STATISTICAL METHODS FOR DATA SCIENCE Mini-Project 4

Duo Group #23

Members: Hima Sri Tipirineni

Nithin Pingili

Contribution of each team member:

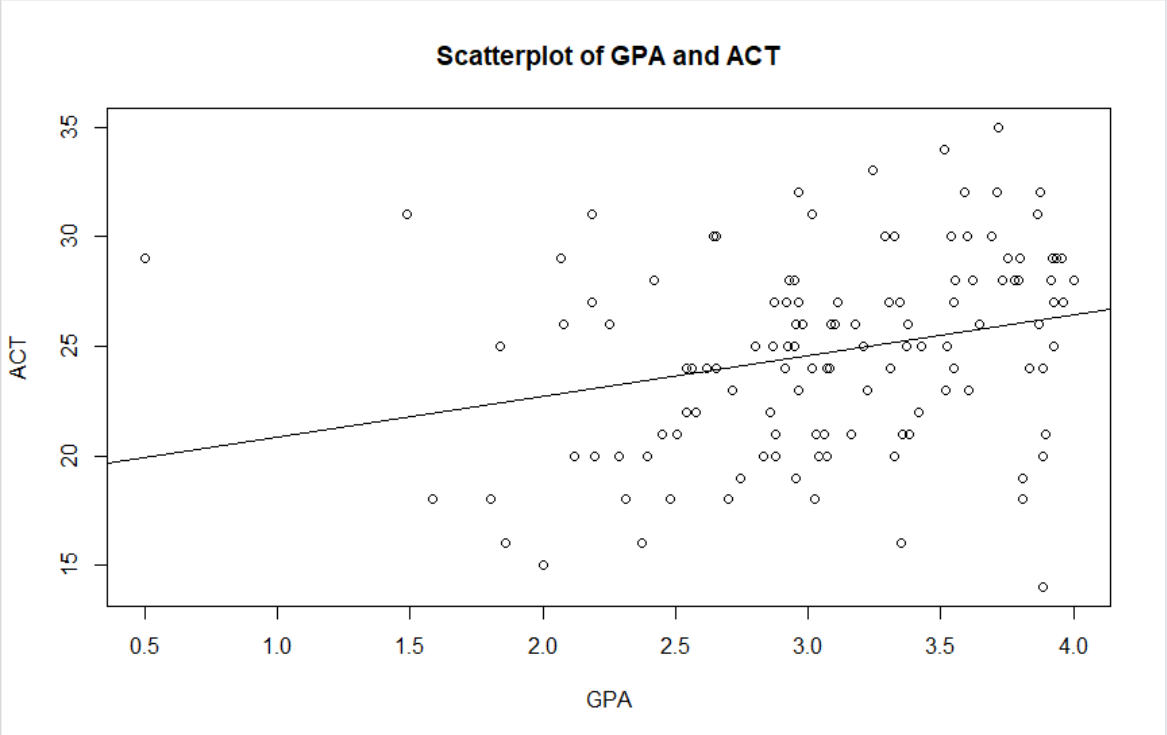
Hima Sri and Nithin worked together to complete both the questions. Collaborated to learn R and then worked on plotting the scatter plots, qq plots, boxplots and histograms for the given questions. Also worked on finding the bootstrap estimates and confidence intervals. Both worked together to answer the questions and report all the findings. Hima Sri wrote R code and annotated the code and Nithin worked to check the accuracy of the R code and added the observations. Both worked efficiently to complete all sections of the project.

Question 1:

(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.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.)

