Mini Project 4

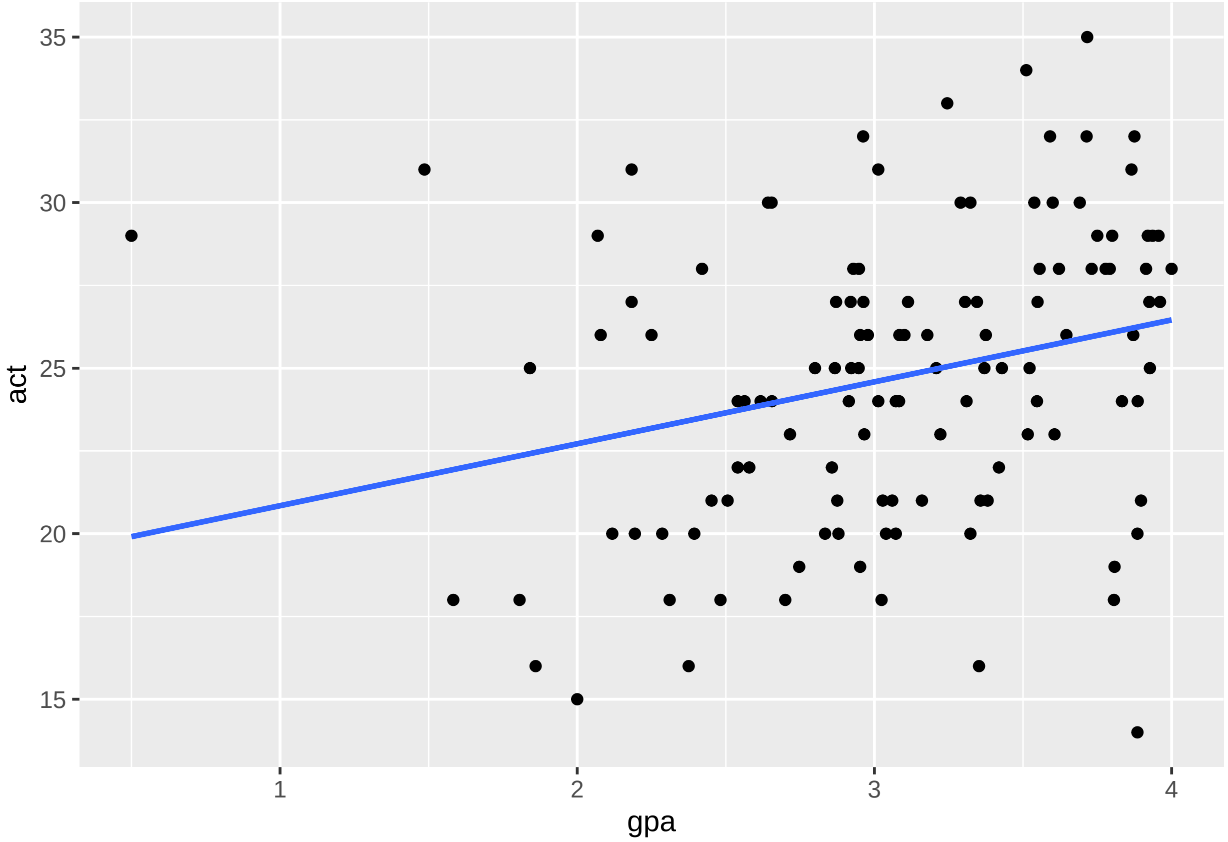
CS6313.001

Rutvij Shah (rds190000)

05 November 2021

Question 1

```
library(ggplot2)

gpa_act = read.csv("~/Downloads/gpa.csv")
gpa = gpa_act$gpa; act = gpa_act$act;

ggplot(gpa_act, aes(x=gpa, y=act)) +
  geom_point() +
  geom_smooth(method="lm", se=F, formula = y~x)
```



```
sample_correlation = round(cor(gpa, act), 3)
cat(
  paste("\nThe sample correlation for GPA and ACT scores is:",
        sample_correlation)
  )
```

```
##
## The sample correlation for GPA and ACT scores is: 0.269
```

