CS 6313.001

Statistical Methods for Data Science

By

Prof. Min Chen

MINI PROJECT-4
DUO GROUP-41

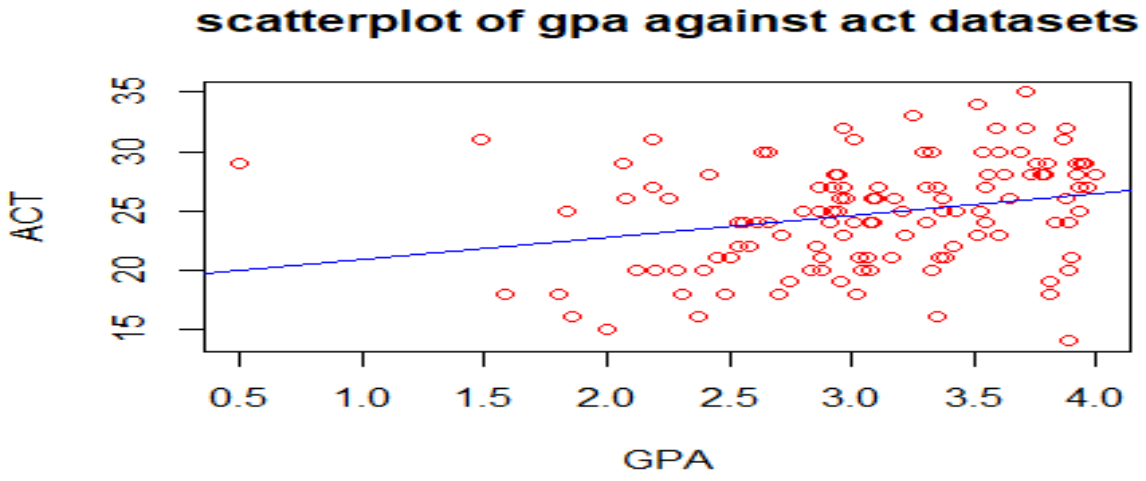
1. Amit Kumar AXK210047
2. Vedant Paresh Shah VXS200021

Contribution of each group member: Both worked together and finished the questions as instructed. First went through all the details required, followed the class Note and textbook, practiced the R-concept, then wrote down the scripts. Both of us worked efficiently to complete the required project and finished it on time.

Question-1: In the class, we talked about bootstrap in the context of one-sample problems. But the idea of nonparametric bootstrap is easily generalized to more general situations. For example, suppose there are two dependent variables X1 and X2 and we have i.i.d. data on (X1, X2) from n independent subjects. In particular, the data consist of (Xi1, Xi2), i = 1, . . . , n, where the observations Xi1 and Xi2 come from the ith subject. Let θ be a parameter of interest — it's a feature of the distribution of (X1, X2). We have an estimator ˆθ of θ that we know how to compute from the data. To obtain a draw from the bootstrap distribution of ˆθ, all we need to do is the following: randomly select n subject IDs with replacement from the original subject IDs, extract the observations for the selected IDs (yielding a resample of the original sample), and compute the estimate from the resampled data. This process can be repeated in the usual manner to get the bootstrap distribution of ˆθ and obtain the desired inference. Now, consider the gpa data stored in the gpa.txt file available on eLearning. The data consist of GPA at the end of freshman year (gpa) and ACT test score (act) for randomly selected 120 students from a new freshman class. Make a scatterplot of gpa against act and comment on the strength of linear relationship between the two variables. Let ρ denote the population correlation between gpa and act. Provide a point estimate of ρ, bootstrap estimates of bias and standard error of the point estimate, and 95% confidence interval computed using percentile bootstrap. Interpret the results. (To review population and sample correlations, look at Sections 3.3.5 and 11.1.4 of the textbook. The sample correlation provides an estimate of the population correlation and can be computed using cor function in R.)

Section: 1

We uploaded the csv file gpa.csv on G-Drive and from there read the csv file. Once the data is read into gpaACTDataSet, we separated the contents of gpa and act into gpaData and actData. Based on these datasets a scatter plot is drawn. To visualize the correlation between the two datasets, abline function is used to add a straight line to the plot. Generated scatterplot is:



