Mini Project #4

Names of group members: (Group-18)

Deepika Mamidipelly

Preethi Pasunuri

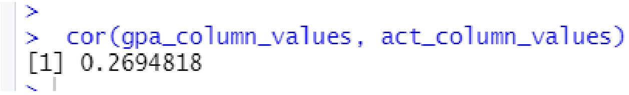
Contribution of each group member

Both group members contributed equally to the inputs for both questions and best is chosen among them to solve these problems. Collaboratively learned R, ran the scripts, and assessed the results. Some of the scripts written by Deepika were analyzed and finalized by Preethi and similarly scripts written by Preethi were assessed and finalized by Deepika. Both group members distribute equal amounts of report documentation, which is then integrated into a single final document. Members of the group worked diligently to meet all the project criteria

Question-1

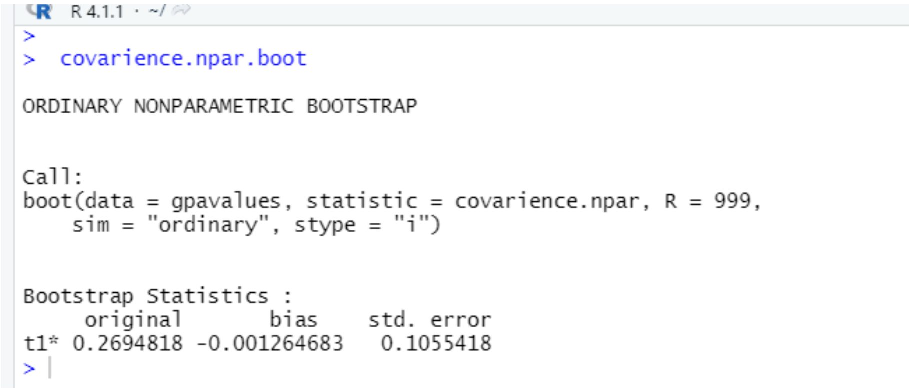
RCode:

```
gpa_values = read.csv("D:\\classes\\Statistics\\Assignments\\gpa.csv")
gpa_column_values <- as.numeric(gpa_values$gpa)
act_column_values <- as.numeric(gpa_values$act)
plot(gpa_column_values, act_column_values, main="GPA and ACT Scatterplot", xlab =
"GPA_Values", ylab = "ACT_Values")
abline(lm(act_column_values~gpa_column_values))
 cor(gpa_column_values, act_column_values)
```



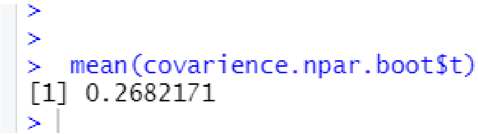
```
>
>  cor(gpa_column_values, act_column_values)
[1] 0.2694818
>
```

```
library(boot)
covarience.npar <- function(gpaset, indices){
 xgpa <- gpaset$gpa[indices]
 xact <- gpaset$act[indices]
 result <- cor(xgpa, xact)
 return(result)
 }
 covarience.npar.boot <- boot(gpavalues,covarience.npar, R = 999, sim = "ordinary", stype =
"i")
 covarience.npar.boot
```



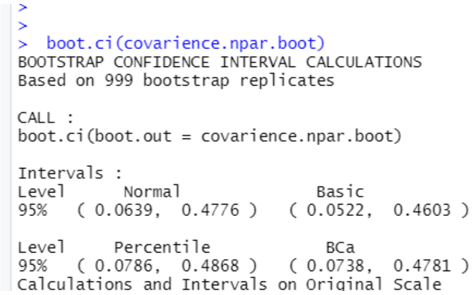
```
>
>  covarience.npar.boot

ORDINARY NONPARAMETRIC BOOTSTRAP


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


Bootstrap Statistics :
     original       bias     std. error
t1* 0.2694818 -0.001264683   0.1055418
>
```

mean(covarience.npar.boot$t)