This suggests there is a slight positive linear correlation between the two samples.

```
library(boot)
cor.npar <- function(x, indices) {
  result <- cor(x[indices,]$gpa, x[indices,]$act)
  return(result)
}

cor.npar.boot <- boot(data=gpa_act, cor.npar, R=999, sim="ordinary", stype="i")

cor.gpa.act <- cor.npar.boot$t0

bootstrap.bias.cor <- mean(cor.npar.boot$t) - cor.npar.boot$t0

bootstrap.se.cor <- sd(cor.npar.boot$t)

# For verification
percentile.95.ci.verif <- boot.ci(boot.out = cor.npar.boot, type="perc")

percentile.95.ci.bot <- quantile(cor.npar.boot$t,0.025, names=F)
percentile.95.ci.top <- quantile(cor.npar.boot$t,0.975, names=F)

cat(paste(
  "\nPoint Estimate of correlation between GPA and ACT scores",
  "\U03C1 =", round(cor.gpa.act, 3), "\n"))
```

```
##
## Point Estimate of correlation between GPA and ACT scores   = 0.269
```

```
cat(paste(
  "\nBootstrap Esitmate of bias for \U03C1 = ",
  round(bootstrap.bias.cor, 3), "\n"))
```

```
##
## Bootstrap Esitmate of bias for   =  0.002
```

```
cat(paste(
  "\nBootstrap Esitmate of SE for \U03C1 = ",
  round(bootstrap.se.cor, 3), "\n"))
```

```
##
## Bootstrap Esitmate of SE for   =  0.106
```

```
ci = paste("[", round(percentile.95.ci.bot, 3),
  ", ", round(percentile.95.ci.top, 3), "]", sep="")

cat(paste(
    "\nThe 95% CI for (percentile bootstrap based) for \U03C1 -> ",
    ci, "\n"))
```

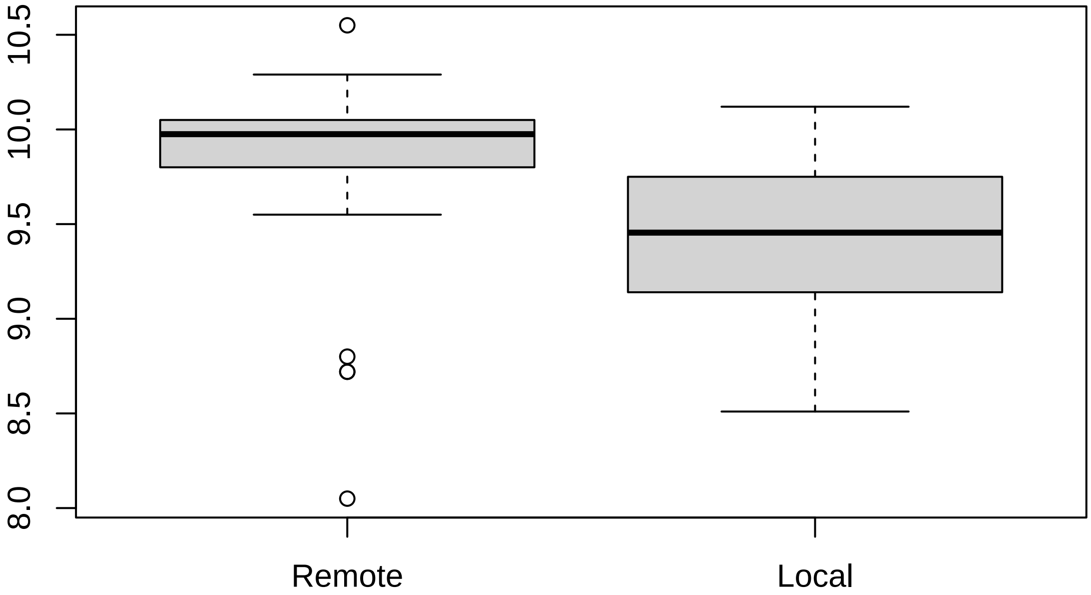
```
##
```

```
## The 95% CI for (percentile bootstrap based) for  ->  [0.061, 0.474]
```

