Mini project #4

Group Member: Chaoran Li, Wenting Wang

Contribution of each member:

Firstly, we discussed the mathematical models and code details together. Then, we divided the project into two part and finished our respective work. Wenting Wang mainly worked on Q1 and analysis in Q2 while Chaoran Li worked on Q3 and code in Q2. Then, we merged our code and solution into one report. Each member makes contribution to each sub task of this project and combines all to finish this project, as the details shown in table 1.

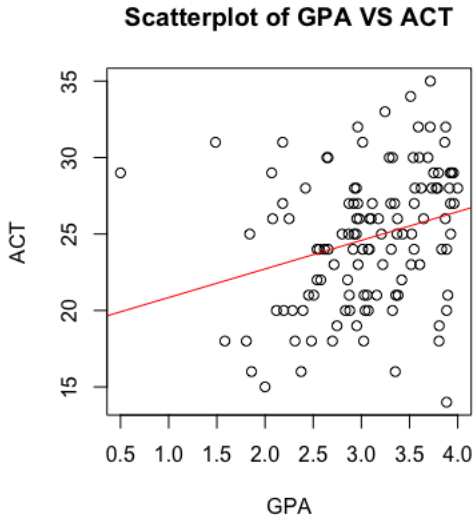
	Question1	Question2	Question3
Chaoran li	20%	50%	20%
Wenting wang	80%	50%	80%

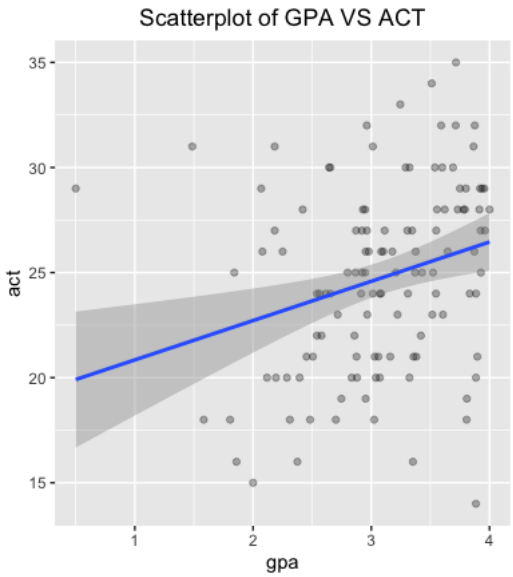
Table 1: Member contribution table

Question 1:

In order to make a scatterplot of GPA against ACT and comment on the strength of linear relationship between the two variables. We first read and store the data into two lists, and then draw the scatter plot as below:

```
> # Take a scatterplot of gpa against act
> student = read.csv(file=file.path("./Mini Project 4/gpa.csv"))
> gpa = student$gpa
> act = student$act
> plot(gpa, act, main="Scatterplot of GPA VS ACT", xlab="GPA", ylab="ACT")
> abline(lm(act~gpa), col="red")
> # Better scatterplot with ggplot2
> library(ggplot2)
> ggplot(data=student, aes(x=gpa, y=act)) + # scatterplot
+   geom_point(alpha=0.3) + # use alpha to represent frequency
+   stat_smooth(method="lm", formula = y ~ x) + # linear relationship
+   labs(title = "Scatterplot of GPA VS ACT") + # add title
+   theme(plot.title = element_text(hjust = 0.5)) # center the title
```





In ggplot2, we can use the shade to mark the overlap of data points. (Shown on the right.)

As we can see from this scatter plot that there exists the positive correlation between GPA and ACT, which means that a student who has a higher GPA would also have a higher ACT score. But as the slope of the red regression line is very flat, so the strength of the linear relationship between the two variables is weak.

```
> # Bootstrap for the correlation between gpa and act
> # Use install.package("boot")
> library(boot)
> cor(gpa, act)
[1] 0.2694818
```

Based on the given sample we can estimate the population correlation using cor function is 0.2694818

```
> # Non-parameteric bootstrap function
> covarience.npar <- function(X, indices) {
+   g = X$gpa[indices]
+   a = X$act[indices]
+   cor(g, a)
+ }
> covarience.npar.boot = boot(student, covarience.npar,
+                             R=999, sim="ordinary", stype="i")
> covarience.npar.boot

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = student, statistic = covarience.npar, R = 999, sim = "ordinary",
    stype = "i")


Bootstrap Statistics :
     original     bias    std. error
t1* 0.2694818 0.00312047   0.1068602
>
```

In order to get more statistics of the correlation, we need to resample the data and use non-parameter bootstrap. The details are shown in the code. Then we can get:

Sample correlation = 0.2694818 Bias= 0.00312047 Standard Error = 0.1068602

```
> # Confidence Interval
> mean(covarience.npar.boot$t)
[1] 0.2726023
> var(covarience.npar.boot$t)
           [,1]
[1,] 0.0114191
> # Get the 95% confidence interval by Percentile Bootstrap Method
> sort(covarience.npar.boot$t)[c(25, 975)]
[1] 0.05593857 0.47224559
>
```

Last, we want to find the 95% confidence interval computed using percentile bootstrap. Thus, we just need to sort the bootstrap correlations, and get the 1st and 3st quartiles is:

[0.05593857, 0.47224559]

```
> # Get the 95% confidence interval by boot.ci to check with percentile method
> boot.ci(covarience.npar.boot, conf=0.95, type="perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

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

Intervals :
Level     Percentile
95%   ( 0.0559,  0.4722 )
Calculations and Intervals on Original Scale
> # Specifically declare to use percentile method to avoid warning:
> # In boot.ci(covarience.npar.boot) :
> #   bootstrap variances needed for studentized intervals
>
```

We also use boot.ci to check this result: the result [0.0559, 0.4722] match the percentile method very good.

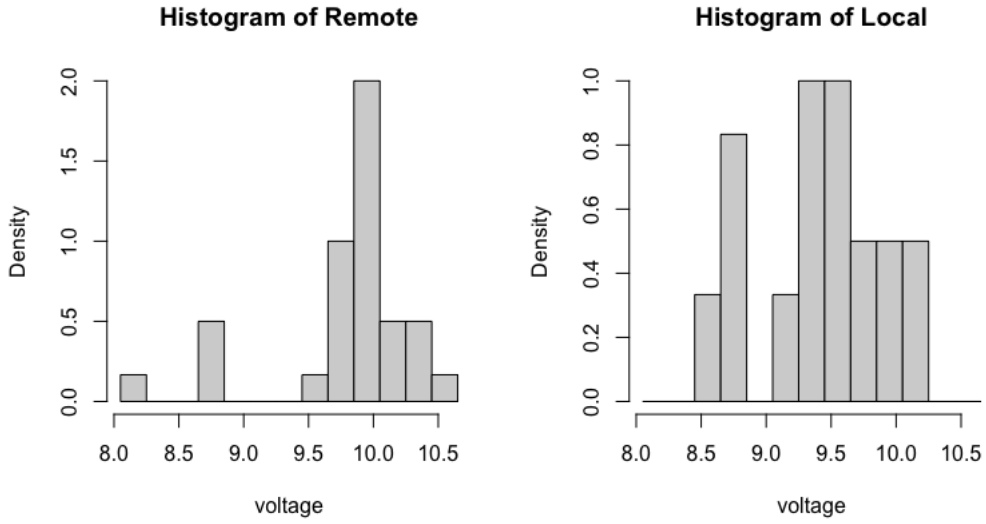
This result also shows that the correlation between GPA and ACT is positive, and the estimate values of both the sample and bootstrap are around 0.27.

Question 2:

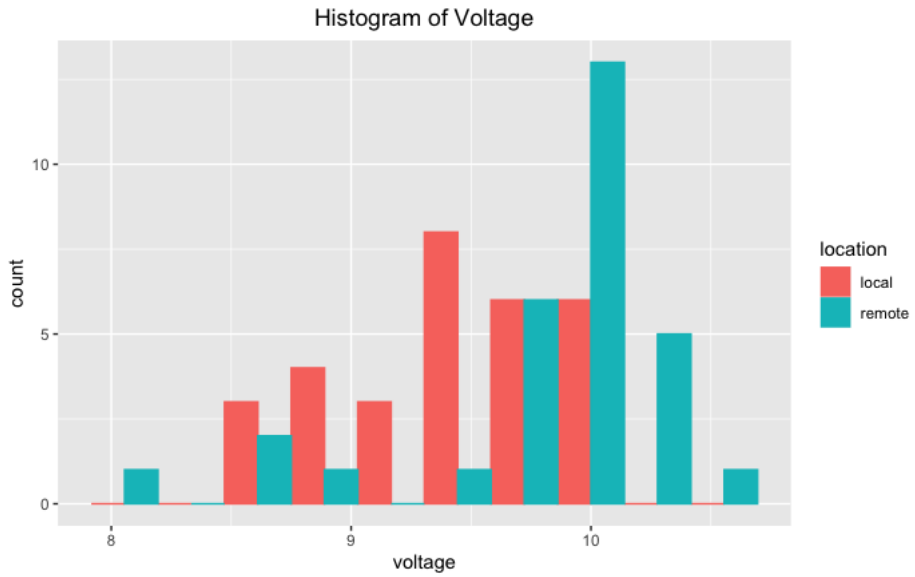
(a) From the boxplot below, we can find that both the mean and variance of the two data sets are different, and the average voltage of remote is higher than local.