Solution:

```
>
> ### Question 1 ####
>
> ### Reading the GPA data into R ###
> gpa_data = read.csv("C:\\Users\\hxt210018\\Downloads\\6313_Prob\\HW\\gpa.csv")
>
>
> ### Drawing scatter plots ###
> gpa = as.numeric(gpa_data$gpa)
> act = as.numeric(gpa_data$act)
>
> plot(gpa,act,main="Scatterplot of GPA and ACT",xlab = "GPA",ylab = "ACT")
> abline(lm(act~gpa))
>
> ## Correlation calculation
> cor(gpa,act)
[1] 0.2694818
>
>
>
```



The scatterplot shows weak positive correlation of the linear relationship between the two variables. The cor(gpa, act) value of 0.2694818 is near 0 which shows weak positive correlation.

```
>
> ### Question 1 ####
>
> ### Reading the GPA data into R ###
> gpa_data = read.csv("C:\\Users\\hxt210018\\Downloads\\6313_Prob\\HW\\gpa.csv")
>
>
> ### Drawing scatter plots ###
> gpa = as.numeric(gpa_data$gpa)
> act = as.numeric(gpa_data$act)
>
> plot(gpa,act,main="Scatterplot of GPA and ACT",xlab = "GPA",ylab = "ACT")
> abline(lm(act~gpa))
>
> ## Correlation calculation
> cor(gpa,act)
[1] 0.2694818
>
> ##Import/attach the boot library ##
> library(boot)
>
> ## Statistic function for correlation ##
>
> covariance_npar = function(gpa_act_data,index) {
+   agpa = gpa_act_data$gpa[index]
+   gact = gpa_act_data$act[index]
+   result = cor(agpa,gact)
+   return (result)
+ }
>
> ## Execute boot function with statistical function ##
> covariance_npar.boot = boot(gpa_data,covariance_npar, R = 999, sim = "ordinary",stype = "i")
>
>
> ## Point estimate of bootstrap value ##
> mean(covariance_npar.boot$t)
[1] 0.275013
>
>
> ##Calculating confidence interval using boot.ci ##
> boot.ci(boot.out = covariance_npar.boot)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

CALL :
boot.ci(boot.out = covariance_npar.boot)

Intervals :
Level      Normal              Basic
95%   ( 0.0497,  0.4782 )   ( 0.0470,  0.4760 )

Level     Percentile            BCa
95%   ( 0.0630,  0.4919 )   ( 0.0410,  0.4716 )
Calculations and Intervals on Original Scale
```

```
> ## bootstrap estimates of bias and standard error of the point estimate ##
> covariance_npar.boot

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = gpa_data, statistic = covariance_npar, R = 999, sim = "ordinary",
    stype = "i")


Bootstrap Statistics :
     original      bias    std. error
t1* 0.2694818 0.003851106    0.108358
```

```
>
>
> ## Verifying the confidence intervals using the percentile bootstrap ##
> sort(covariance_npar.boot$t)[c(25, 975)]
[1] 0.06298279 0.49192774
>
```

The point estimate ρ = 0.2694818, bias = 0.003851106, standard error = 0.108358

The 95% confidence interval computed using percentile bootstrap is [0.0630, 0.4919]. The confidence interval shows that the point estimate ρ is likely between 0.063 and 0.04919

It can be interpreted that the point estimate of bootstrap value is close to the actual correlation of the sample. Also, the 95% confidence interval computed using percentile bootstrap captures the population correlation in the 95% intervals calculated from population.

R code :

```
### QUestion 1 ####

### Reading the GPA data into R ###

gpa_data = read.csv("C:\\Users\\hxt210018\\Downloads\\6313_Prob\\HW\\gpa.csv")

### Drawing scatter plots ###

gpa = as.numeric(gpa_data$gpa)

act = as.numeric(gpa_data$act)

plot(gpa,act,main="Scatterplot of GPA and ACT",xlab = "GPA",ylab = "ACT")

abline(lm(act~gpa))

## Correlation calculation

cor(gpa,act)

##Import/attach the boot library ##

library(boot)

## Statistic function for correlation ##

covariance_npar = function(gpa_act_data,index) {

 agpa = gpa_act_data$gpa[index]

  gact = gpa_act_data$act[index]

  result = cor(agpa,gact)

  return (result)

}

## Execute boot function with statistical function ##

covariance_npar.boot = boot(gpa_data,covariance_npar, R = 999, sim = "ordinary",stype = "i")

covariance_npar.boot

## Point estimate of bootstrap value ##
```

mean(covariance_npar.boot$t)