From the above attached scatterplot, it is evident that the line that is drawn in the scatterplot has a slope greater than zero, which says that there is a positive association between the gpa and act dataset. It would mean that the strength of the linear relationship is weak. Now, we used the cor() function to find the correlation of the two datasets. Using these datasets, the correlation value obtained is: 0.2694818. Now to resampling of data and correlation estimation we used the boot function. We created a statistical function to calculate correlation using cor() function. Point estimate is also taken as expected value t* from samples of bootstrap. The values we get from the functions are Estimates: 0.2706756, Bias: 0.001193824, std. error: 0.1062264. To find out the confidence interval (CI) we used boot.ci function. And the CI values is: [0.0734, 0.4815]. After that the bootstrap correlation is sorted and the 1st and 3rd quartiles resulted in (0.07340128 0.48150066). It verifies that the confidence interval is correct. We found that the point estimate of correlation from bootstrap is approximately close to the correlation value from the samples as well as the confidence interval from boot.ci is approximately close to the quantile values from sorted bootstrap data. Also, the correlation value is approximately 0.27 which states that there is a positive association in the scatter plot.

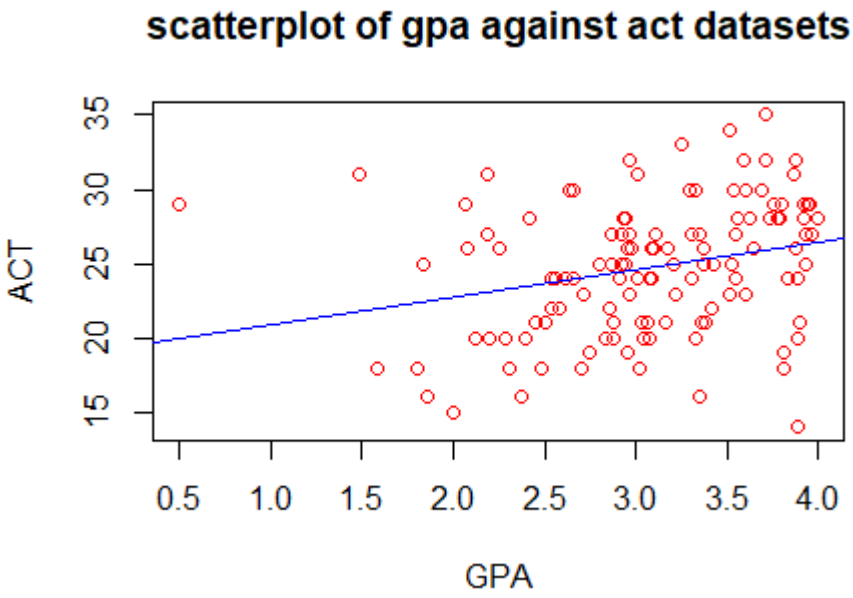
Section: 2

R-Code

```
> library(boot) # import the boot library

# Read the data from gpa.csv file
> id <- "1Ezp3FP1cDhZtbH3EDF_a9f6TM4YvudIW" # G-Drive file ID
gpaACTDataSet <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download", id))

# Data set summary
> summary(gpaACTDataSet)
```

```
      gpa              act       
 Min.   :0.500   Min.   :14.00  
 1st Qu.:2.689   1st Qu.:21.00  
 Median :3.078   Median :25.00  
 Mean   :3.074   Mean   :24.73  
 3rd Qu.:3.593   3rd Qu.:28.00  
 Max.   :4.000   Max.   :35.00  
```

```
# Separate the datset for GPA and ACT
> gpaData <- as.numeric(gpaACTDataSet$gpa)
> actData <- as.numeric(gpaACTDataSet$act)



# gpa and ACT plots
> plot(gpaData, actData, main= "scatterplot of gpa against act datasets", xlab = "GPA", ylab = "ACT",
col="red")
# Add Straight Lines to a Plot
> abline(lm(actData~gpaData), col="blue")
```



 # Correlation of gpaData & actData

```
> cor(gpaData, actData)
[1] 0.2694818
```

 # Etatistical function for correlation

```
> covarience.npar <- function(gpaset, indices){
+         xgpa <- gpaset$gpa[indices]
+         xact <- gpaset$act[indices]
+         result <- cor(xgpa, xact)
+         return(result)
+ }
```

#Execute boot function with statistical function

```
> covarience.npar.boot <- boot(gpaACTDataSet, covarience.npar,  R = 999, sim = "ordinary", stype ="i")
> covarience.npar.boot

ORDINARY NONPARAMETRIC BOOTSTRAP


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


Bootstrap Statistics :
     original      bias    std. error
t1* 0.2694818 0.003235887   0.1079692
```

```
> mean(covarience.npar.boot$t)
[1] 0.2723805
```

 # Getting confidence interval using boot.ci

```
> 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%   ( 0.0596,  0.4735 )   ( 0.0642,  0.4832 )

Level     Percentile            BCa
95%   ( 0.0558,  0.4747 )   ( 0.0419,  0.4614 )
Calculations and Intervals on Original Scale
```

 # Verifying confidence interval by extracting quantiles

```
> sort(covarience.npar.boot$t)[c(25, 975)]
[1] 0.05576736 0.47474205
```

Question-2: Consider the data stored in the file VOLTAGE.DAT on eLearning. These data come from a Harris Corporation/University of Florida study to determine whether a manufacturing process performed at a remote location can be established locally. Test devices (pilots) were set up at both the remote and the local locations and voltage readings on 30 separate production runs at each location were obtained. In the dataset, the remote and local locations are indicated as 0 and 1, respectively.