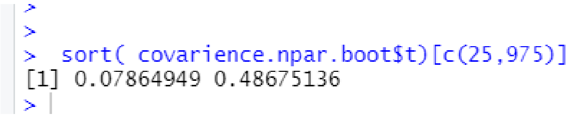
```
>
>
>  mean(covarience.npar.boot$t)
[1] 0.2682171
>
```

boot.ci(covarience.npar.boot)



```
>
>
>  boot.ci(covarience.npar.boot)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

CALL :
boot.ci(boot.out = covarience.npar.boot)

Intervals :
Level      Normal              Basic
95%   ( 0.0639,  0.4776 )   ( 0.0522,  0.4603 )

Level     Percentile            BCa
95%   ( 0.0786,  0.4868 )   ( 0.0738,  0.4781 )
Calculations and Intervals on Original Scale
```

sort(covarience.npar.boot$t)[c(25,975)]

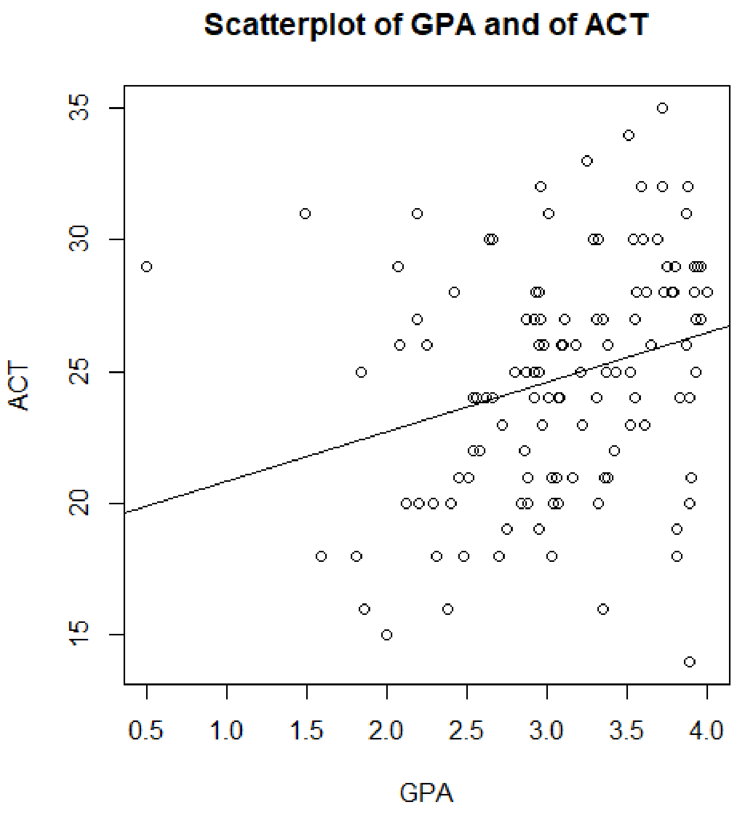


```
>
>
>  sort( covarience.npar.boot$t)[c(25,975)]
[1] 0.07864949 0.48675136
>
```

Observations:

First read the data file and two column values gpa,act are extracted and stored into different variables(gpa_column_values and act_column_values). Once the two datasets are separated , scatter plots are drawn for them. abline is implemented to see the correlation between the two datasets.

Scatter plot that gets generated:



From the graph, it can be concluded that the line that is drawn in the scatter plot has a positive slope greater than zero. This indicates that there is a positive association amongst the gpa and act. This would mean that the strength of the linear relationship is weak.

Next,we used cor function in R to find the correlation between two datasets(gpa and act). The R output of the datsets for correlation s is…. Based on the given datasets, the correlation ended up being 0.2694818.

Then the boot function is used in order to resample and find estimates for the correlation. Then, statistical functions are made in order to calculate correlation using the cor function once again which is returned. The expected value t* from bootstrap samples is also used as a point estimate. Values returned from the functions for the data is:

Estimate = 0.2694818, Bias = -0.001264683 and Standard Error = 0.1055418.

