CS 6313 Statistical Methods in Data Science
Mini Project 4

Gaurav Joshi

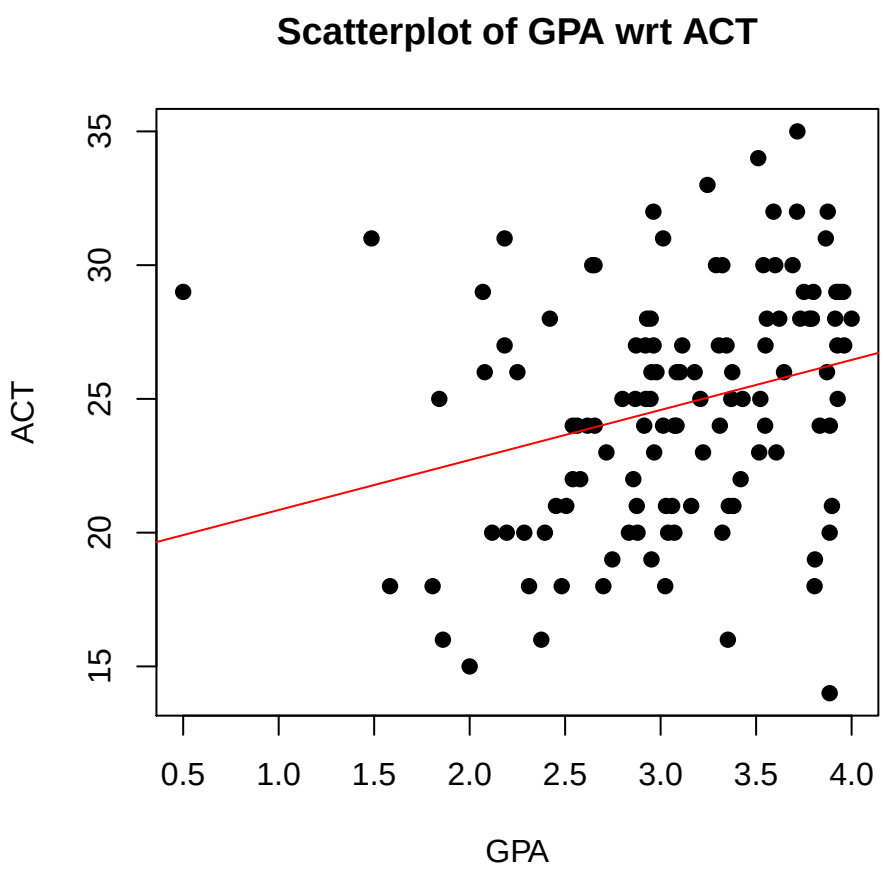
April 11, 2022

Question 1

The R Code for to draw scatter plot is given below.

R Code:

```
# Part 1: Draw scatter plot with the line

# Read gpa values from gpa.csv
gpa_values <- read.csv("gpa.csv")

# Get the gpa and act from the data
gpa <- gpa_values$gpa
act <- gpa_values$act

# draw a scatter plot of gpa wrt to act
plot(gpa, act, main = "Scatterplot of GPA wrt ACT",
     xlab = "GPA", ylab = "ACT", pch = 19)

# add a straight line to the scatter plot
abline(lm(act~gpa, data = gpa_values), col = "red")
```



From the scatter plot, it is clear that gpa and act have a weak linear relationship. This can be proved by the `abline` function which plots the straight line in the scatter plot with a positive slope.

The next step is to calculate bootstrap estimates. This can be done by computing correlation between gpa and act for the entire dataset. (shown on the next page).

```r
# Part 2: Get the bootstrap estimates for percentile

# import the boot library to calculate bootstrap statistics
library(boot)

# Get the correlation of gpa and act
cor(gpa, act)
```

```
## [1] 0.2694818
```

```r
statistic <- function(gpa_values, i) {
  return (cor(gpa_values$gpa[i], gpa_values$act[i]))
}

# The bootstrap re-sampling takes the following parameters
# data
# statistic (user defined function)
# Number of bootstrap replications
# simulation type which is non-parametric by default
# Character string of type indices
covarience <- boot(gpa_values, statistic = statistic,
     R = 999, sim = "ordinary", stype = "i")

# Output of covarience
covarience
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = gpa_values, statistic = statistic, R = 999, sim = "ordinary",
##     stype = "i")
##
##
## Bootstrap Statistics :
##      original      bias    std. error
## t1* 0.2694818 0.004101003   0.1081609
```

```r
# Calculate point estimate
mean(covarience$t)
```

```
## [1] 0.2735828
```

```r
# Calculate confidence interval
boot.ci(covarience, type = c("norm", "basic", "perc", "bca"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 999 bootstrap replicates
##
## CALL :
```

```
## boot.ci(boot.out = covarience, type = c("norm", "basic", "perc",
##     "bca"))
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.0534,  0.4774 )   ( 0.0573,  0.4750 )
##
## Level     Percentile            BCa
## 95%   ( 0.0640,  0.4817 )   ( 0.0325,  0.4571 )
## Calculations and Intervals on Original Scale
```

```
# Verify confidence intervals using quantile
sort(covarience$t)[c(25, 975)]
```

```
## [1] 0.06397271 0.48169241
```