##Calculating confidence interval using boot.ci ##

boot.ci(boot.out = covariance_npar.boot)

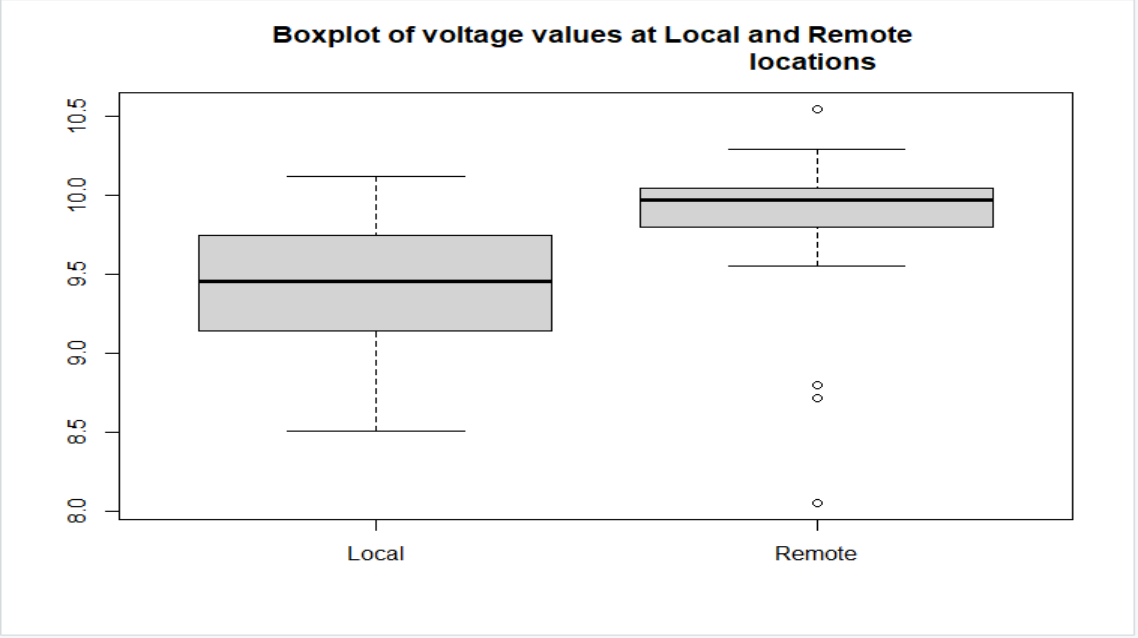
Verifying the confidence intervals using the percentile bootstrap

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

Question 2: (7 points) 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.