To obtain the confidence interval, the boot.ci function is used. The value obtained after running the code snippets is: (0.0786, 0.4868). Then the bootstrap correlation is sorted and the 1st and 3rd quartiles result in (0.07864949 0.48675136) which verifies that the confidence interval is correct.

The point estimate of correlation from bootstrap was interpreted to be approximately close to the correlation value from the samples. Furthermore, the confidence interval from boot.ci is nearly identical to the quantile values from sorted bootstrap data. Finally, the correlation value is around 0.3, indicating a positive association in the scatter plot.

Question-2:

#R code for Question 2

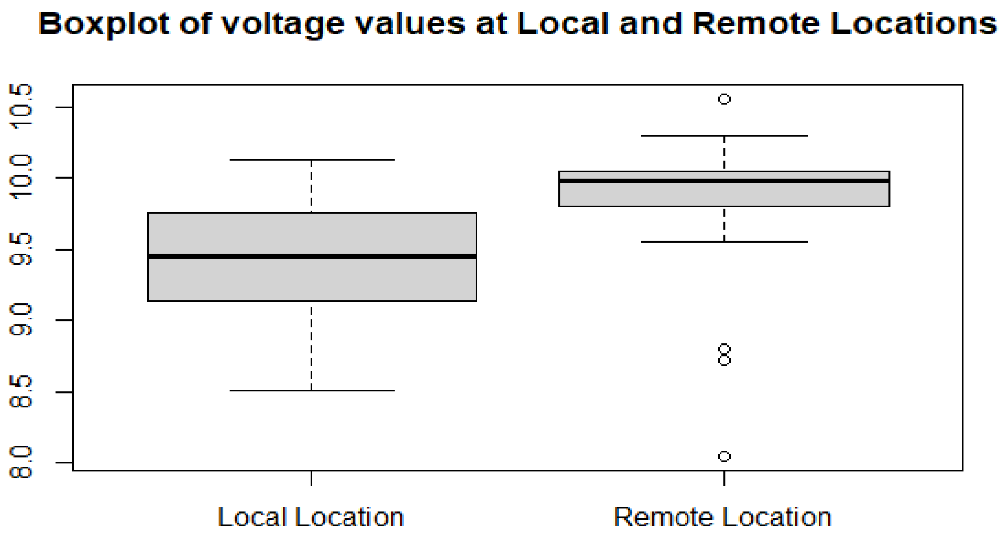
```
> #Read data from csv
> voltage <- read.csv('D:\\SUB_CS_6313\\Assignments\\VOLTAGE.csv')
> #Separate datasets based upon location
> voltage.remote = voltage$voltage[which(voltage$location == 0)]
> voltage.local = voltage$voltage[which(voltage$location == 1)]
> #Draw boxplots and summary of datasets
> boxplot(voltage.local, voltage.remote, names  = c("Local Location", "Remote Location"), main = "Boxplot of voltage values at Local and Remote Locations", range = 1.5)
>
> summary(voltage.remote)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.050   9.800   9.975   9.804  10.050  10.550
> summary(voltage.local)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.510   9.152   9.455   9.422   9.738  10.120
>
> #Draw qqplots for datasets
> par(mfrow = c(1,2))
> qqnorm(voltage.local, main = "Local")
> qqline(voltage.local)
> qqnorm(voltage.remote, main = "Remote")
> qqline(voltage.remote)
> #Calculate mean, variance, standard error and confidence interval > var(voltage. local)
> var(voltage.remote)
[1] 0.2925895
>
> se <- sqrt(var(voltage.local)/30 + var(voltage.remote)/ 30)
> se
[1] 0.1318979
>
> diff = mean(voltage.remote) - mean(voltage.local)
> diff + c(-1,1) * qnorm(0.975) * 0.1318979
[1] 0.1228182 0.6398485
>
> #Calculate confidence interval using t test
> t.test(voltage.remote, voltage.local, alternative = "two.sided", paired = FALSE,var.equal = FALSE, conf.level = 0.95)
```

```
Welch Two Sample t-test

data:  voltage.remote and voltage.local
t = 2.8911, df = 57.16, p-value = 0.005419
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1172284 0.6454382
sample estimates:
mean of x mean of y
 9.803667  9.422333
```

