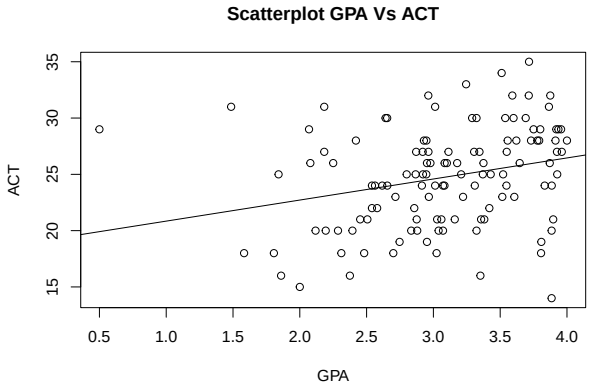
# Mini-Project 4 - Solo Group 10

```
library(boot)
```

```
read_csv_func = function(x){
  df = read.csv(x, header=TRUE) # Read CSV
  return (df)
}
```

**Solution 1**

```
gpa_df = read_csv_func("/Users/karthik_ragunath/Desktop/Stats/gpa.csv")
gpa_score = as.numeric(gpa_df$gpa)
act_score = as.numeric(gpa_df$act)
plot(gpa_score, act_score, xlab = "GPA", ylab = "ACT", main="Scatterplot GPA Vs ACT")
abline(lm(act_score ~ gpa_score))
```




## Inference

From the scatter plot we can infer that the slope is positive and the strength of the linear relationship between the two given variables is weak.

The correlation value is:

```
correlation_gpa_act = cor(gpa_score, act_score)
correlation_gpa_act
```

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

The Bootstrap Estimates Summary:

```
covariance.npar = function(gpa_df, indexes){
   gpa = gpa_df$gpa[indexes]
   act = gpa_df$act[indexes]
   result = cor(gpa, act)
   return(result)
}
covariance.npar.boot = boot(gpa_df, covariance.npar, R = 999, sim = 'ordinary', stype = "i")
covariance.npar.boot
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = gpa_df, statistic = covariance.npar, R = 999, sim = "ordinary",
##     stype = "i")
##
##
## Bootstrap Statistics :
##      original      bias    std. error
## t1* 0.2694818 0.003430071   0.1047612
```

The mean of the boostrap estimate

```
mean(covariance.npar.boot$t)
```

```
## [1] 0.2729119
```

The 95% confidence interval summary of the bootstrap estimate

```
boot.ci(covariance.npar.boot)
```

```
## Warning in boot.ci(covariance.npar.boot): bootstrap variances needed for
## studentized intervals
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 999 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = covariance.npar.boot)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.0607,  0.4714 )   ( 0.0658,  0.4648 )
##
## Level     Percentile            BCa
## 95%   ( 0.0741,  0.4732 )   ( 0.0488,  0.4557 )
## Calculations and Intervals on Original Scale
```

The 95% confidence interval computed using percentile bootstrap estimate

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

```
## [1] 0.07412646 0.47315738
```

# Inference

The estimate of correlation from bootstrap data is approximately the same as
the correlation value from the sample and the confidence interval from boot.ci
is approximately the same as the confidence interval computed theorotically.
As the correlation value is approximately 0.269, there is a positive association in
the scatter plot.

---

**Solution (2)**

**(a)**

Exploratory data analysis of voltage consumption in local and remote locations

Summary Statistics

```
voltage_df = read_csv_func("/Users/karthik_ragunath/Desktop/Stats/VOLTAGE.csv")
remote = voltage_df$voltage[which(voltage_df$location == 0)]
local = voltage_df$voltage[which(voltage_df$location == 1)]

summary(remote)
```

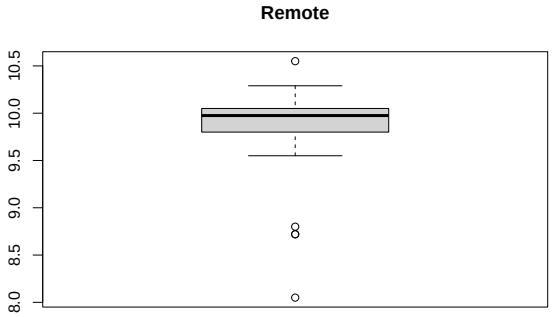
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.050   9.800   9.975   9.804  10.050  10.550
```