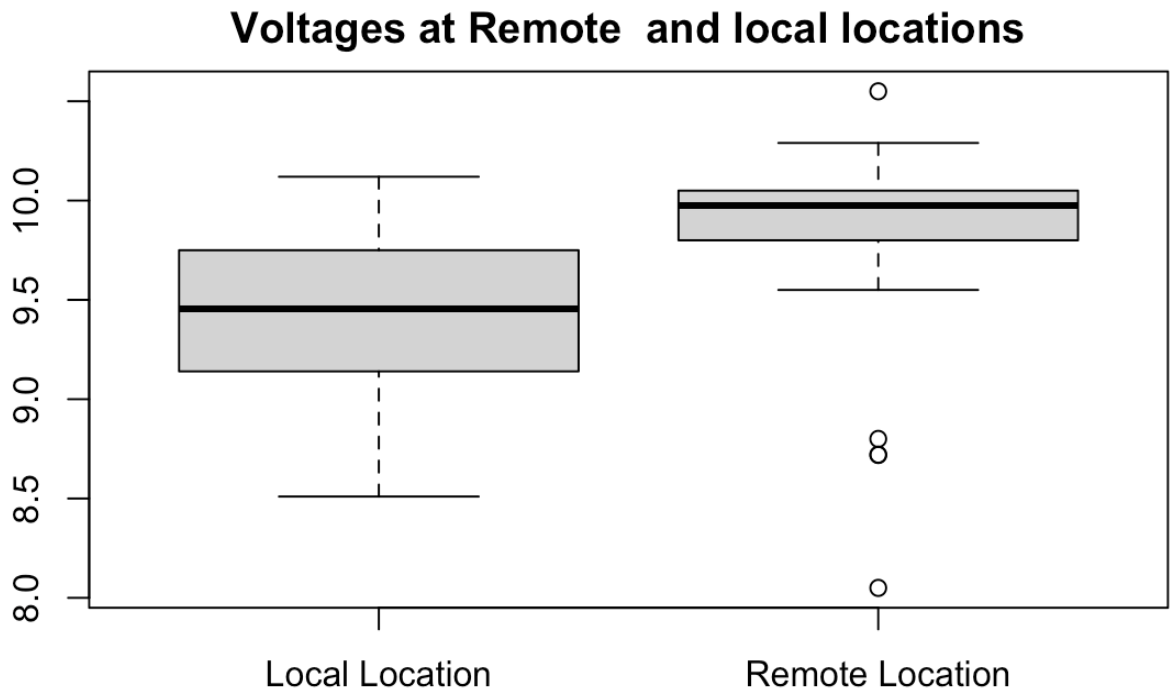
(a) (1 points) Perform an exploratory analysis of the data by examining the distributions of the voltage readings at the two locations. Comment on what you see. Do the two distributions seem similar? Justify your answer.