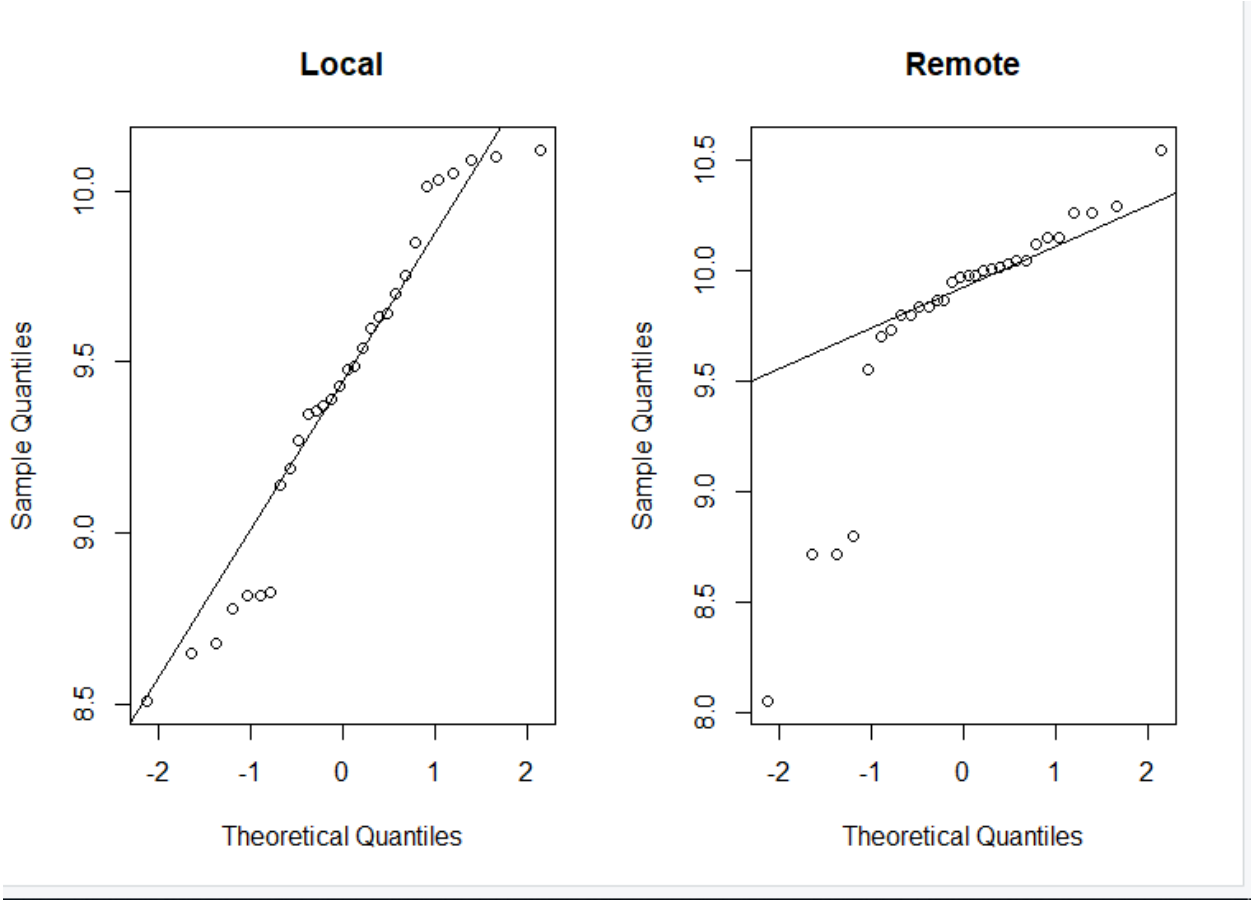
```
>
>
> ### Question 2 ###
> ## Reading the voltage data into R ##
> voltage_data = read.csv("C:\\Users\\hxt210018\\Downloads\\6313_Prob\\HW\\VOLTAGE.csv")
> voltage_data.remote = voltage_data$voltage[which(voltage_data$location == 0)]
> voltage_data.local = voltage_data$voltage[which(voltage_data$location == 1)]
>
> ## Box plots to conduct the exploratory analysis ##
> boxplot(voltage_data.local,voltage_data.remote, names = c("Local","Remote"),main = "Boxplot of voltage values at Local and Remote
+                                                    locations",range = 1.5)
>
> summary(voltage_data.remote)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.050   9.800   9.975   9.804  10.050  10.550
> summary(voltage_data.local)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.510   9.152   9.455   9.422   9.738  10.120
>
```



The boxplots show that the two distributions are not similar. The voltage values at remote locations are higher than the voltage values at local locations because the first quartile, median, third quartile values at remote locations are greater than that at local locations. The distribution of voltage values at local locations is symmetric because the median is approximately equal to the mean. The distribution of voltage values at remote locations is left skewed because the median is greater than the mean and the median is closer to the third quartile. The boxplot of voltage values at local location has no outliers and the boxplot of voltage values at remote location has four outliers.