The `cor` function in R calculates correlation between ACT and GPA. The confidence interval is calculated using boot from the boot library with the user defined `covariance` function as its main argument.

From the results we can say that the point estimate of bootstrap (given by the mean of covarience for variable t) is close to the correlation of the samples. The confidence interval calculated using the boot library is also close to quantile value of sorted bootstrap.

Question 2

(a)

The two distributions are compared using side-by-side boxplot. The R code and the box plot are given below.

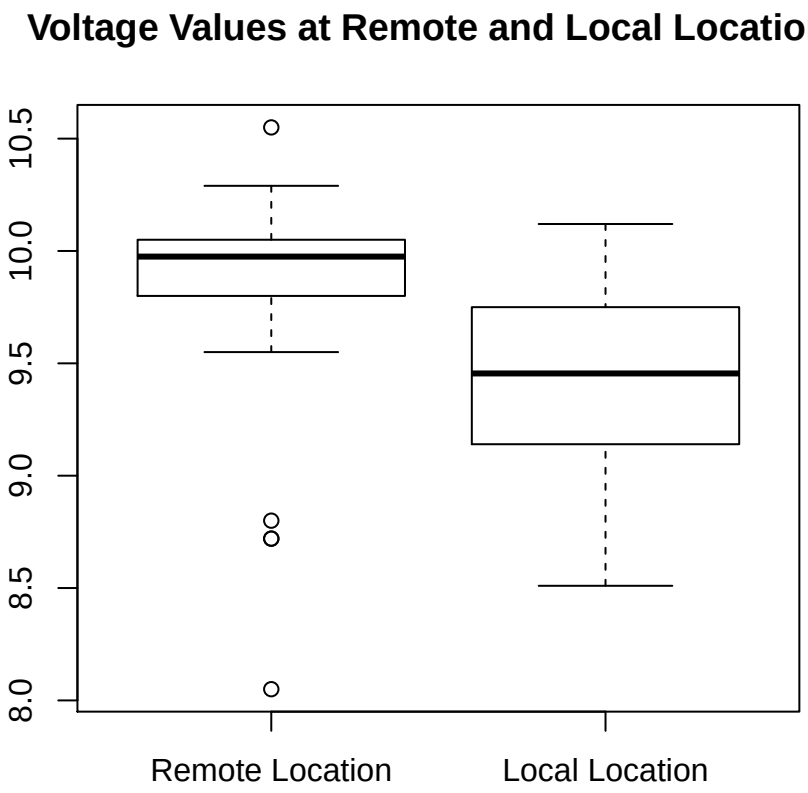
```
# Part 1: Draw a Box Plot

# read csv file
voltage_data <- read.csv("VOLTAGE.csv")

# Store the information about remote and local voltage
remote_voltage <- voltage_data$voltage[which(voltage_data$location == 0)]
local_voltage <- voltage_data$voltage[which(voltage_data$location == 1)]

# Summary of remote and local voltage
summary(remote_voltage)
```

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

```
summary(local_voltage)
```

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

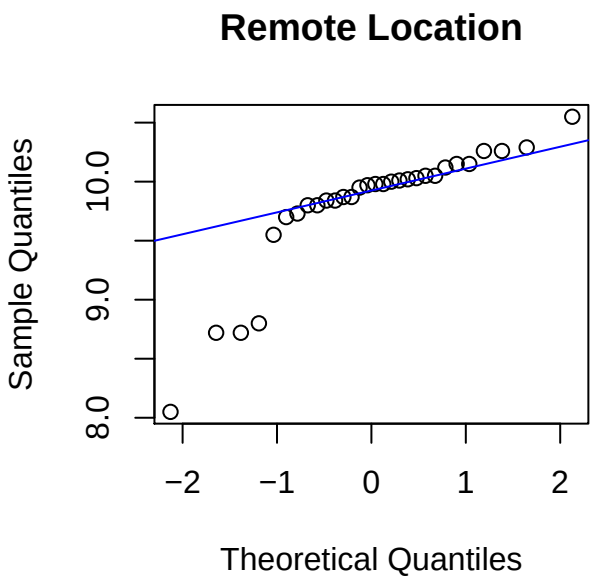
```
# Draw a Box Plot of Local and Remote Voltage Values
boxplot(remote_voltage, local_voltage,
        names = c("Remote Location", "Local Location"),
        main = "Voltage Values at Remote and Local Locations", range = 1.5)
```