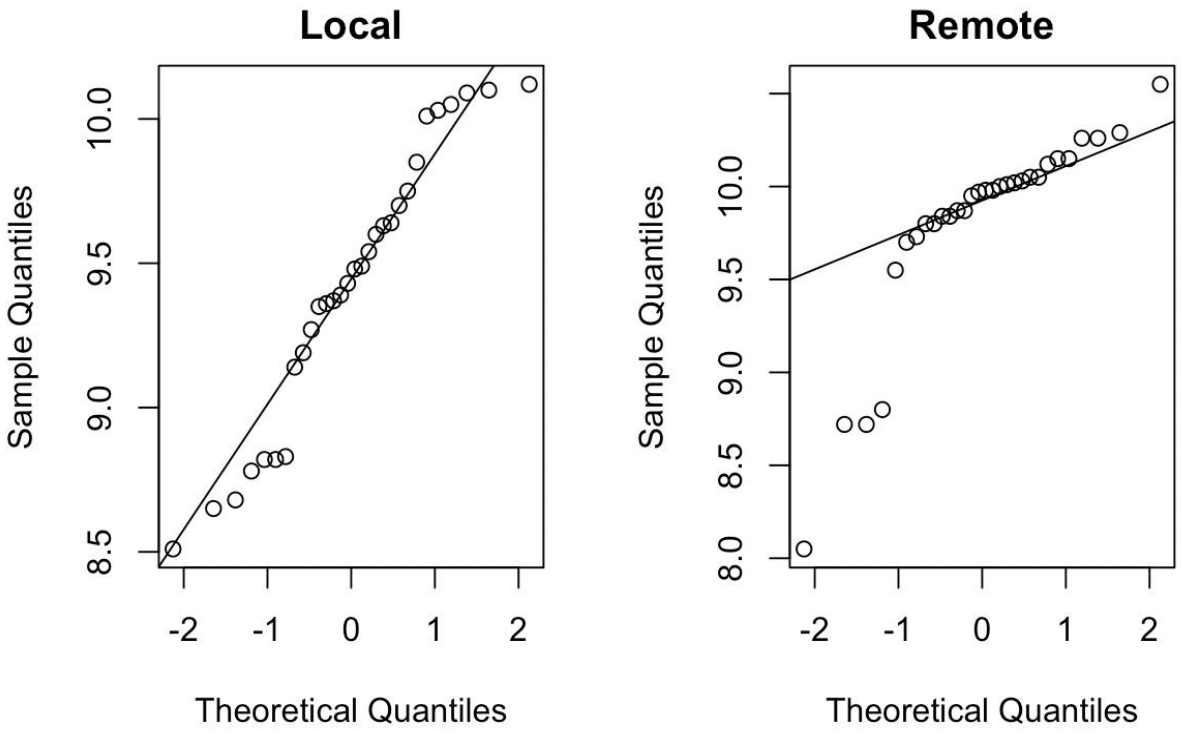
(b) (5 points) The manufacturing process can be established locally if there is no difference in the population means of voltage readings at the two locations. Does it appear that the manufacturing process can be established locally? Answer this question by constructing an appropriate confidence interval. Clearly state the assumptions, if any, you may be making and be sure to verify the assumptions.

(c) (1 point) How does your conclusion in (b) compare with what you expected from the exploratory analysis in (a)?

Section: 1

(a) We Downloaded the csv file VOLTAGE.csv on C-Drive and from there read the csv file. Once the data is read into voltage, we separated the contents of voltage into Remote and Local. Based on these datasets a box plot and scatter plot is drawn. To visualize the correlation between the two datasets, a line function is used to add a straight line to the scatter plot. Generated scatterplot is:





From both these plots we can see that the distribution of both these datasets is same and we can confirm that by also seeing the Summary of these datasets.

```
> summary(voltage.local)
   Min.  1st Qu.  Median    Mean   3rd Qu.    Max.
 8.510   9.152    9.455     9.422    9.738    10.120
> summary(voltage.remote)
   Min.  1st Qu.  Median    Mean   3rd Qu.    Max.
 8.050   9.800    9.975     9.804   10.050   10.550
```

As seen in the 5 point Summary both the datasets have a Median greater than the Mean which indicates that these datasets are left-skewed. Also, there are outliers that exist in the remote dataset which can be clearly seen in the qqplot.

It can also be observed that the both plot have some data points coinciding with the line. So hence it can be assumed that the both datasets are normalized.

(b) Here the question says that we can establish a local manufacturing process if there is no difference persists between population means of both datasets.

So first, we will treat both datasets as independent samples. Also, we have assumed that datasets have a normal distribution from the QQplots. Now as both datasets have distinct ranges, we will assume that they have unequal variance.

So, now we will use Satterwhite's Approximation and T-distribution for calculating the Confidence interval for the population mean. For convenience we have taken confidence interval of 95%.

Estimated Mean = Mean of Remote Location – Mean of Local Location = Mr - Ml =0.381333

Variance of Remote Location = Sr^2 = 0.2925895

Variance of Local Location = Sl^2 = 0.229322

Size of Remote Location Dataset = n = 30

Size of Local Location Dataset = n = 30

Standard Error of Estimated Mean = SDE =Sqrt((Sr^2)/n + (Sl^2)/m) = 0.1318979

We know that for 95% Confidence Interval we have Value of Z (qnorm) as 1.96, so now we will calculate the Confidence interval:

Upper Bound = Estimated Mean + qnorm(.975)*SDE = 0.6398484

Lower Bound = Estimated Mean - qnorm(.975)*SDE = 0.1228182

The Confidence Interval we have calculated is (0.1228182, 0.6398484)

To verify the Confidence interval, we use the T – distribution and we get the following Confidence Interval (0.1172284, 0.6454382)

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

This means that the difference between population means of 2 datasets is not zero. Which means that it rejects the null hypothesis that the manufacturing can be established at the local location. It thus means that the manufacturing process cannot be established at the Local location.