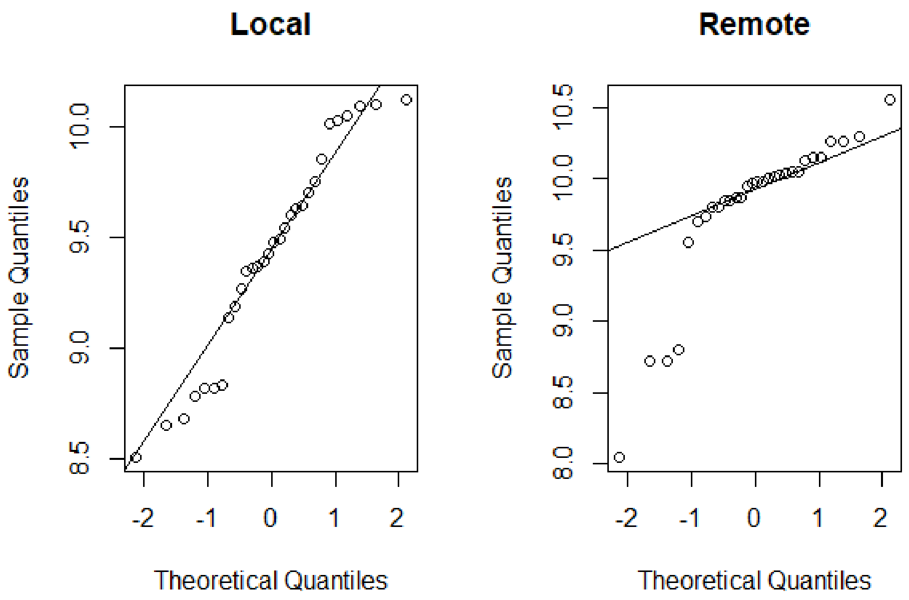
Observation:

The data is first read in and separated into two separate variables based on location.

a) Boxplots are examined to compare the two distributions:



The voltage readings at remote locations are notably higher than those at local locations. With the 5-point summary, it is clear that both graphs are skewed to the left because the medians are greater than the mean. Outliers can also be clearly seen in the remote location graph. QQplots for both datasets:



It can be seen that for some values, the data points and line coincide, implying that the data sets are normalized.

b) It is assumed that the manufacturing process will be established locally if there is no difference between the population means. As a result, the null hypothesis is: Difference = 0 => sample mean of remote - sample mean of local = 0 And the Alternative Hypothesis would be: Difference!= 0 => remote sample mean - local sample mean!= 0

To begin, the two samples must be viewed as independent entities. Based on the graphed QQ plots, it is assumed that the data is normal. Now, because the IQR are vastly different, assuming that population variances are equal cannot be assumed. As a result, Satterthwaite's approximation and t-distributions must be performed. Calculating the variance and mean:

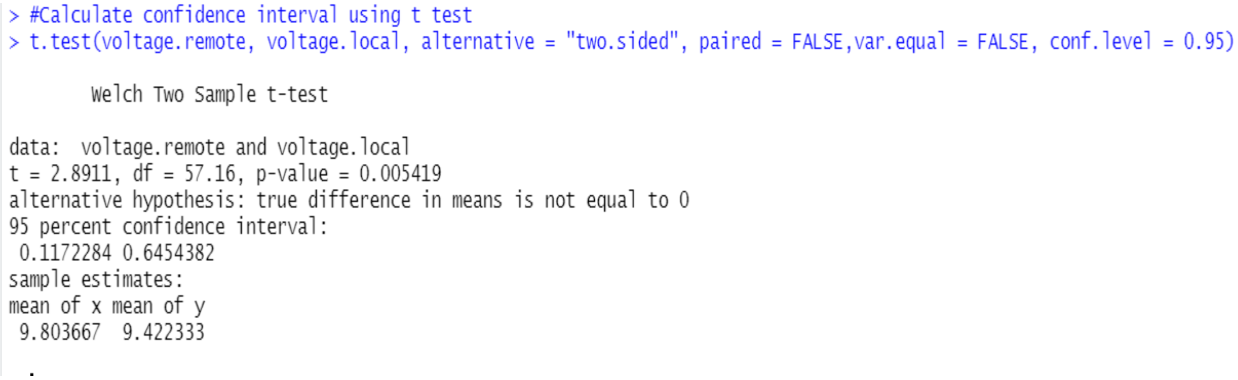
$$\bar{r} - \bar{l} = 0.3813333$$
$$S_r^2 = 0.2925895$$
$$S_l^2 = 0.229322$$
$$\widehat{SE}(\bar{r} - \bar{l}) = \sqrt{\left(\frac{S_r^2}{n_r}\right) + \left(\frac{S_l^2}{n_l}\right)}$$
$$= \sqrt{\frac{0.2925895}{30} + \frac{0.229322}{30}}$$
$$= \sqrt{\frac{0.5219115}{30}}$$
$$= 0.1318979$$

It is known that the 95% confidence interval for the Z value is 1.96, so calculating the confidence interval:

$$\text{Lower Bound: } (\bar{r} - \bar{l}) - Z_{\frac{\alpha}{2}} \times \widehat{SE}(\bar{r} - \bar{l})$$
$$: 0.3813333 - 1.96 * 0.1318979$$
$$: 0.1228182$$
$$\text{Upper Bound: } (\bar{r} - \bar{l}) + Z_{\frac{\alpha}{2}} \times \widehat{SE}(\bar{r} - \bar{l})$$
$$: 0.3813333 + 1.96 * 0.1318979$$
$$: 0.6398485$$

The confidence interval we calculated is (0.1228182, 0.6398485)

A t test is used to validate the confidence interval. The values are determined by the t test result are (0.1172284, 0.6454382). As a result, the confidence interval is appropriate, and the normal assumptions hold. Because 0 does not fall within the confidence interval (as determined by the t test), the null hypothesis is rejected. This means that the difference in means between the two locations is not zero. As a result, the manufacturing process cannot be established in local locations.



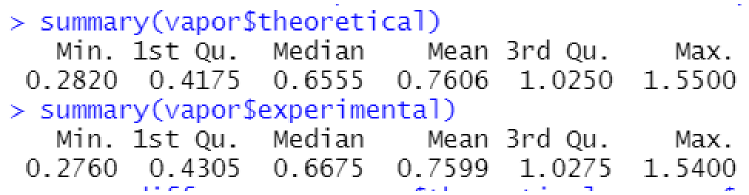
```
> #Calculate confidence interval using t test
> t.test(voltage.remote, voltage.local, alternative = "two.sided", paired = FALSE,var.equal = FALSE, conf.level = 0.95)

	Welch Two Sample t-test

data:  voltage.remote and voltage.local
t = 2.8911, df = 57.16, p-value = 0.005419
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1172284 0.6454382
sample estimates:
mean of x mean of y 
 9.803667  9.422333 
  .
```

c)Part a shows that voltage readings at remote are greater than those at local. It suffices to say that high voltage is required to power heavy equipment in any manufacturing process. As a result of parts A and B, it is clear that the manufacturing process must be located in a remote location.