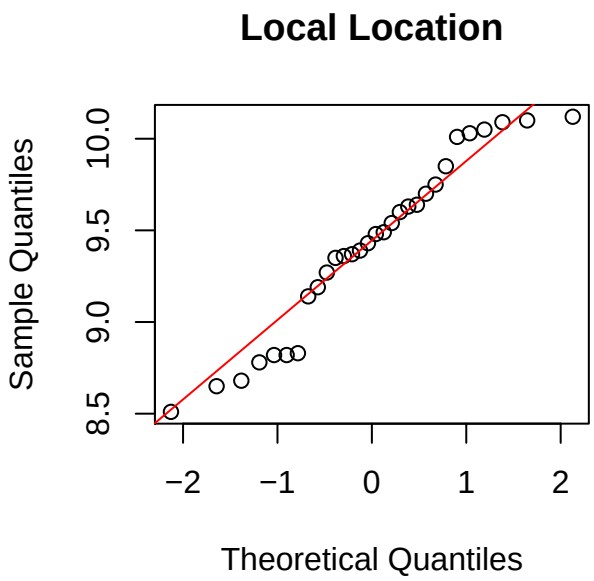


We can clearly see that the remote location data is left skewed, whereas the local location data is symmetric. The remote location data alos has more outliers than the local location data.

We now draw the QQ Plot.

```
qqnorm(remote_voltage, main = "Remote Location", pch=1)
qqline(remote_voltage, col="blue")
qqnorm(local_voltage, main = "Local Location", pch=1)
qqline(local_voltage, col="red")
```





The sample quantile appears a little higher in the remote location than the local location.

(b)

The null hypothesis can be formulated using the statement '`there is no difference in the population means of voltage readings at the two locations.`' from the question.

The null hypothesis would be:

$$\text{Sample Mean of Remote Locations} - \text{Sample Mean of Local Locations} = 0$$

The alternative hypothesis would be:

$$\text{Sample Mean of Remote Locations} - \text{Sample Mean of Local Locations} \neq 0$$

We cannot assume that the population variance is equal based on the results of the QQ Plot. Thus, we use the Satterthwaite's Approximation with t-distribution for computing the confidence interval.

$$\overline{SE}(\bar{r} - \bar{l}) = \sqrt{\frac{S_r^2}{n_r} + \frac{S_l^2}{n_l}}$$

We calculate the variance and mean which gives the output:

$$S_r^2 = 0.2925895, S_l^2 = 0.229322, n_r = n_l = 30$$

Solving Satterthwaite's Approximation with t-distribution:

$$\begin{aligned}
\overline{SE}(\bar{r} - \bar{l}) &= \sqrt{\frac{S_r^2}{n_r} + \frac{S_l^2}{n_l}} \\
&= \sqrt{\frac{0.2925895}{30} + \frac{0.229322}{30}} \\
&= \sqrt{\frac{0.52191}{30}} \\
&= 0.13189
\end{aligned}$$

We use the standard error to calculate lower bound and upper bound of the confidence intervals. The difference of mean $\bar{r} - \bar{l} = 0.3813333$. The value of $Z_{\frac{\alpha}{2}} = 1.96$

$$\begin{aligned}
\text{Lower bound of CI} &= (\bar{r} - \bar{l}) - Z_{\frac{\alpha}{2}} * \overline{SE}(\bar{r} - \bar{l}) \\
&= 0.381333 - 1.96 * 0.13189 \\
&= 0.12282
\end{aligned}$$

$$\begin{aligned}
\text{Upper bound of CI} &= (\bar{r} - \bar{l}) + Z_{\frac{\alpha}{2}} * \overline{SE}(\bar{r} - \bar{l}) \\
&= 0.381333 + 1.96 * 0.13189 \\
&= 0.63984
\end{aligned}$$