(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.

```
>
>
> ## Drawing QQ plots for the datasets ##
>
> par(mfrow=c(1,2))
> qqnorm(voltage_data.local,main= "Local")
> qqline(voltage_data.local)
> qqnorm(voltage_data.remote, main = "Remote")
> qqline(voltage_data.remote)
>
```



On the QQ plot for local most of the points fall on the straight line which shows that the distribution for local is assumed to be normal. However, the distribution for remote locations is not assumed to be normal because most of the points do not fall on the straight line.

```
>
> ## Calculating mean, variance, standard error and confidence intervals ##
> var_l = var(voltage_data.local)
> var_l
[1] 0.229322
>
> var_r = var(voltage_data.remote)
> var_r
[1] 0.2925895
>
>
> standard_error = sqrt(var_l/30 + var_r/30)
> standard_error
[1] 0.1318979
>
>
> difference_of_means = mean(voltage_data.remote) - mean(voltage_data.local)
> CI = difference_of_means + c(-1,1) * qnorm(0.975) * standard_error
> CI
[1] 0.1228182 0.6398484
>
>
> ## Calculating the confidence interval using the t test ##
> t.test(voltage_data.remote,voltage_data.local,alternative = "two.sided",paired = FALSE, var.equal = FALSE, conf.level = 0.95)

        Welch Two Sample t-test

data:  voltage_data.remote and voltage_data.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

> |
```

Null hypothesis: difference in the population means of voltage readings at the two locations is equal to 0

Alternative hypothesis: difference in the population means of voltage readings at the two locations is not equal to 0

Assume the data is normal based on the QQ plots. We don't assume the population variances are equal because the interquartile ranges are very different. Therefore, we can find the confidence intervals using the Satterthwaite approximation and t-distribution.

The 95% confidence interval is [0.1228182, 0.6398484]. The confidence interval shows that the mean voltage for remote locations is greater than that of local locations by value between 0.1228182 and 0.6398484

To verify the 95% confidence interval, a t test must be performed. The t test gives the values [0.1172284, 0.6454382]. So, we can conclude that the confidence interval is appropriate and assumptions are valid. Because 0 does not lie in the confidence interval from the t test, the null hypothesis is rejected. Therefore, the manufacturing process cannot be established locally.

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

In part a, we observed that the first quartile, median and third quartile voltage values at remote locations are higher than that of local locations. This suggests that the mean of distribution of voltage values at remote location is greater than that of distribution of voltage values at local location. The result in part b also confirms this part. In part b we also concluded that the manufacturing process

cannot be established locally. The manufacturing process requires higher voltages of power and remote location has higher voltages than local location.

R code :

```
### Question 2 ###

## Reading the voltage data into R ##

voltage_data = read.csv("C:\\Users\\hxt210018\\Downloads\\6313_Prob\\HW\\VOLTAGE.csv")

voltage_data.remote = voltage_data$voltage[which(voltage_data$location == 0)]

voltage_data.local = voltage_data$voltage[which(voltage_data$location == 1)]

## Box plots to conduct the exploratory analysis ##

boxplot(voltage_data.local,voltage_data.remote, names = c("Local","Remote"),main = "Boxplot of voltage values at Local and Remote locations",range = 1.5)

summary(voltage_data.remote)

summary(voltage_data.local)

## Drawing QQ plots for the datasets ##

par(mfrow=c(1,2))

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

qqline(voltage_data.local)

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

qqline(voltage_data.remote)

## Calculating mean, variance, standard error and confidence intervals ##

var_l = var(voltage_data.local)

var_l

var_r = var(voltage_data.remote)

var_r

standard_error = sqrt(var_l/30 + var_r/30)

standard_error

difference_of_means = mean(voltage_data.remote) - mean(voltage_data.local)

CI = difference_of_means + c(-1,1) * qnorm(0.975) * standard_error

CI
```

