric bootstrap.

Define a function that would calculate mean of difference values.

```
# construct a confidence interval using non-parametric bootstrap
# define function to calculate mean

mean.npar = function(x,indices){
  result = mean(x[indices])
  return(result)
}
```

Calculate 95% Confidence Interval.

```
mean.npar.boot = boot(data = diff, mean.npar, R = 999, sim = "ordinary", stype = "i")

# calculate 95% confidence interval using percentile bootstrap
boot.ci(mean.npar.boot,conf = 0.95, type = "perc")
```

The following output was obtained.

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates
```

```
CALL :
boot.ci(boot.out = mean.npar.boot, conf = 0.95, type = "perc")

Intervals :
Level     Percentile
95%   (-0.0064,  0.0074 )
Calculations and Intervals on Original Scale
```

Therefore, based on our assumptions, the Confidence Interval obtained was [-0.0064, 0.0074]. Now, as 0 lies in this interval, we may conclude that the theoretical vapor pressure is a good model of reality.

Section – 2
R – Code

Q(1)

```
# Question 1
library(boot)

# read "gpa.csv" file
gpacsv = read.csv("D:/Users/rao29/Documents/Sem 5/Stats for Data Science/MP4/gpa.csv", header = TRUE, sep = ',')

# plot scatter plot of gpa against act
plot(gpacsv$gpa, gpacsv$act, main = "Scatterplot of GPA vs ACT", xlab = "GPA", ylab = "ACT", pch=20)

# point estimate of population correlation(rho) between GPA and ACT
rhop = cor(gpacsv$gpa,gpacsv$act)
print(paste("Point estimate of correlation between GPA and ACT = ", rhop))

# define function to find correlation between GPA and ACT
corr.npar = function(x,indices){
  result = cor(gpacsv$gpa[indices], gpacsv$act[indices])
  return(result)
}

# calculate point estimate, bias and standard error values
(corr.npar.boot = boot(data = gpacsv, corr.npar, R = 999, sim = "ordinary", stype = "i"))

# calculate 95% confidence interval using percentile bootstrap
boot.ci(corr.npar.boot,conf = 0.95, type = "perc")
```

Q(2)

```
# Question 2

# read "VOLTAGE.csv" file
data = read.csv("D:/Users/rao29/Documents/Sem 5/Stats for Data Science/MP4/VOLTAGE.csv", header = TRUE, sep = ",")

# obtain remote and local values separately
remoteVal = subset(data, location == "0")
localVal = subset(data, location == "1")


# Question 2(a)
# plot side by side boxplots of voltage distributions at both locations
boxplot(remoteVal$voltage, localVal$voltage, main = "Voltage Readings", names = c("Remote", "Local"), xlab = "Location", ylab = "Voltage")

# get summary statistics for voltage at remote location
summary(remoteVal$voltage)
IQR(remoteVal$voltage)
sd(remoteVal$voltage)

# get summary statistics for voltage at local location
summary(localVal$voltage)
IQR(localVal$voltage)
sd(localVal$voltage)

# obtain normal QQ plots
par(mfrow = c(1,2))
qqnorm(remoteVal$voltage, main = "Remote")
qqline(remoteVal$voltage)
qqnorm(localVal$voltage, main = "Local")
qqline(localVal$voltage)

# Question 2(b)
# construct appropriate confidence interval with assumption
# consider 95% CI, alpha = 0.05

alpha = 1-0.95

# mean of voltages at remote
(remoteVal.mean = mean(remoteVal$voltage))
```

```
# mean of voltages at local
(localVal.mean = mean(localVal$voltage))

# variance of voltages at remote
(remoteVal.var = var(remoteVal$voltage))

# variance of voltages at local
(localVal.var = var(localVal$voltage))

# number of remote locations
(n = nrow(remoteVal))

# number of local locations
(m = nrow(localVal))

# CI for difference in means
diff_means = remoteVal.mean - localVal.mean + c(-1,1)*qnorm(1-(alpha/2))*sqrt((remoteVal.var/n) +
(localVal.var/m))
diff_means
```

Q(3)

```
# Question 3
library(boot)
set.seed(123)

# read "VAPOR.csv" file
vapdata = read.csv("D:/Users/rao29/Documents/Sem 5/Stats for Data Science/MP4/VAPOR.csv",
header = TRUE, sep = ",")

# histogram for theoretical and experimental values
hist(vapdata$theoretical, main = "Histogram of Theoretical Values", xlab = "Theoretical Values")
hist(vapdata$experimental, main = "Histogram of Experimental Values", xlab = "Experimental Values")

# QQ plot for theoretical values
qqnorm(vapdata$theoretical,main = "Theoretical Values")
qqline(vapdata$theoretical)

# QQ plot for experimental values
qqnorm(vapdata$experimental, main = "Experimental Values")
qqline(vapdata$experimental)
```

```
# side by side boxplots for theoretical and experimental values
boxplot(vapdata$theoretical,vapdata$experimental, main ="Vapor pressure readings", names = c("Theoretical", "Experimental"), ylab = "Vapor Pressure")

# Summary statistics of theoretical values
summary(vapdata$theoretical)

# Summary statistics of experimental values
summary(vapdata$experimental)

# obtain difference between Theoretical values and Experimental values
diff = vapdata$theoretical - vapdata$experimental

# Histogram and QQ plot for difference
hist(diff, main = "Difference of two values")
qqnorm(diff, main = "Difference of two values")
qqline(diff)

# construct a confidence interval using non-parametric bootstrap
# define function to calculate mean
mean.npar = function(x,indices){
  result = mean(x[indices])
  return(result)
}

(mean.npar.boot = boot(data = diff, mean.npar, R = 999, sim = "ordinary", stype = "i"))

# calculate 95% confidence interval using percentile bootstrap
boot.ci(mean.npar.boot,conf = 0.95, type = "perc")
```