Question 2

(a)

```
voltage_data = read.csv("~/Downloads/voltage.csv")
v_remote = voltage_data[voltage_data$location == 0,][,'voltage']
v_local = voltage_data[voltage_data$location == 1,][,'voltage']


boxplot(v_remote, v_local, names = c("Remote", "Local"), range = 1.5)
```



```
cat("Five point summary + mean for remote voltage.\n")
```

```
## Five point summary + mean for remote voltage.
```

```
summary(v_remote)
```

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

```
cat("Five point summary + mean for local voltage.\n")
```

```
## Five point summary + mean for local voltage.
```

```
summary(v_local)
```

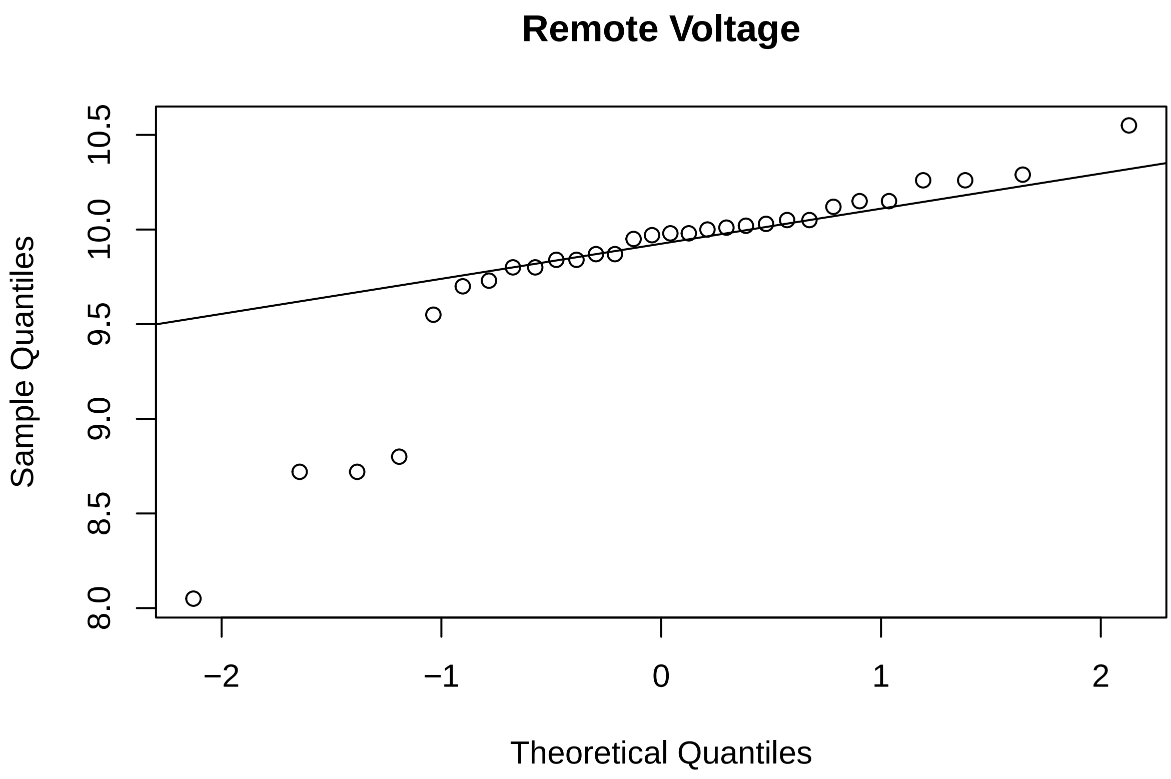
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.510   9.152   9.455   9.422   9.738  10.120
```

The readings for the remote location are higher, on average when compared to those which are local. And the five point summary reiterates the graphical observation from the box plot.

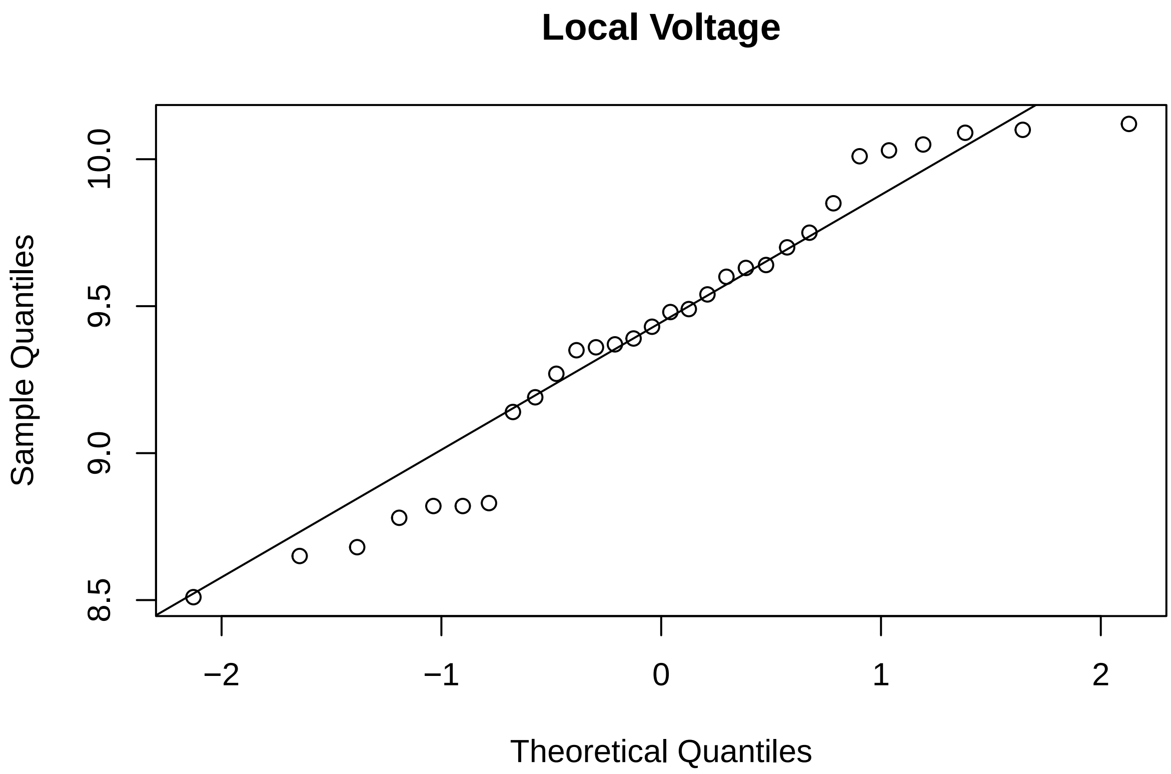
For both sets of data, the mean is less than the median, thereby suggesting the presence of a left-skew.

Now to check the normality of the data.