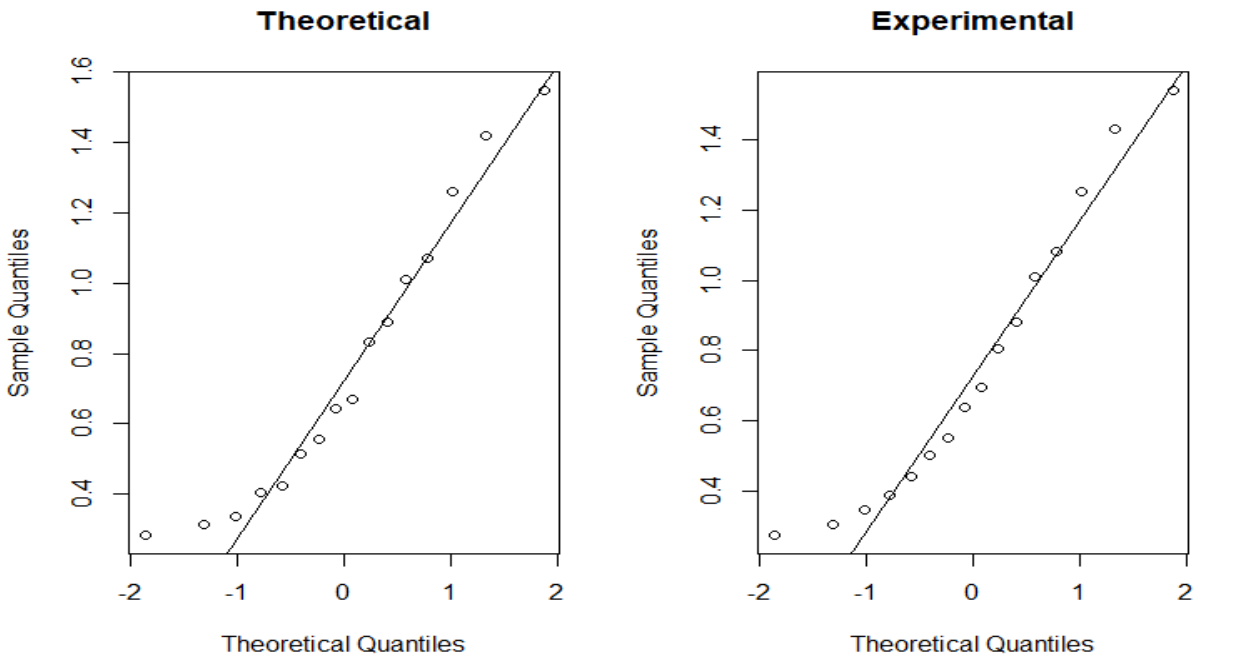
Calculating the confidence interval using the t test

