Project 4

Name of group members: 1. Ajay Danda (AXD180068) 2. Satyam Bhikadiya (SXB180124)

Contribution of each group member:

Ajay Danda:

- Equal Contribution in solving Q1
- Equal Contribution in solving Q2
- Equal Contribution in solving Q3
- Equal Contribution in Documenting the Report

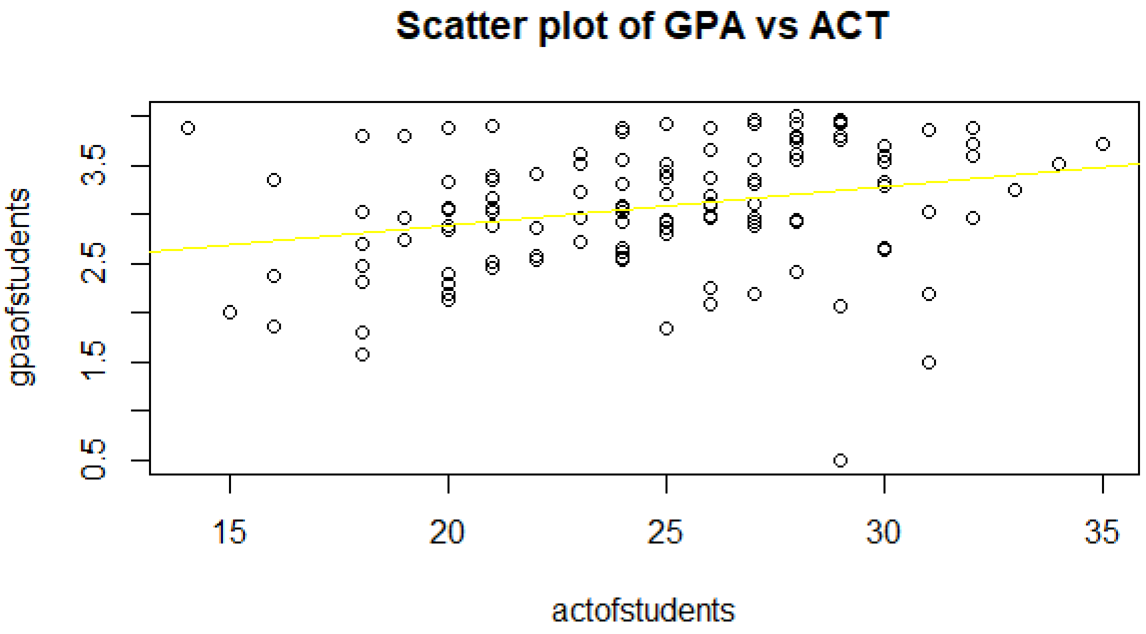
Satyam Bhikadiya:

- Equal Contribution in solving Q1
- Equal Contribution in solving Q2
- Equal Contribution in solving Q3
- Equal Contribution in Documenting the Report

<u>Question1:</u>

(6 points) 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.csv 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.)