```
summary(local)
```

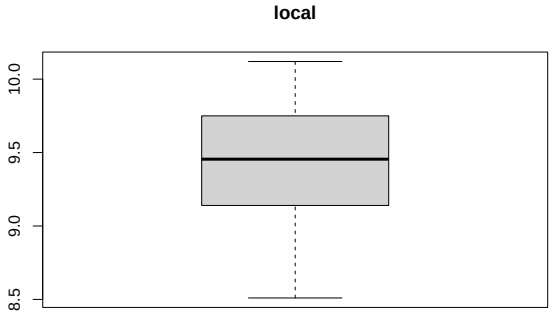
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.510   9.152   9.455   9.422   9.738  10.120
```

Box-Plot of Local and Remote Voltage Consumption Statistics

```
boxplot(remote,main="Remote")
```




```
boxplot(local,main="local")
```



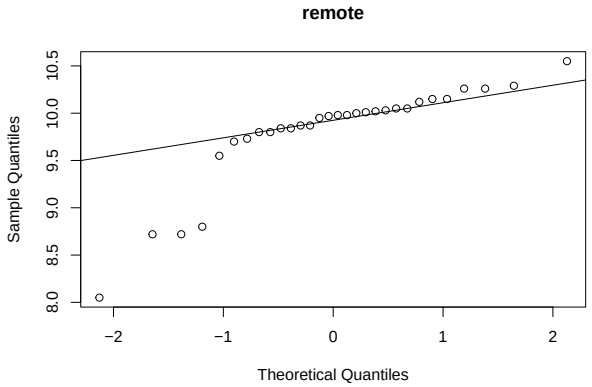

# Inferences

We can observe that voltage readings are higher for remote locations when compared to local locations. From the summary statistics we see that both local and remote locations are left skewed because their medians are greater than their corresponding mean. Also, there are outliers in remote data boxplot.

QQ-Plots

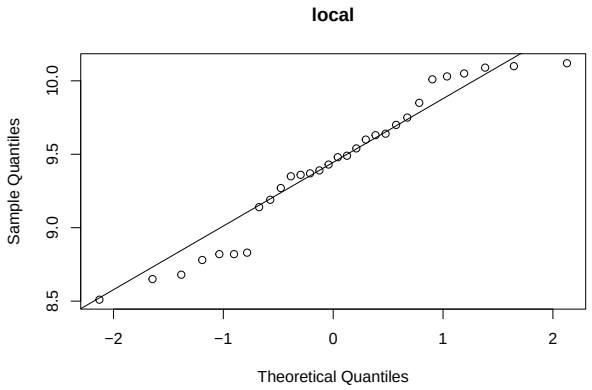
Remote

```
qqnorm(remote,main="remote")
qqline(remote)
```