```
qqnorm(v_remote, main = "Remote Voltage"); qqline(v_remote)
```



```
qqnorm(v_local, main = "Local Voltage"); qqline(v_local)
```



We can observe that both distribution's qqplots suggest that the distributions can be considered to be approximations of a normal distribution.

(b)

Null Hypothesis: $\mu_{remote} - \mu_{local} = 0$ Alternate Hypothesis: $\mu_{remote} - \mu_{local} \neq 0$

Assuming the samples are i.i.d (proof of normalization based on qqplots). Though, since IQRs have a large difference, we cannot assume the population variances are equal.

Thus, we must use the Satterthwaite approximation to approximate a T-distribution.

Degrees of freedom for CI of a difference of means, given unequal, unknown standard deviations:

$$\nu = \left[(\frac{s_X^2}{n} + \frac{s_Y^2}{m})^2 \middle/ \frac{s_X^4}{n^2(n-1)} + \frac{s_Y^4}{m^2(m-1)} \right]$$

Where

```
cat(paste("s\U2093 = "), round(sd(v_remote), 3))
```

```
## s  =  0.541
```

```
cat(paste("\nn =", length(v_remote)))
```

```
##
## n = 30
```

```
cat(paste("\n\ns\U1d67 = "), round(sd(v_local), 3))
```

```
##
##
## s  =  0.479
```

```
cat(paste("\nm =", length(v_local)))
```

```
##
## m = 30
```

Therefore,

$$\nu = \left[(\frac{0.522}{30})^2 \middle/ \frac{0.138}{26100} \right]$$

$$\nu = \frac{0.272 * 26100}{0.138 * 900}$$

Degrees of freedom, $\nu = 57.16$

Assuming we want 95% CI, $\frac{\alpha}{2} = .025$

The CI is,

$$\bar{X} - \bar{Y} \pm t_{0.025}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