<u>Answer:</u>

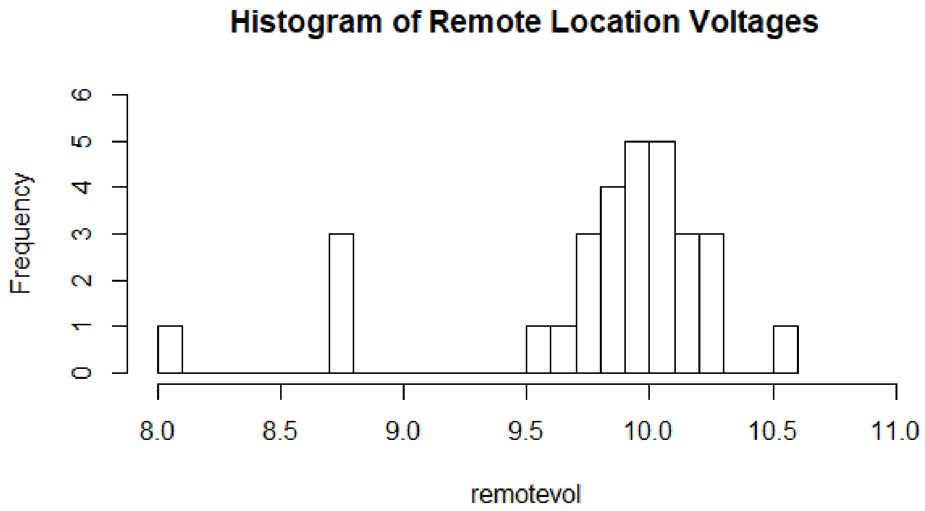


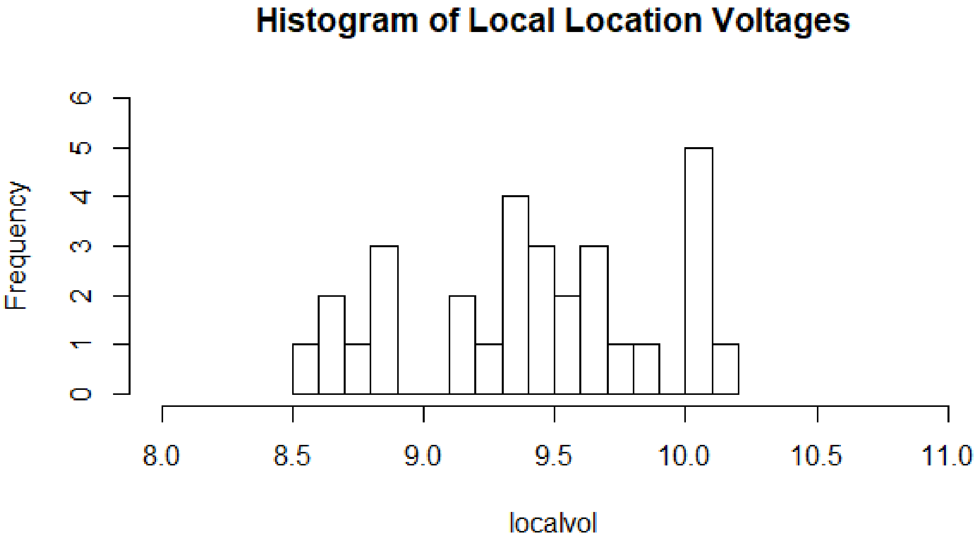
From the above scatterplot we can observe that, there lies ample points between GPA’s and ACT Scores which implies an indication about linearity amongst the two. The estimate of the population correlation is given by the sample correlation, sample correlation provides a point estimate of 0.2694818 and based upon this estimate we can conclude that the linear relation is not very strong as correlation varies between [-1,1]. The 95% confidence interval generated by percentile bootstrap is (0.0642, 0.4778), bootstrap estimate of bias is 0.003487544 and the bootstrap estimate standard error is 0.105556. The calculations are based on 10,000 bootstrap replicates, if they are large enough to generate an accurate bootstrap distribution. Our estimator can be considered quite accurate as the standard errors and the bias produced by the bootstrap samples are very small. As the range of confidence interval is quite wide, it can be inferred that there exists a positive correlation between the GPA and ACT scores but the extent of its strength is opaque as the interval is wide. However, the upper value of confidence interval is 0.4778 which is very small. The linear relationship which appears to be positive varies from very weak to moderately strong. There exists a linear relationship between GPA and ACT score but the strength of its relationship is not strong enough to compute one value from the other.

Question2:

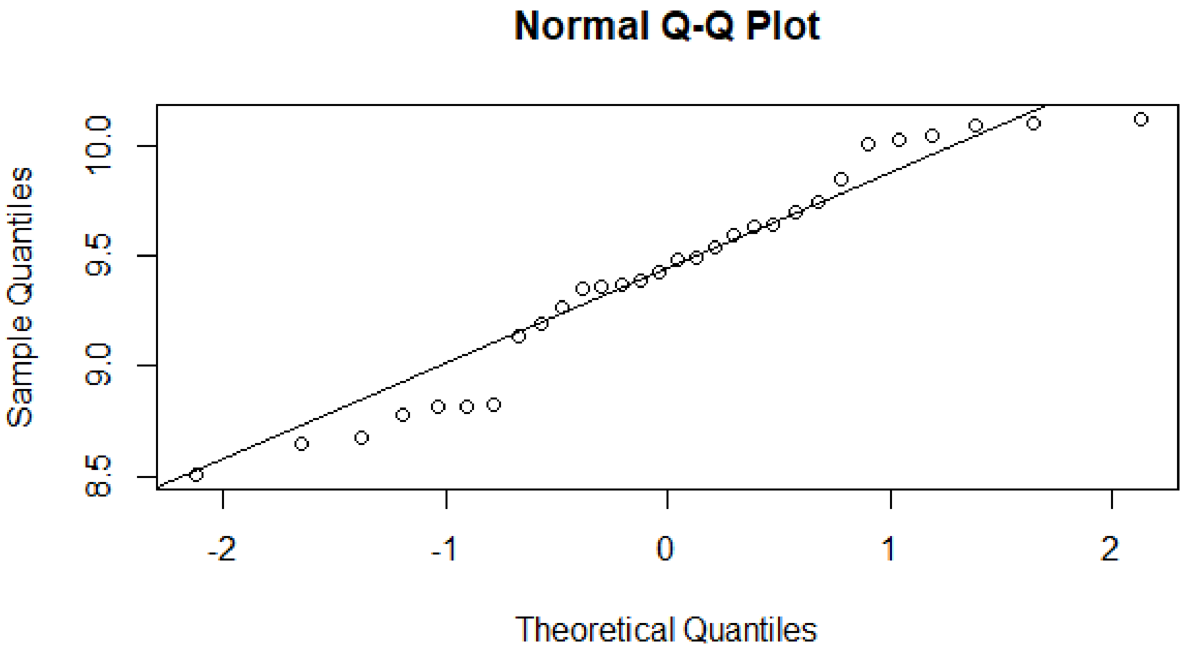
(7 points) Consider the data stored in the file VOLTAGE.csv on eLearning. These data come from a Harris Corporation/University of Florida study to determine whether a manufacturing process performed at a remote location can be established locally. Test devices (pilots) were set up at both the remote and the local locations and voltage readings on 30 separate production runs at each location was 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.*
Answer:

A. Frequency Histograms: After creating Frequency Histograms of the voltages, we observe that the remote location voltages are more discontinuous. However, this is not the case for local location voltages. The two distributions do not seem similar is the conclusion drawn after observing the frequency histograms.