Local

```
qqnorm(local,main="local")
qqline(local)
```



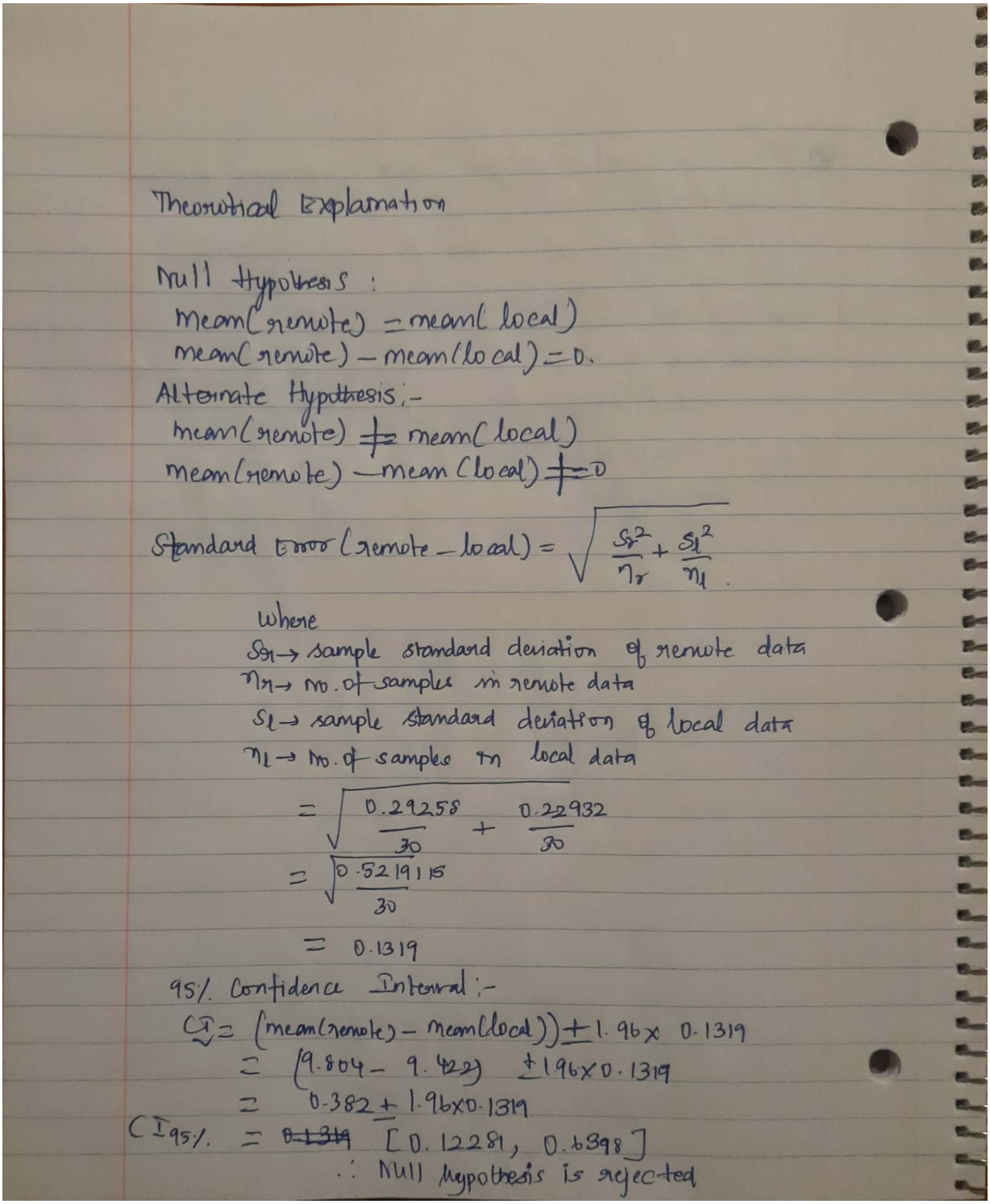

# Inference

The data points and the line of QQ-Plots coincide. Hence we can assume that the data are normalized. Therefore, we can conclude that both the distribution are similar in the sense that they follow normal distribution.

---

**(b)**

We will compute 95% confidence interval values theorotically and also through programming means with the given and perform hypothesis testing assuming normal distribution.




Scanned with CamScanner

Figure 1: Theorotical Computation

---

**Computing Confidence Interval Programmatically**

Computing for 95% Confidence Interval assuming Normal Distribution

```
standard_error = sqrt(var(local)/30 + var(remote)/ 30)
standard_error
```

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

```
confidence_interval = (mean(remote) - mean(local)) + c(-1,1)*qnorm(0.975) * standard_error
confidence_interval
```

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

Performing T-Test

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

```
##
##  Welch Two Sample t-test
##
## data:  remote and 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
```

# Inference

In order to verify the Confidence Interval computed under normal distribution assumption, we perform the T test. The T-Test results verify our CI computation. Therefore, our computed CI is correct and the data can be assumed to be normalized. Moreover, 0 doesn't lie in the CI interval so the null hypothesis is rejected which implies that the mean difference between the remote and local location is not zero. Hence we can conclude that the manufacturing process cannot be established locally.

---

(c)

It is clear from (b) that the voltage requirements are higher for remote locations when compared to local needs since the Confidence Interval of mean difference between remote and local data doesn't include 0.
From (a), it is clear from box-plots and summary statistics that remote location has

higher voltage requirement than local locations.
Therefore findings from (a) and (b) corresponds with each other.
From our findings, we can conclude that the manufacturing unit must be setup at remote location.