```
cat(paste("T's Critical Value", round(qt(.975, 57.16),3)))
```

```
## T's Critical Value 2.002
```

```
cat(paste("\nMean of Remote Voltage Sample", round(mean(v_remote),3)))
```

```
##
## Mean of Remote Voltage Sample 9.804
```

```
cat(paste("\nMean of Local Voltage Sample", round(mean(v_local),3)))
```

```
##
## Mean of Local Voltage Sample 9.422
```

```
# diff = mean(v_remote) - mean(v_local)
# ci = diff + c(-1, +1) * qt(.975, 57.16) * sqrt(0.522/30)
# t = diff / sqrt(0.522/30)
```

$$9.804 - 9.422 \pm 2.002 * \sqrt{\frac{0.522}{30}}$$

$$0.381 \pm 0.264$$

ANS i.e. the CI is [0.117, 0.645]

For a two-sided alternative, we reject H_0 if $|t| \geq t_{\alpha/2}$ and accept otherwise.

$$t = \frac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

$$t = \frac{9.804 - 9.422}{\sqrt{\frac{0.522}{30}}} = \frac{0.062}{0.132}$$

Since $t = 2.89$ is within the rejection region of $(-\infty, -2.002] \cup [2.002, \infty)$ we **reject** the null hypothesis & state that the remote process cannot be localized. Alternatively, the p value is significantly less than $\alpha = 0.05$, we **reject** H_0