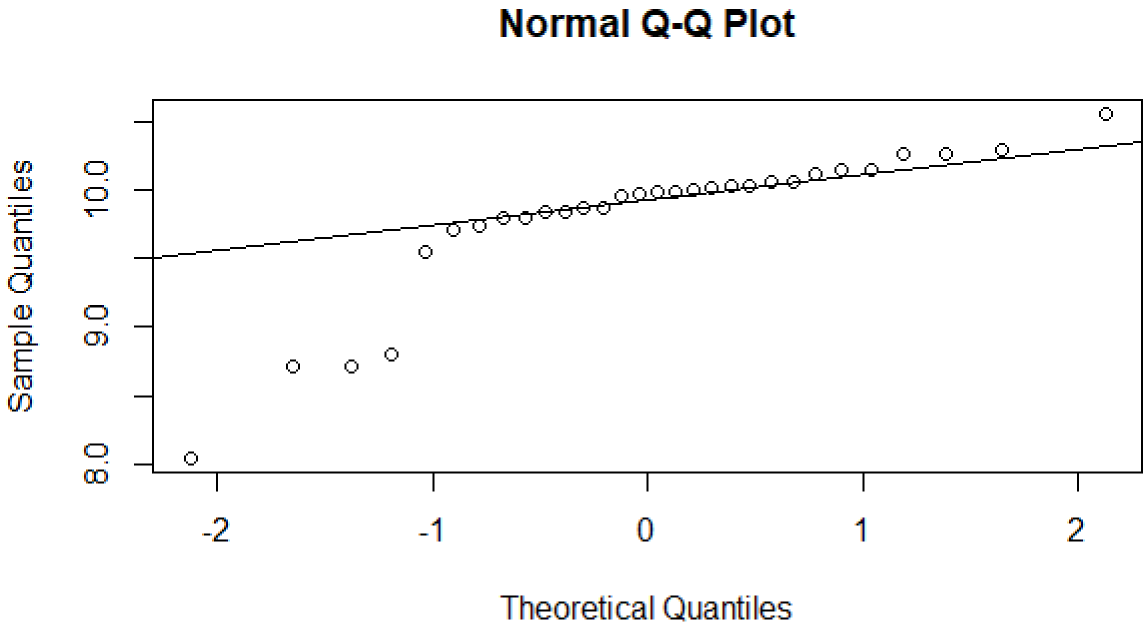


B. Q-Q Plot to find if local location voltages are Normal:



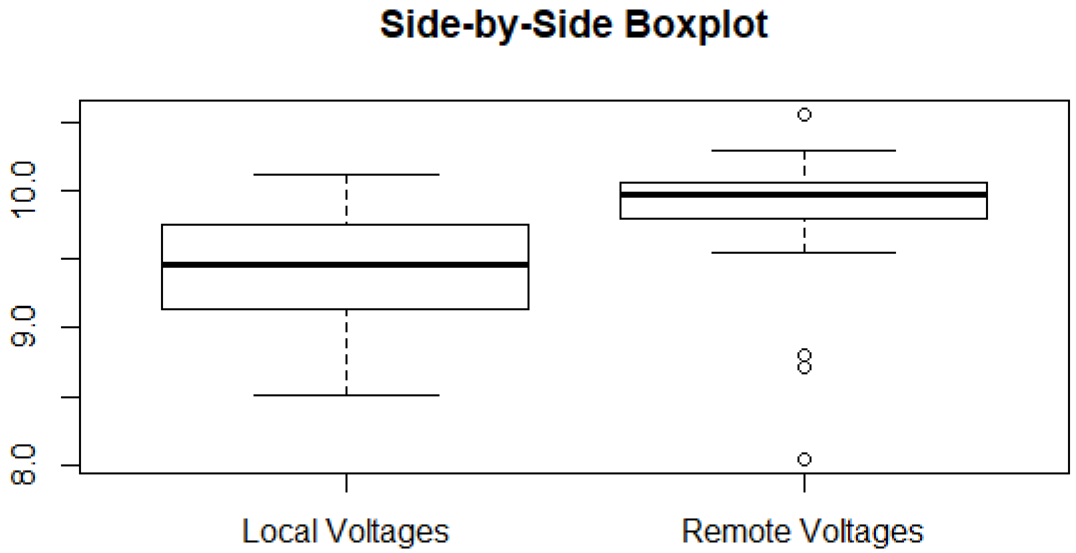
After plotting the QQ plot we observe that the sample distribution is not Normal, but we can say that the sample size is large enough to approximate a normal distribution.

Q-Q Plot to find if remote location voltages are Normal:



After plotting the QQ plot we observe that the sample distribution is not Normal, but we can say that the sample size is large enough to approximate a normal distribution.

C. Side-By-Side Boxplots:



We get a better comparison of both distributions after performing the Side by Side boxplots. We observe that, outliers are present only in the remote location voltages. We also observe that local location voltages have a wider IQR and larger variance in comparison to the remote location voltages and thus have no outliers. The IQR of remote location voltages is much smaller and has less variance. The means of the two distributions are also unequal and are different from each other. This also proved when we compute the summary statistics of the two distributions.

(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.*

Answer:

After performing the exploratory analysis in part (a) it is quite clear that the data does not follow a normal distribution. However, as n=30 for both distributions, we can assume that the sample sizes are large enough to have normal approximation.

After observing the IQR of both the distribution from the boxplot, we know that the variances of the both the distribution are unequal. There doesn't exists any pairing between each of these two data distributions and hence they form two independent samples.

On assuming large sample size and unequal variances for the given data the 95% confidence interval is (0.1172284 0.6454382). On an average the mean of the voltages measured at the remote location is greater than the voltage measured at the local location is concluded from the confidence interval. Hence, it can be concluded that the manufacturing process cannot be established locally. This conclusion is in accordance with the explanatory analysis in Part a.