(c) It can be observed from (a) that the remote location the voltage readings are higher than that of local location and from (b) it can be observed that the manufacturing process cannot be established at the local location.

So, we can conclude from both (a) and (b) that manufacturing process can be done in remote location.

Section: 2

 # Reading the Voltage.csv file

```
voltage <- read.csv(file="/Users/vedantshah/Downloads/VOLTAGE.csv")
```

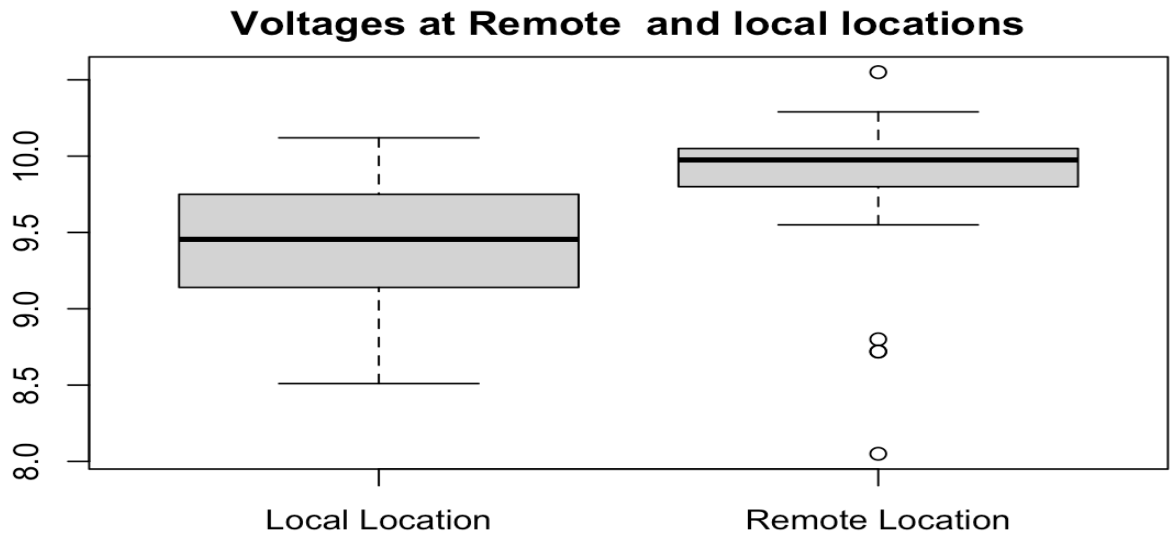
 # Getting the theoretical and experimental values

```
voltage.remote<-voltage$voltage[which(voltage$location==0)]

voltage.local<-voltage$voltage[which(voltage$location==1)]
```

#Boxplot

```
par(mfrow=c(1,1))

boxplot(voltage.local,voltage.remote,range=1.5,main="Voltages at Remote  and local locations",

names = c("Local Location", "Remote Location"))
```

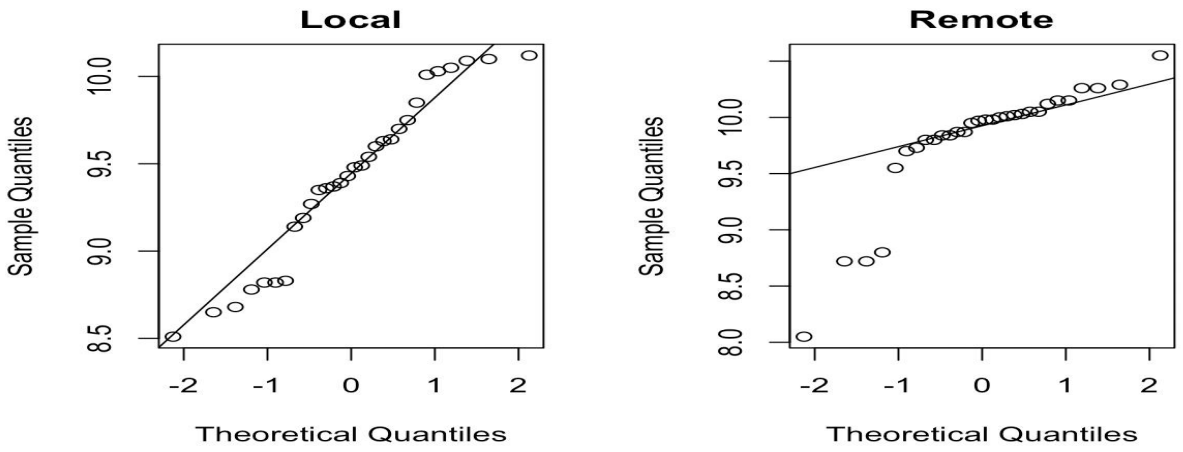
#Output:



#QQplots

```
par(mfrow=c(1,2))

qqnorm(voltage.local, main="Local")

qqline(voltage.local)

qqnorm(voltage.remote, main="Remote")

qqline(voltage.remote)
```

#Output:



#Summary

print("summary(voltage.local)")

summary(voltage.local)

print("summmary(voltage.remote)")

summary(voltage.remote)

#Output:

```
> #Summary
> print("summary(voltage.local)")
[1] "summary(voltage.local)"
> summary(voltage.local)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.510   9.152   9.455   9.422   9.738  10.120
>
> print("summmary(voltage.remote)")
[1] "summmary(voltage.remote)"
> summary(voltage.remote)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.050   9.800   9.975   9.804  10.050  10.550
```

#Calculating Confidence Interval with unequal Variance

print("var(voltage.local)")

var(voltage.local)

print("var(voltage.remote)")

var(voltage.remote)

print("Standard Error")

sde=sqrt(var(voltage.local)/30+var(voltage.remote)/30)

print(sde)

print("Estimated Mean")

```
est=mean(voltage.remote)-mean(voltage.local)
print(est)
print("Confidence Interval Range")
M=est+c(-1,1)*qnorm(0.975)*sde
print(M)
```

#Output:

```
[1] "var(voltage.local)"
[1] 0.229322
[1] "var(voltage.remote)"
[1] 0.2925895
[1] "Standard Error"
[1] 0.1318979
[1] "Estimated Mean"
[1] 0.3813333
[1] "Confidence Interval Range"
[1] 0.1228182 0.6398484
```

#Calculating Confidence Interval using t test with unequal Variance

```
t.test(voltage.remote, voltage.local, alternative="two.sided", paired= FALSE, var.equal= FALSE, conf.level=0.95 )
```

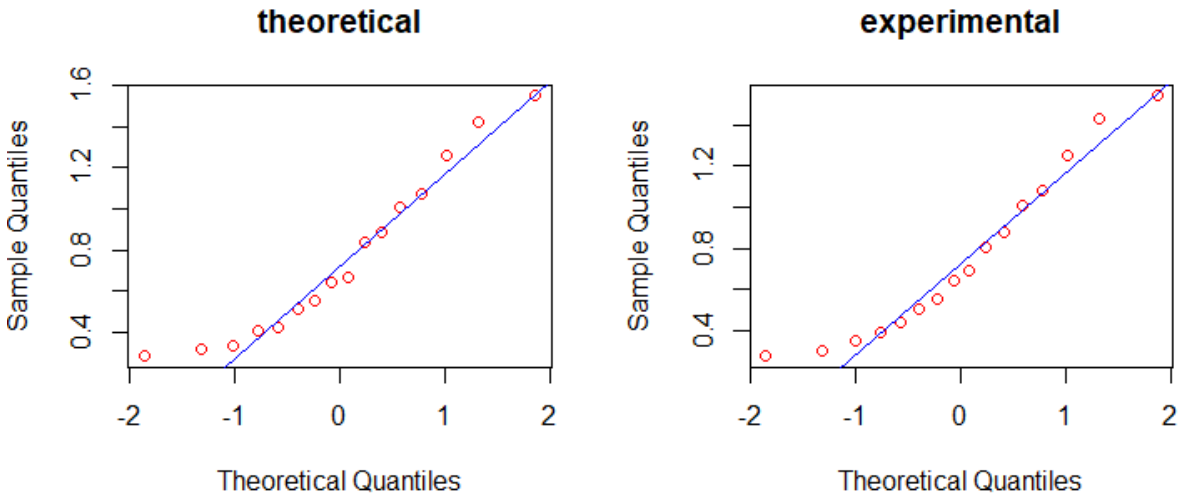
#Output:

```
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
```

Question-3: The file VAPOR.DAT on eLearning provide data on theoretical (calculated) and experimental values of the vapor pressure for dibenzothiophene, a heterocycloaromatic compound similar to those found in coal tar, at given values of temperature. If the theoretical model for vapor pressure is a good model of reality, the true mean difference between the experimental and calculated values of vapor pressure will be zero. Perform an appropriate analysis of these data to see whether or not this is the case. Be sure to justify all the steps in the analysis.