```
# confirmation of manual calculations & findings
t.test(v_remote, v_local)
```

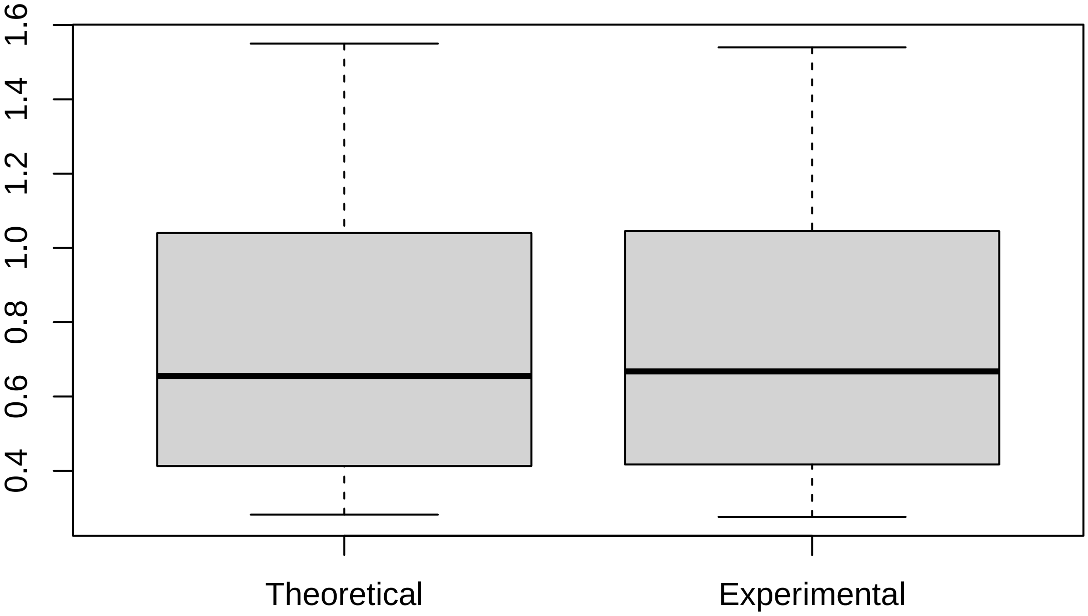
```
##
##  Welch Two Sample t-test
##
## data:  v_remote and v_local
## t = 2.8911, df = 57.16, p-value = 0.005419
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1172284 0.6454382
## sample estimates:
## mean of x mean of y
##  9.803667  9.422333
```

(c)

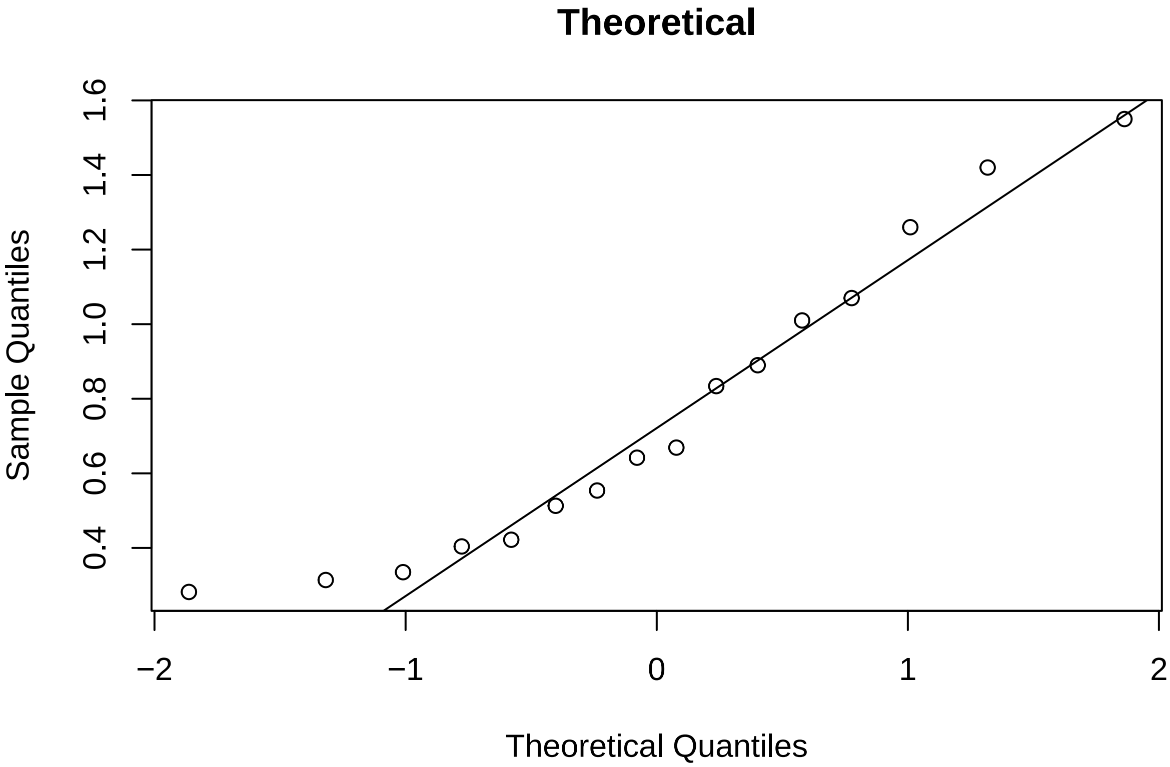
Part (a) suggested that the voltage readings at the remote location are consistently higher than those taken locally. The conclusion from the hypothesis test confirms that observation and makes it statistically evident that the process cannot be localized without significant error/change in metric.

Question 3

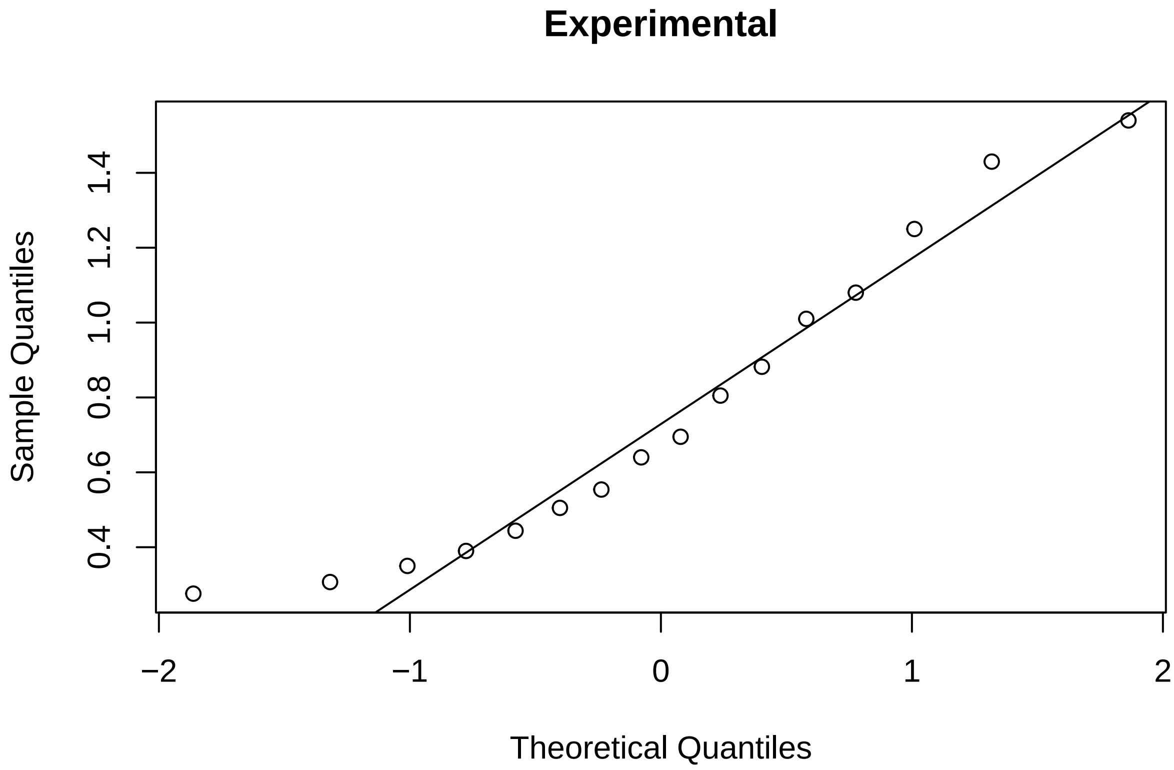
```
vapor_data = read.csv("~/Downloads/vapor.csv")