Question-3:

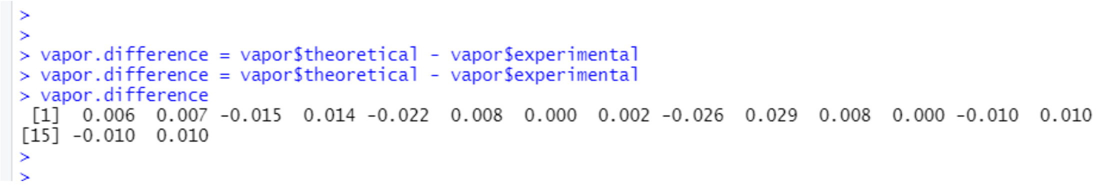
R-code:

```
vapor = read.csv("D:\\classes\\Statistics\\Assignments\\VAPOR.CSV")

par(mfrow=c(1,2))

qqnorm(vapor$experimental, main = "Experimental Values")
qqline(vapor$experimental)

qqnorm(vapor$theoretical, main = "Theoretical Values")
qqline(vapor$theoretical)
boxplot(vapor$theoretical, vapor$experimental, names = c("Theoretical", "Experimental"),
main = "Boxplot of Theoretical and Experimental Values")
```



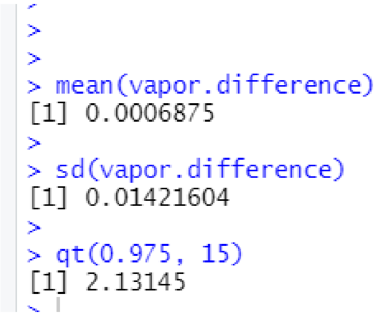
```
> summary(vapor$theoretical)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2820  0.4175  0.6555  0.7606  1.0250  1.5500 
> summary(vapor$experimental)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2760  0.4305  0.6675  0.7599  1.0275  1.5400 
```

```
vapor.difference = vapor$theoretical - vapor$experimental
vapor.difference
```



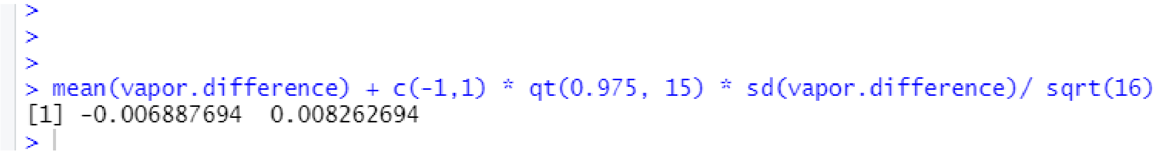
```
>
>
> vapor.difference = vapor$theoretical - vapor$experimental
> vapor.difference = vapor$theoretical - vapor$experimental
> vapor.difference
 [1]  0.006  0.007 -0.015  0.014 -0.022  0.008  0.000  0.002 -0.026  0.029  0.008  0.000 -0.010  0.010
[15] -0.010  0.010
>
>
```

```
mean(vapor.difference)
sd(vapor.difference)
qt(0.975, 15)
```



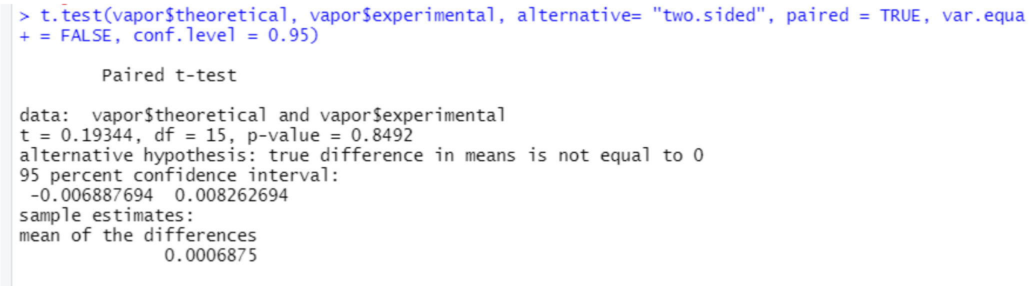
```
>
>
> mean(vapor.difference)
[1] 0.0006875
>
> sd(vapor.difference)
[1] 0.01421604
>
> qt(0.975, 15)
[1] 2.13145
```