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

Question 3 : (7 points) 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

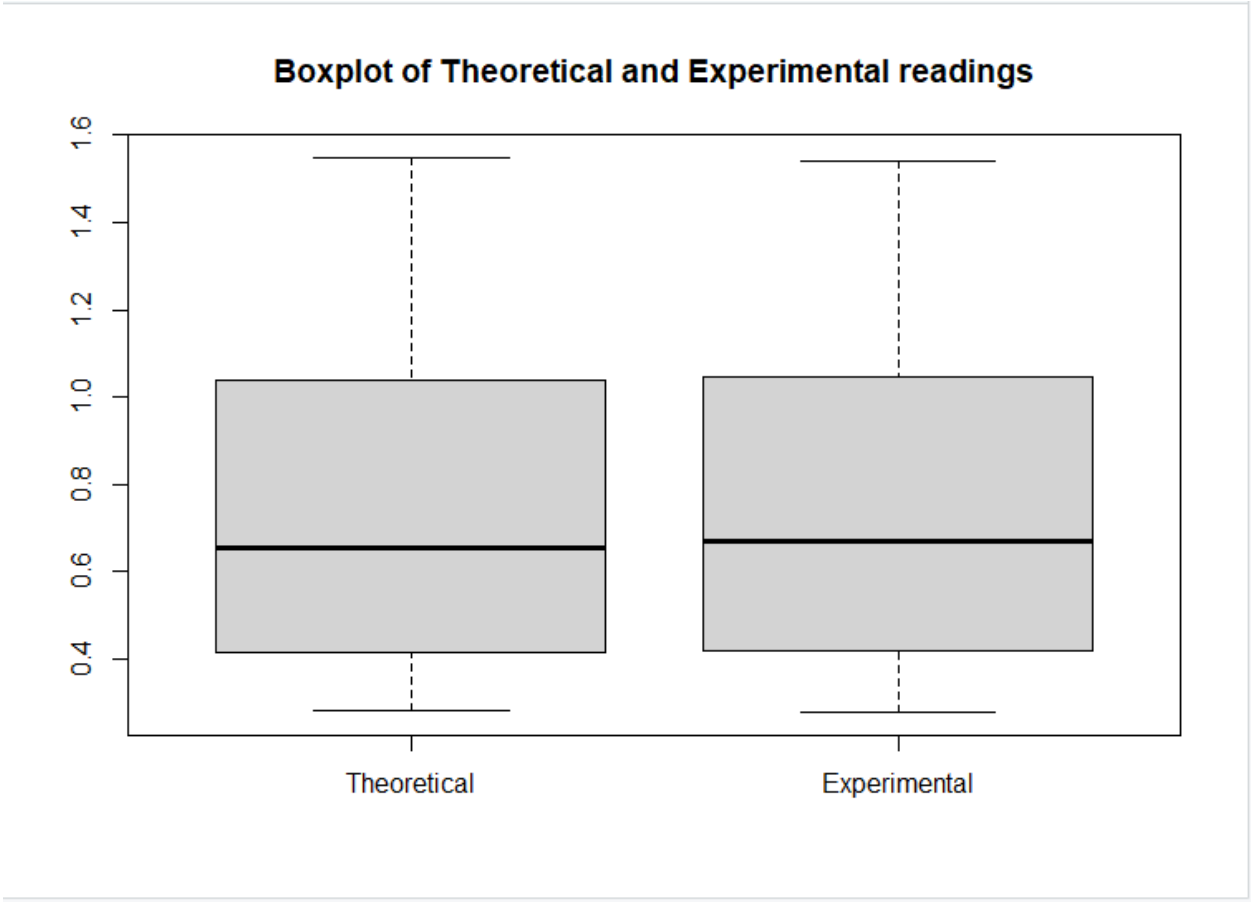
Solution:

```
>
> ### Question 3 ###
>
> ## Read the vapor data for the file into R ##
> vapor = read.csv("C:\\Users\\hxt210018\\Downloads\\6313_Prob\\HW\\VAPOR.csv")
>
> ## Drawing the QQ plots ##
> par(mfrow = c(1,2))
> qqnorm(vapor$theoretical, main = "Theoretical")
> qqline(vapor$theoretical)
> qqnorm(vapor$experimental, main = "Experimental")
> qqline(vapor$experimental)
>
```