boxplot(vapor_data$theoretical, vapor_data$experimental,
        names=c("Theoretical", "Experimental"))
```



```
qqnorm(vapor_data$theoretical, main="Theoretical"); qqline(vapor_data$theoretical)
```



```
qqnorm(vapor_data$experimental, main="Experimental"); qqline(vapor_data$experimental)
```



```
summary(vapor_data[,-1])
```

```
##   theoretical      experimental   
##  Min.   :0.2820   Min.   :0.2760  
##  1st Qu.:0.4175   1st Qu.:0.4305  
##  Median :0.6555   Median :0.6675  
##  Mean   :0.7606   Mean   :0.7599  
##  3rd Qu.:1.0250   3rd Qu.:1.0275  
##  Max.   :1.5500   Max.   :1.5400  
```

The boxplots suggest that both sets of data have similar distributions & the qqplots suggest that they are close to normally distributed.

The summary statistics suggest that they're right skewed distributions (mean > median) and backup the observation that their SDs/Variances are similar.

Thus, our problem becomes that of finding a confidence interval (and testing a hypothesis) for the difference of means for a small sample (16 readings) from two distributions which have equal but unknown standard deviations.

Which necessitates the use of a T-Test, with the given conditions.

Null Hypothesis = $\mu_{theoretical} = \mu_{experimental}$ Alternate Hypothesis = $\mu_{theoretical} \neq \mu_{experimental}$

```
t.test(vapor_data$theoretical, vapor_data$experimental, var.equal =TRUE, paired = TRUE)
```

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

The Two-Sample T test, confirms the Null Hypothesis, the mean of the difference between experimental & theoretical mean is close to 0 and a p-value of 0.8492 derived from a t value of 0.019 suggests that with a high degree of confidence.

Thus, we can indeed say that the theoretical model for vapor pressure is a good model of reality.