---

**Solution (3)**

Exploratory Data Analysis

Summary Statistics

```
vapor = read.csv("/Users/karthik_ragunath/Desktop/Stats/VAPOR.csv")
theoretical = vapor$theoretical
experimental = vapor$experimental
summary(theoretical)
```

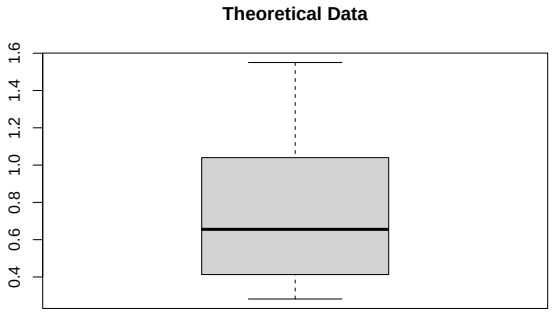
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2820  0.4175  0.6555  0.7606  1.0250  1.5500
```

```
summary(experimental)
```

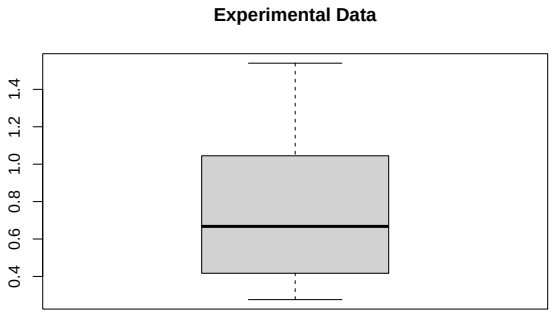
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2760  0.4305  0.6675  0.7599  1.0275  1.5400
```

Plotting box plots

```
boxplot(theoretical,main="Theoretical Data")
```




```
boxplot(experimental,main="Experimental Data")
```

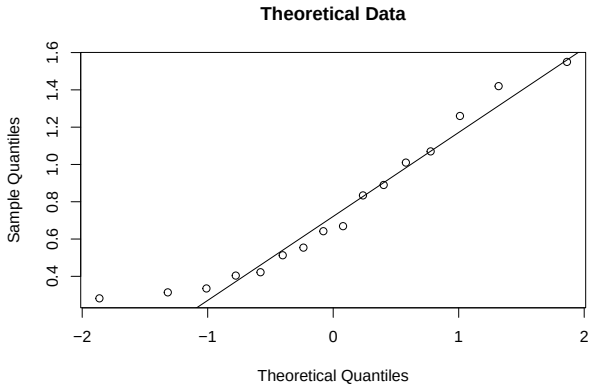



# Inference

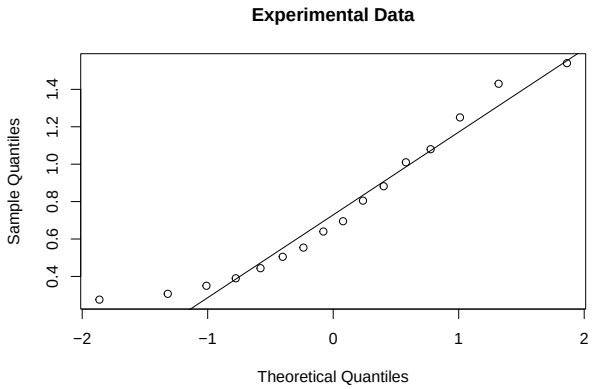
Both theorotical and experimental data distributions are right skewed since their their median are lesser than their corresponding mean.

Plotting QQ-Plots

```
qqnorm(theoretical,main="Theoretical Data")
qqline(theoretical)
```




```
qqnorm(experimental,main="Experimental Data")
qqline(experimental)
```




# Inference

The data points and the line coincide hence we assume that the data are normalized.

Computing 95% Confidence Interval

```
data_diff = theoretical-experimental
conf_interval=(mean(theoretical)-mean(experimental))+
   c(-1,1)*qt(0.975, 15)*sd(data_diff)/sqrt(16)
conf_interval
```

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

Performing T-Test for verification

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

```
##
##  Paired t-test
##
## data:  theoretical and 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
```

# Inference

Null hypothesis- The  mean difference between the experimental and theoretical values are zero. Alternative hypothesis - The mean  difference between the

experimental and theoretical values are not zero.
Since 0 is present in the confidence interval, the null hypothesis is accepted which implies that the sample mean is approximately same as theorotical mean. Therefore, we can conclude that the null hypothesis case that the experimental and theorotical mean are equal is true.

---