Thus, the 95% confidence interval is given by $(0.12282, 0.63984)$ We verify this, using t-distribution by the R code given below.

```r
# Step 1: Get the variance of remote and local voltage data
var_remote <- var(remote_voltage)
var_local <- var(local_voltage)

var_remote
```

```
## [1] 0.2925895
```

```r
var_local
```

```
## [1] 0.229322
```

```r
# Store length of remote and local voltage for better readability
remote_size <- length(remote_voltage)
local_size <- length(local_voltage)

# Step 2: Use the variance to calculate Standard Error
standard_err <- sqrt(var_local/local_size + var_remote/remote_size)

standard_err
```

```
## [1] 0.1318979
```

```r
# Step 3: Calculate difference of mean
# Using qnorm, create confidence interval
diff_mean <- mean(remote_voltage) - mean(local_voltage)

diff_mean
```

```
## [1] 0.3813333
```

```r
output <- diff_mean + c(-1, 1) * qnorm(0.975) * standard_err

output
```

```
## [1] 0.1228182 0.6398484
```

```r
t.test(remote_voltage, local_voltage, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  remote_voltage and local_voltage
## 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
```

The confidence interval given by R is $(0.1172284, 0.6454382)$. The null hypothesis is rejected because 0 does not lie in the confidence interval.

Thus, the difference in sample mean of remote and local locations is non-zero. We can conclude that we cannot establish manufacturing process at local locations.

(c)

In part b, we concluded that we cannot establish manufacturing process at local locations. From part a, we see that the voltage values in remote locals are high than that of the local locations.

Since we are establishing manufacturing process, intuitively we want to have higher voltage to sustain machinery. It would thus make sense to establish the process in remote locations based on the information we understood from parts a and b.

Question 3

Just like question 2, we compute qqplot and box plots using R.

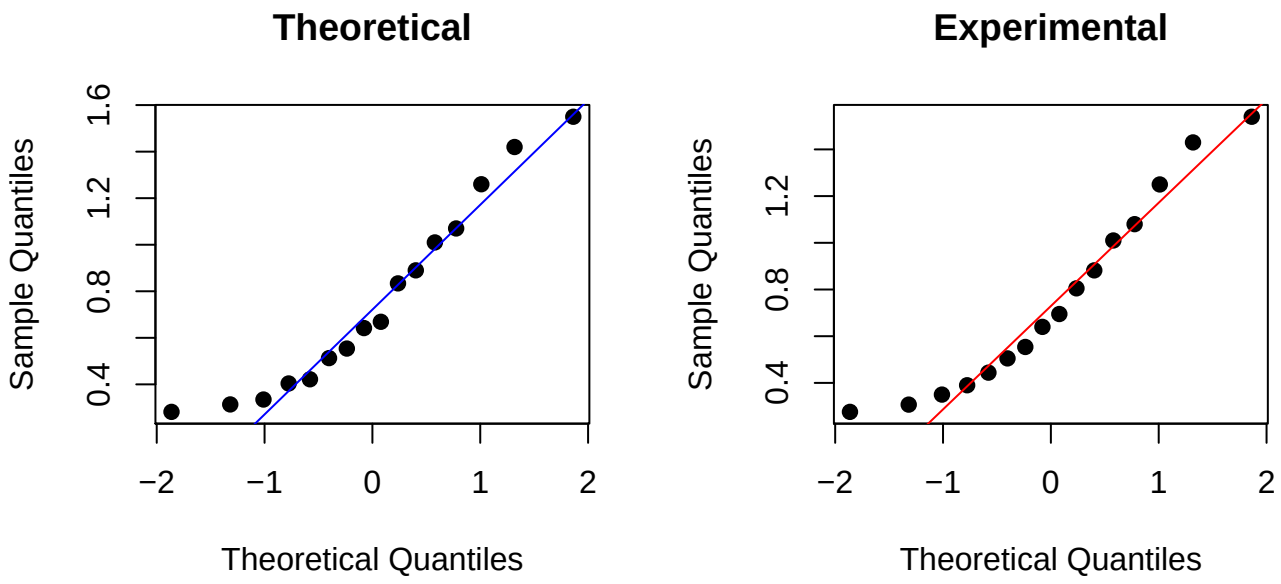
```
# read the csv file
vapor_data <- read.csv("VAPOR.csv")

# Store the exp and theoretical data
exp_data <- vapor_data$experimental
theoretical_data <- vapor_data$theoretical

# Get the summary of experimental and theoretical data
summary(exp_data)
```

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

```
summary(theoretical_data)
```