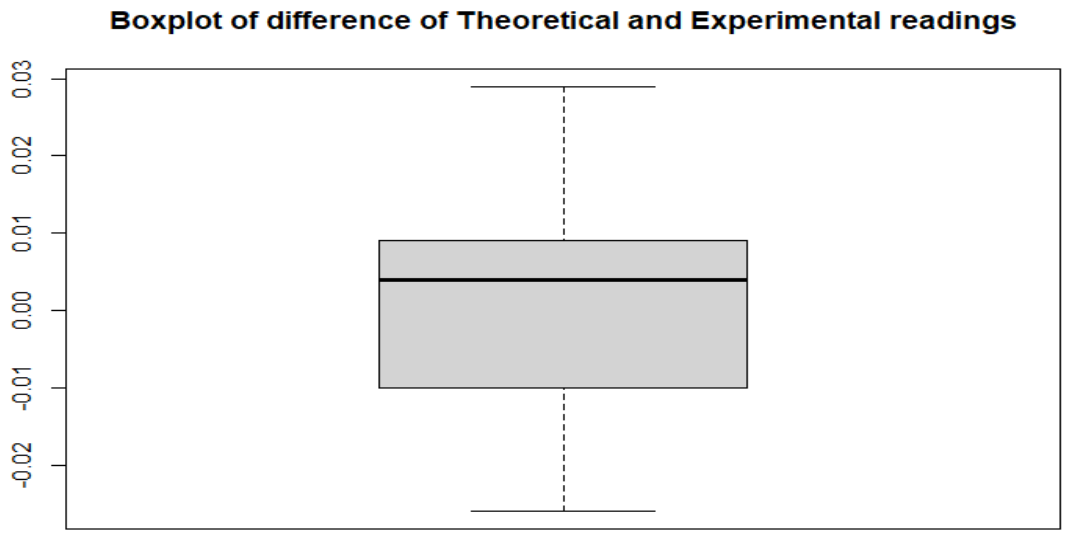


On the QQ plots most of the points fall on the straight line which shows that the distributions for theoretical and experimental are assumed to be normal.

```
>
> ## Drawing the boxplots and summaries ##
> par(mfrow = c(1,1))
> boxplot(vapor$theoretical,vapor$experimental,names = c("Theoretical","Experimental"),
+         main = "Boxplot of Theoretical and Experimental readings")
>
```



```
> diff = vapor$theoretical - vapor$experimental
>
> boxplot(diff, main = "Boxplot of difference of Theoretical and Experimental readings")
>
> 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
>
> summary(diff)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.0260000 -0.0100000  0.0040000  0.0006875  0.0085000  0.0290000
>
```



It can be concluded from the boxplots that the distributions of theoretical and exponential readings are similar because the first quartile, median, third quartile values for theoretical and experimental distributions are very similar. The difference between the 5 number summaries of both distributions is negligible. Also, both distributions are right skewed because the mean is greater than the median, the median is closer to first quartile than third quartile and the difference between first quartile and minimum is less than the difference between third quartile and maximum.

```
# calculating mean, standard deviation and confidence interval

>
> mean_d = mean(diff)
> mean_d
[1] 0.0006875
>
> sd_d = sd(diff)
> sd_d
[1] 0.01421604
>
> qt(0.975,15)
[1] 2.13145
>
> cI = mean_d + c(-1,1) * qt(0.975,15) * sd_d/sqrt(16)
> cI
[1] -0.006887694  0.008262694
>
> ## Using t test calculating confidence intervals ##
>
> t.test(vapor$theoretical, vapor$experimental, alternative = "two.sided", paired = TRUE, var.equal = 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

> |
```

Null hypothesis: true mean difference between the experimental and theoretical values of vapor pressure is equal to zero

Alternative hypothesis: true mean difference between the experimental and theoretical values of vapor pressure is not equal to zero

The calculated mean and standard deviation are:

mean = 0.0006875, standard deviation = 0.01421604

The 95% confidence interval is [-0.006887694, 0.008262694]. The confidence interval shows that the true mean difference between the experimental and theoretical values of vapor pressure is between -0. 006887694 and 0.008262694

To verify the confidence interval, a t test must be performed. The t test gives the values [-0.006887694, 0.008262694]. So, we can conclude that the confidence interval is appropriate. Because 0 lies in the confidence interval from the t test, the null hypothesis is accepted. Therefore, the true mean difference between the experimental and theoretical values of vapor pressure is zero.

R code :

```
### Question 3 ###

## Read the vapor data for the file into R ##

vapor = read.csv("C:\\Users\\hxt210018\\Downloads\\6313_Prob\\HW\\VAPOR.csv")
```

```
## Drawing the QQ plots ##
par(mfrow = c(1,2))
qqnorm(vapor$theoretical, main = "Theoretical")
qqline(vapor$theoretical)
qqnorm(vapor$experimental, main = "Experimental")
qqline(vapor$experimental)
## Drawing the boxplots and summaries ##
par(mfrow = c(1,1))
boxplot(vapor$theoretical,vapor$experimental,names = c("Theoretical","Experimental"),
      main = "Boxplot of Theoretical and Experimental readings")
diff = vapor$theoretical - vapor$experimental
boxplot(diff, main = "Boxplot of difference of Theoretical and Experimental readings")
summary(vapor$theoretical)
summary(vapor$experimental)
summary(diff)
## Calculating mean, standard error and confidence intervals ##
mean_d = mean(diff)
mean_d
sd_d = sd(diff)
sd_d
qt(0.975,15)
cI = mean_d + c(-1,1) * qt(0.975,15) * sd_d/sqrt(16)
cI
## Using t test calculating confidence intervals ##
t.test(vapor$theoretical, vapor$experimental, alternative = "two.sided", paired = TRUE, var.equal =
FALSE, conf.level = 0.95)
```