mean(vapor.difference) + c(-1,1) * qt(0.975, 15) * sd(vapor.difference)/ sqrt(16)



```
>
>
>
> mean(vapor.difference) + c(-1,1) * qt(0.975, 15) * sd(vapor.difference)/ sqrt(16)
[1] -0.006887694  0.008262694
> 
```

t.test(vapor$theoretical, vapor$experimental, alternative= "two.sided", paired = TRUE, var.equal = FALSE, conf.level = 0.95)



```
> t.test(vapor$theoretical, vapor$experimental, alternative= "two.sided", paired = TRUE, var.equa
+ = FALSE, conf.level = 0.95)

        Paired t-test

data:  vapor$theoretical and vapor$experimental
t = 0.19344, df = 15, p-value = 0.8492
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.006887694  0.008262694
sample estimates:
mean of the differences
              0.0006875
```

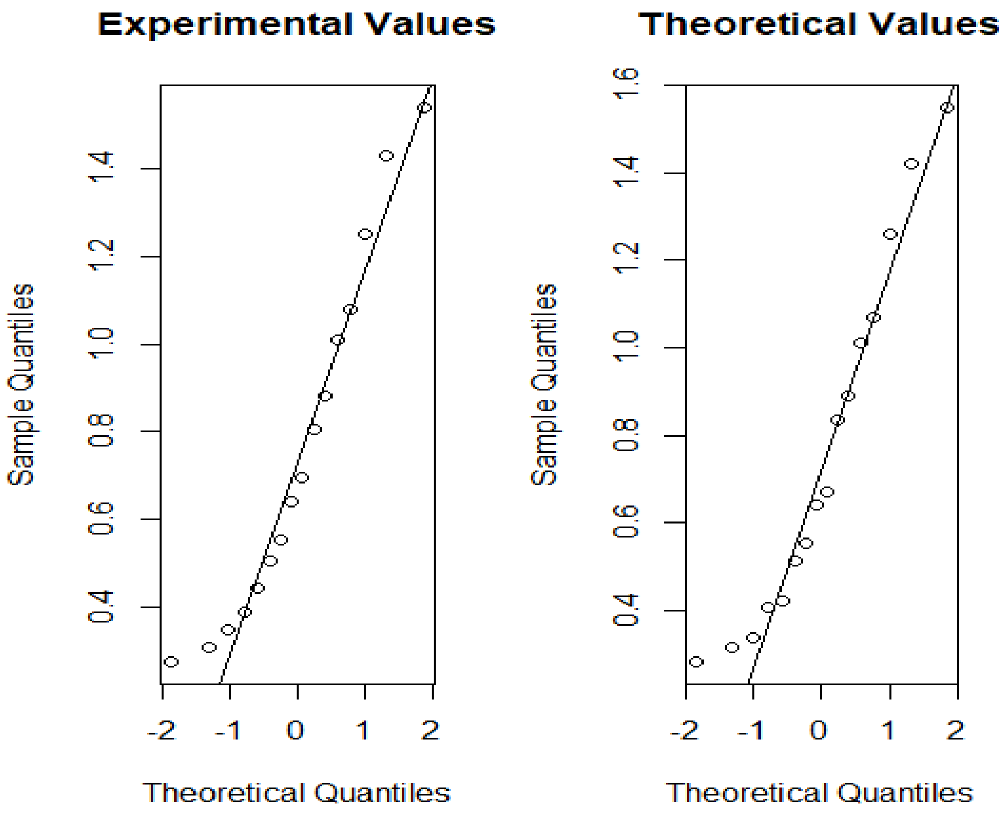
Observations:

First read the data file and column values are extracted and stored into different variables.