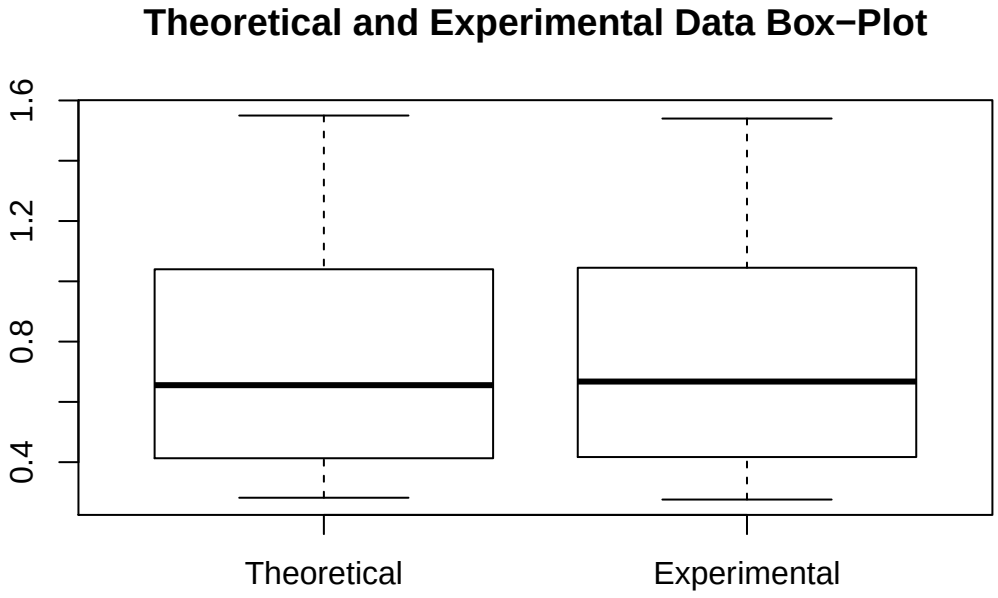
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2820  0.4175  0.6555  0.7606  1.0250  1.5500
```

```
# Draw qq plot with line for theoretical data
qqnorm(theoretical_data, main = "Theoretical", pch = 19)
qqline(theoretical_data, col = "blue")

# Draw qq plot with line for exp data
qqnorm(exp_data, main = "Experimental", pch = 19)
qqline(exp_data, col = "red")
```



We then draw side-by-side box plot for theoretical and experimental data.

```
boxplot(theoretical_data, exp_data,
        names = c("Theoretical", "Experimental"),
        main = "Theoretical and Experimental Data Box-Plot")
```



Since, both the box plot of the theoretical is nearly identical to box plot of experimental data , we can conclude that the two datasets are very similar. We can also see that both the datasets are right skewed.

The null hypothesis can be formulated using the statement `‘the true mean difference between the experimental and calculated values of vapor pressure will be zero.‘` from the question.

The null hypothesis will be $\bar{t} - \bar{e} = 0$

The alternative hypothesis will be: $\bar{t} - \bar{e} \neq 0$

Calculating, $t = 2.13145, \text{mean} = 0.0006875, \text{standard deviation} = 0.014216$

$$\text{Lower bound of CI} = (\bar{d} - t_{\frac{\alpha}{2},n-1}) * \frac{S_\Delta}{\sqrt{n}}$$

$$= 0.0006875 - 2.13145 * \frac{0.014216}{4}$$

$$= -0.006887$$

$$\text{Upper bound of CI} = (\bar{d} + t_{\frac{\alpha}{2},n-1}) * \frac{S_\Delta}{\sqrt{n}}$$

$$= 0.0006875 + 2.13145 * \frac{0.014216}{4}$$

$$= 0.0082626$$

Thus, we get the confidence interval $(-0.006887, 0.0082626)$ from the above calculations.

We verify the confidence interval using t-test in the R code given below (contd next page)

```r
# Calculate difference between theoretical and experimental data
diff_theoretical_exp <- theoretical_data - exp_data;
diff_theoretical_exp
```

```
##  [1]  0.006  0.007 -0.015  0.014 -0.022  0.008  0.000  0.002 -0.026  0.029
## [11]  0.008  0.000 -0.010  0.010 -0.010  0.010
```

```r
# Store length of diff as a separate variable for readability
size <- length(diff_theoretical_exp)
size
```

```
## [1] 16
```

```r
# Calculate Standard Error
standard_err <- sd(diff_theoretical_exp)/sqrt(size)
standard_err
```

```
## [1] 0.00355401
```

```r
# Calculate CI
res <- mean(diff_theoretical_exp) + c(-1, 1) * qt(0.975, size - 1) * standard_err
res
```

```
## [1] -0.006887694  0.008262694
```

```r
t.test(theoretical_data, exp_data, conf.level = 0.95, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  theoretical_data and exp_data
## 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
```

Unlike the question 2, 0 lies in the confidence interval $(-0.006887694, 0.008262694)$. Thus, our null hypothesis is acceptable and $\bar{t} - \bar{e} = 0$ is true. Thus, the mean difference between the experimental and calculated values of vapor pressure will be zero.