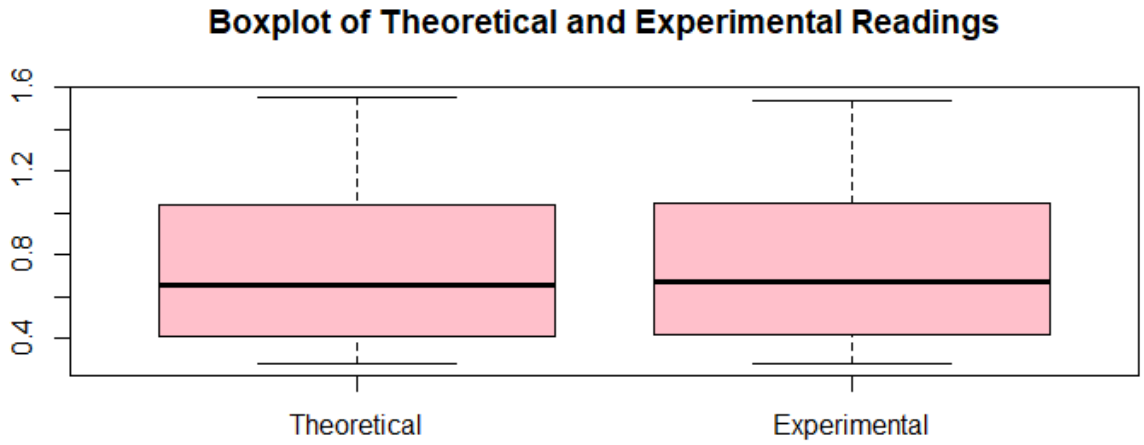
Section: 1

We uploaded the .dat file VAPOR.DAT on G-Drive and from there read the data. Once the data is read into vapourDataSet, we calculated the summary of the data using summary () function and a qqplot graph is drawn using qqnorm() and qqline() function based on theoretical and experimental value set.



From the above plot we can see sample data can be treated as approximately normal.

We also used the theoretical and experimental value set to draw a boxplot, which is attached below:



From the above plot, It is evident that the two datasets are very similar, and the differences are very minimal which can be ignored. The IQR, and 5-plot summary also supports it. Both distributions are right skewed, as the mean is slightly than the median (From summary () data attached in section 2).

Now, we calculated the mean value of difference between theoretical and experimental values.
Null Hypothesis: True mean difference between $\bar{T} - \bar{E}$ == 0
Alternative Hypothesis: True mean difference between $\bar{T} - \bar{E}$!= 0
The mean, standard dev results in:
mean = 0.0006875, standard dev = 0.01421604, t = 2.13145

Now, the confidence interval based on the t distribution is:

$Lower\ bound: \bar{d} - t_{\frac{\alpha}{2},\ n-1} * \frac{S_d}{\sqrt{n}}$ = 0.0006875 - 2.13145 * $\frac{0.01421604}{\sqrt{16}}$ = - 0.00688769461

$Upper\ bound: \bar{d} + t_{\frac{\alpha}{2},\ n-1} * \frac{S_d}{\sqrt{n}}$ = 0.0006875 + 2.13145 * $\frac{0.01421604}{\sqrt{16}}$ = 0.00826269461

In order to verify the result of confidence interval a t test was performed. The observed interval is (-0.006887694, 0.008262694). It proves that the interval is correct. We can see that the value 0 lies between the confidence interval we calculated, which means that the $\bar{T} - \bar{E}$ = 0. And hence, the null hypothesis is accepted. Which is also being supported by the boxplot.