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

Answer:

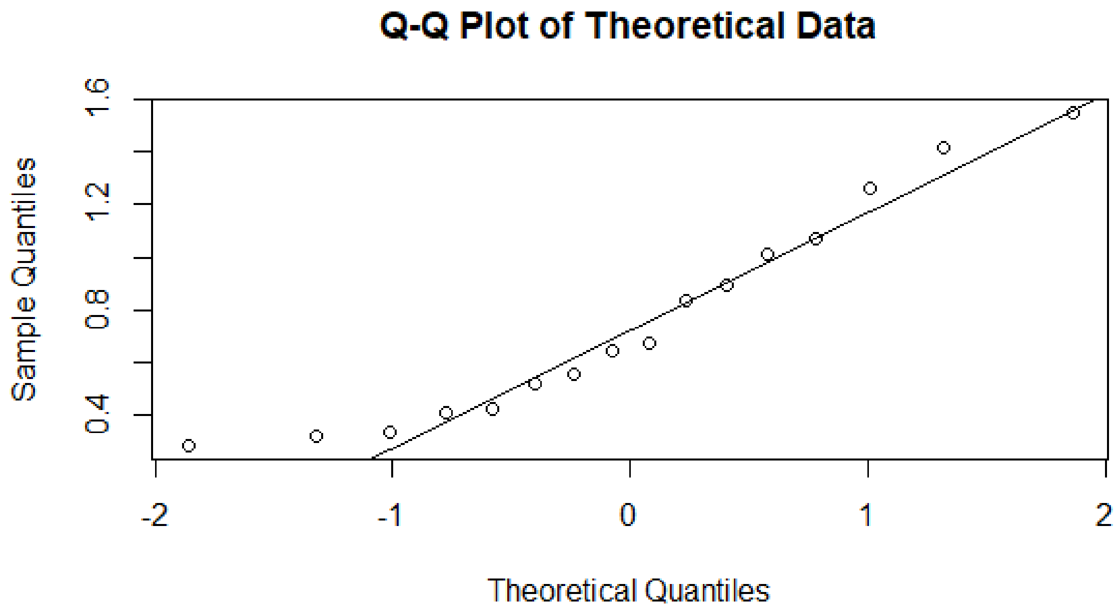
The sample means of the voltages at the remote and local locations have quite significant differences is seen from the summary statistics of the two distributions. The mean differed by 0.382, is quite large for the dataset. The centre of the confidence interval can never be 0 based on the given samples and becomes very difficult to conclude that the means of populations are equal. Hence, the conclusion that is obtained from Part b is same as that obtained from Part a.

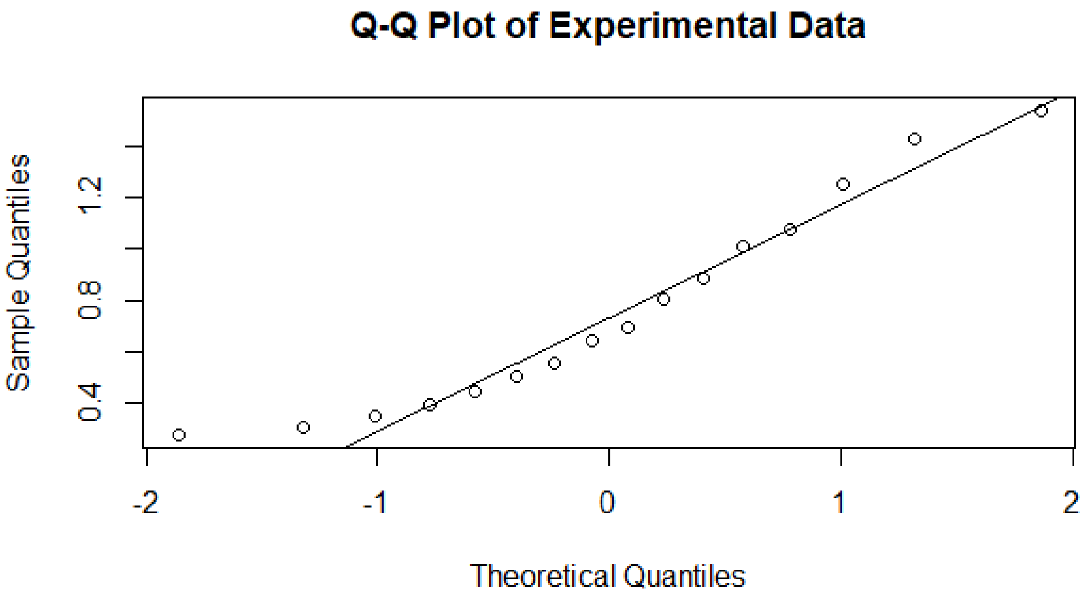
Question3:

(7 points) The file VAPOR.csv 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.

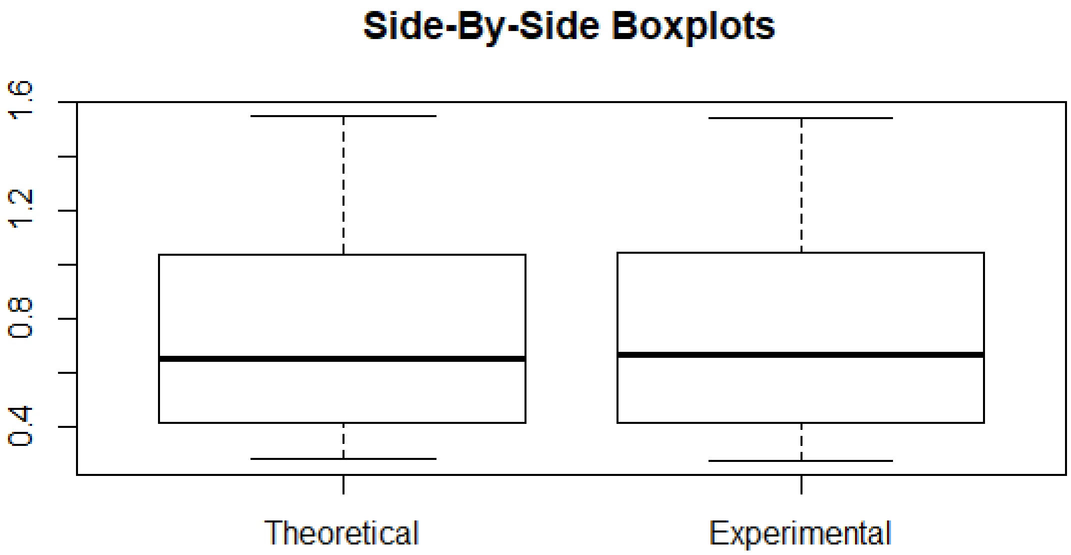
Answer:

The sample given for experimental vapor pressure and the theoretical vapor pressure is a paired sample from two distinct populations. For each of the given temperature values the experimentally obtained vapour pressure sample is paired with its theoretical counterpart.
Observing the QQ plot for Theoretical and Experimental data,