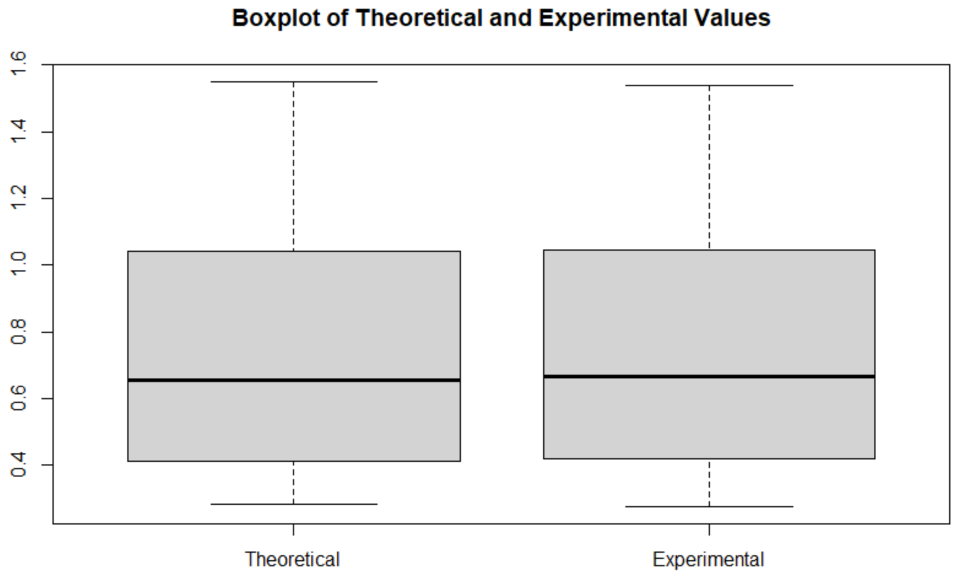
‘T’ is used to denote theoretical values and ‘E’ will be used to denote experimental values.

T(bar) represents the sample value of T used to estimate the population mean, and e(bar) represents the sample mean of E used to estimate the population mean.

QQplot’s for theoretical and experimental values:



As a result of the qqplots, it is clear that the samples are approximately normal. Theoretical and experimental values are plotted in a boxplot:



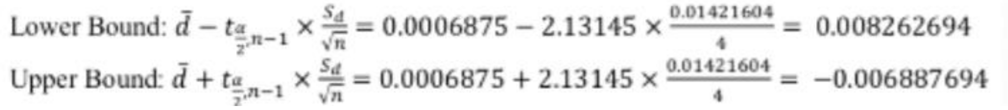
The two datasets are clearly very similar, and the differences are almost insignificant. This conclusion is supported by the IQR and 5-plot summary. Both distributions are skewed to the right because their mean is greater than their median. Let us now examine the mean difference between theoretical and experimental values.

Null Hypothesis: True mean difference between t(bar) and e(bar) == 0

Alternative Hypothesis: True mean difference between t(bar) and e(bar) != 0.

Now, the confidence interval is calculated using the t distribution.

Calculating the mean, standard dev results in: mean = 0.0006875, standard dev = 0.01421604, t = 2.13145



Lower Bound: $\bar{d} - t_{\frac{\alpha}{2},n-1} \times \frac{S_d}{\sqrt{n}} = 0.0006875 - 2.13145 \times \frac{0.01421604}{4} = 0.008262694$

Upper Bound: $\bar{d} + t_{\frac{\alpha}{2},n-1} \times \frac{S_d}{\sqrt{n}} = 0.0006875 + 2.13145 \times \frac{0.01421604}{4} = -0.006887694$

In order to verify the confidence interval a t test is conducted.

The observed interval is (--0.006887694 0.008262694). This means that the interval is appropriate. Since the value 0 lies within the found interval, it means that the t(bar) - e(bar) = 0.

So, the null hypothesis is accepted , so the true mean difference of theoretical and experimental values is zero. This is also supported by the boxplot.