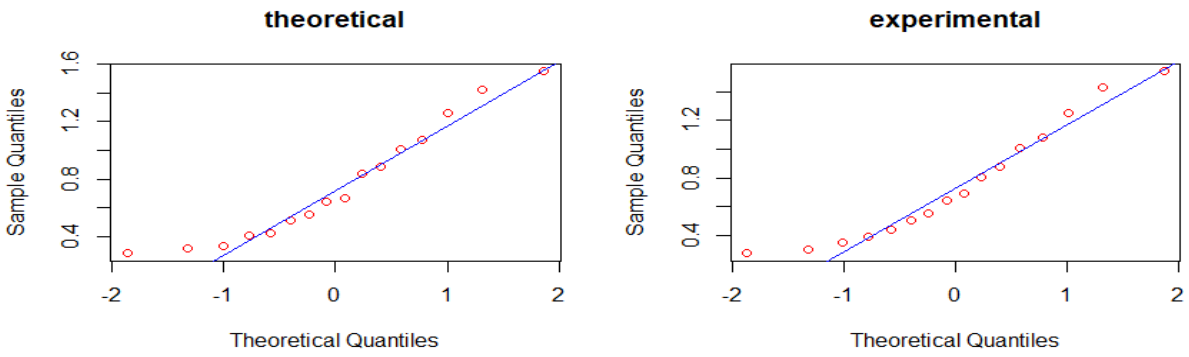
Section: 2

R-Code

```
# Read data from .dat file
> id <- "1ExMKDlVunMxC3w5bBx8ycyPS0S7DnUVe" # G-Drive file ID
> vapourDataSet <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download", id))
> summary(vapourDataSet)
```

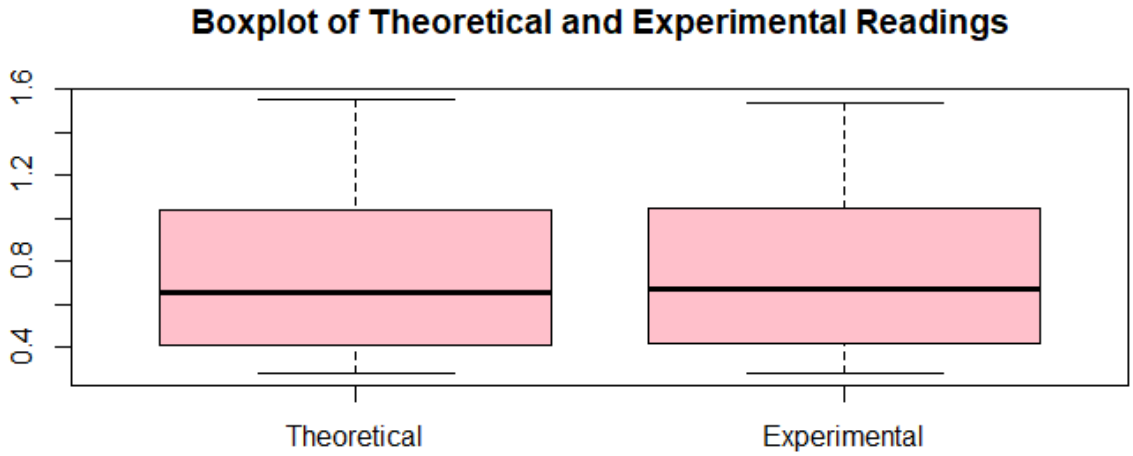
```
  temperature      theoretical      experimental
Min.   :100.6   Min.   :0.2820   Min.   :0.2760
1st Qu.:108.1   1st Qu.:0.4175   1st Qu.:0.4305
Median :116.0   Median :0.6555   Median :0.6675
Mean   :116.0   Mean   :0.7606   Mean   :0.7599
3rd Qu.:123.9   3rd Qu.:1.0250   3rd Qu.:1.0275
Max.   :131.8   Max.   :1.5500   Max.   :1.5400
```

```
# Draw qqplots based on the data loaded in vapourDataSet
> par (mfrow = c (1,2))
> qqnorm (vapourDataSet$theoretical, col = "red", main = "theoretical")
> qqline (vapourDataSet$theoretical, col = "blue")
> qqnorm (vapourDataSet$experimental, col = "red", main = "experimental")
> qqline (vapourDataSet$experimental, col = "blue")
```



Draw boxplots and summaries from dataset vapourDataSet

> boxplot (vapourDataSet$theoretical, vapourDataSet$experimental, names = c("Theoretical", "Experimental"), main = "Boxplot of Theoretical and Experimental Readings", col = "pink")



Mean, Standard deviation, t(n-1) val, and confidence interval

```
> vapourDataSet.difference = vapourDataSet$theoretical - vapourDataSet$experimental
> meanValue <- mean(vapourDataSet.difference)
> meanValue
[1] 0.0006875
> sdValue <- sd(vapourDataSet.difference)
> sdValue
[1] 0.01421604
> Tn_1Value <- qt(0.975, 15)
> Tn_1Value
[1] 2.13145
> CI_vapourDataSet <- meanValue + c(-1,1) * Tn_1Value * sdValue/ sqrt(16)
> CI_vapourDataSet
[1] -0.006887694 0.008262694
```

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

```
        Paired t-test

data:  vapourDataSet$theoretical and vapourDataSet$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
```