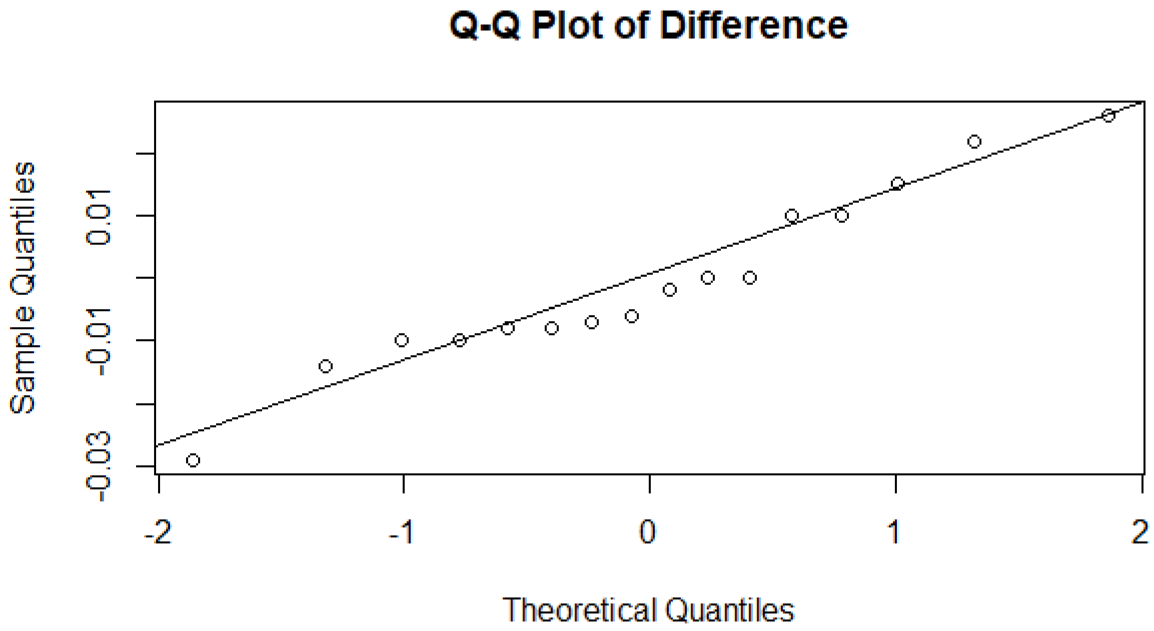
The distributions of the two sample populations are quite similar. In order to understand the difference between the means of the two populations we can plot a Side by Side boxplot for both distributions.



The difference between their means also seems small that is observed from the above Side by Side boxplot.

As the two-population samples are paired, the distribution of their difference can be obtained by subtracting the paired values at each of the given temperatures. Hence, we obtain a single sample containing the distribution of their differences and this distribution can be used to calculate the confidence interval.

Observing the QQ Plot of this difference distribution:



Since the number of samples are few and the above QQ plot does not observe Normal Distribution, we use one-sample bootstrap to obtain the 95% Confidence Interval.

We obtain the following 95% confidence interval: (-0.0072, 0.0064), after performing the percentile bootstrap on the one-sample difference distribution. We observe that the confidence interval obtained by the percentile bootstrap is very small. The bootstrap statistics also show that the 0 value is close to the centre of the Confidence Interval. Hence, we can say that there is no difference between their population means. Therefore, to conclude the theoretical model is good model of reality because the true mean difference between the experimental and calculated values of vapor pressure is approximately zero.

R CODE Question 1:

```
> ques1=read.csv("gpa.csv", header=TRUE, sep = ",") #Reading csv file
> gpaofstudents=as.numeric(as.character((ques1[,1])))
> actofstudents=as.numeric(as.character((ques1[,2])))

#Plotting Scatter plot of GPA vs ACT of students, and trying to fit a line.
> plot(x=actofstudents,y=gpaofstudents,main="Scatter plot of GPA vs ACT")
> abline(lm(gpaofstudents~actofstudents), col="yellow")

> cor(actofstudents,gpaofstudents) #Generating point estimate from the sample
[1] 0.2694818

> library(boot) #Importing the boot library

# Function to calculate correlation from the sampled distribution
> nonparacor=function(a,index){answer=cor(a[index,1],a[index,2]); return(answer)}

#Generating bootstrap distribution
> finalanswer=boot(ques1, nonparacor, 10000, sim = "ordinary", stype = "i")
> finalanswer

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = ques1, statistic = nonparacor, R = 10000, sim = "ordinary",
    stype = "i")


Bootstrap Statistics :
     original      bias    std. error
t1* 0.2694818 0.003487544    0.105556


> mean(finalanswer$t) #Comparing Actual mean with the mean from bootstrap statistics
[1] 0.2729693

#Percentile confidence interval of bootstrap distribution
boot.ci(boot.out = finalanswer, conf = 0.95, type = "perc")

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

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

Intervals :
Level     Percentile
95%   ( 0.0642,  0.4778 )
Calculations and Intervals on Original Scale
```

R CODE QUESTION 2:

Part a

```
> fulldata=read.csv('voltage.csv', header = TRUE, sep = ",")#Reading csv file
> reomtedata=(fulldata$location==0)
> remotedata=which(fulldata$location==0) #Check locations of remote site
> remotedata
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
[27] 27 28 29 30
> remote=fulldata[1:30, ]#Observed data shows that 1st 30 are remote locations
> local=fulldata[31:60, ] #Next 30 entries will be local locations
> remotevol=remote[,2] #Storing values of voltages of remote locations
> localvol=local[,2] #Storing values of voltages of local locations

> hist(remotevol, main="Histogram of Remote Location Voltages", breaks = 20, xlim = c(
8,11), ylim = c(0,6))#Plotting histogram of Remote location voltages

> hist(localvol, main="Histogram of Local Location Voltages", breaks = 20, xlim = c(8,
11), ylim = c(0,6)) # Plotting histogram of Local location voltages

> qqnorm(localvol) #Plotting QQ Plot for Local voltages
> qqline(localvol)

> qqnorm(remotevol) #Plotting QQ Plot for Remote Voltages
> qqline(remotevol)

#Summary Statistics to compute difference in Means
> summary(localvol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  8.510   9.152   9.455   9.422   9.738  10.120 
> summary(remotevol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  8.050   9.800   9.975   9.804  10.050  10.550 
> sd(localvol)
[1] 0.4788757
> sd(remotevol)
[1] 0.5409155


> boxplot(localvol,remotevol, main="Side-by-Side Boxplot", names=c("Local Voltages","R
emote Voltages"))
#Plotting side-by-side boxplots of the two voltages
```

Part b

```
#Using Satterthwaite's approximation because nx=ny=30 sample and unequal
variances
>nx=ny=30
v=((sd(remotevol)^2/nx)+((sd(localvol)^2)/ny))^2/((sd(remotevol)^4)/((nx^2)*(nx-1))+(s
d(localvol)^4/((ny^2)*(ny-1))))
> v
[1] 57.16003

#Computing 95% confidence interval
ci=mean(remotevol)-mean(localvol)+c(-1,1)*qt(1-(1-0.95)/2,v)*sqrt((sd(remotevol)^2/nx)
+sd(localvol)^2/ny)
> ci
[1] 0.1172284 0.6454382
```

R CODE QUESTION 3:

```
> fulldata=read.csv("Vapor.csv", header=TRUE, sep=",") #Reading csv file
> tempdata=fulldata[,1]
> theorydata=fulldata[,2] #Storing theoretical observations
> expdata=fulldata[,3] #Storing experimental observations

#Plotting QQ Plot for Theoretical observations
> qqnorm(theorydata, main="Q-Q Plot of Theoretical Data")
> qqline(theorydata)


# Plotting QQ Plot for Experimental Observations
> qqnorm(expdata, main="Q-Q Plot of Experimental Data")
> qqline(expdata)


> boxplot(theorydata,expdata, main="Side-By-Side Boxplots", names=c("Theoretical", "Ex
perimental")) # Plotting Side by side Boxplots for comparison


> diff=expdata-theorydata #Computing Difference distribution
> diff
 [1] -0.006 -0.007  0.015 -0.014  0.022 -0.008  0.000 -0.002  0.026 -0.029 -0.008
[12]  0.000  0.010 -0.010  0.010 -0.010
> mean(diff)
[1] -0.0006875

#Plotting QQ Plot for Difference Distribution
> qqnorm(diff, main = "Q-Q Plot of Difference")
> qqline(diff)



> differencemean=function(a,index){answer=mean(a[index); return(answer)} # Function to
calculate mean from the resampled diff distribution

# Computing bootstrap distribution
> bootdistribution=boot(diff,differencemean,1000, sim = "ordinary", stype = "i")
> bootdistribution

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = diff, statistic = differencemean, R = 1000, sim = "ordinary",
    stype = "i")


Bootstrap Statistics :
      original      bias    std. error
t1* -0.0006875 8.59375e-05 0.003437996

#Computing Confidence Interval
> boot.ci(boot.out = bootdistribution, conf = 0.95, type = "perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

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

Intervals :
Level     Percentile
95%   (-0.0072,  0.0064 )
Calculations and Intervals on Original Scale
```