Histograms:

```
> # (a)
> voltage = read.csv(file=file.path("./Mini Project 4/VOLTAGE.csv"))
>
> voltage.remote = voltage$voltage[voltage$location == 0]
> voltage.local = voltage$voltage[voltage$location == 1]
>
> voltage.histInterval = 0.2
> voltage.breaks = seq(min(min(voltage.remote), min(voltage.local)),
+                      max(max(voltage.remote), max(voltage.local)) +
+                        voltage.histInterval,
+                      voltage.histInterval)
> hist(voltage.remote, breaks=voltage.breaks, probability=T,
+      xlab="voltage", main="Histogram of Remote")
> hist(voltage.local, breaks=voltage.breaks, probability=T,
+      xlab="voltage", main="Histogram of Local")
```

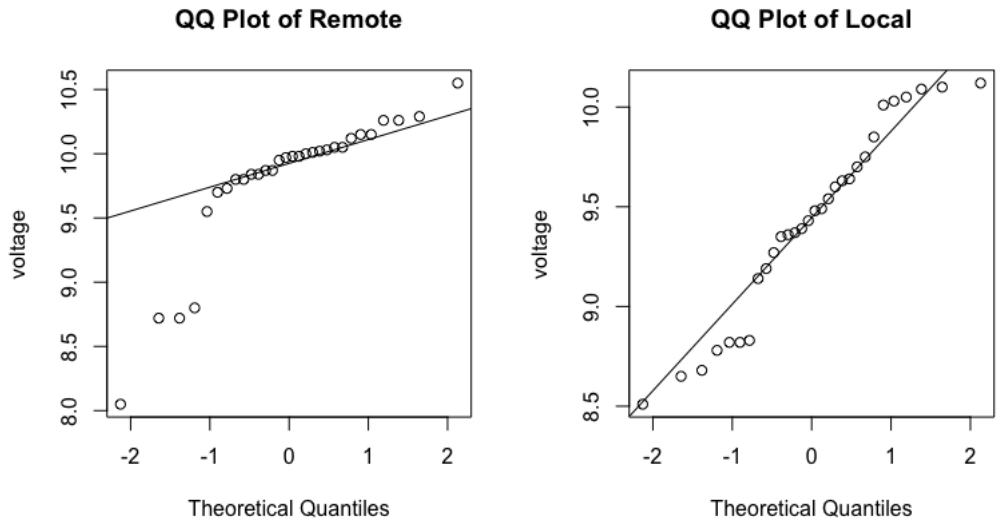



```
> # Draw two histograms together with ggplot2
> voltage.locationName = ifelse(voltage$location == 0, "remote", "local")
> df <- data.frame(
+   location = voltage.locationName,
+   voltage = voltage$voltage
+ )
> ggplot(df, aes(x=voltage, color=location, fill=location)) +
+   geom_histogram(bins=10, position="dodge") + # histgram
+   labs(title = "Histogram of Voltage") + # add title
+   theme(plot.title = element_text(hjust = 0.5)) # center the title
>
```



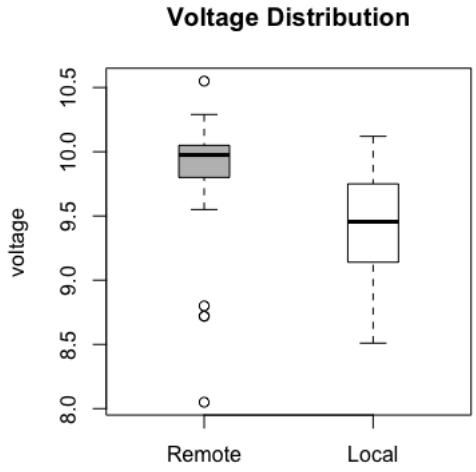
QQplots:

```
> qqnorm(voltage.remote, ylab="voltage", main="QQ Plot of Remote")
> qqline(voltage.remote)
> qqnorm(voltage.local, ylab="voltage", main="QQ Plot of Local")
> qqline(voltage.local)
```



From the QQplots, we can see that for both the two datasets, most points are just on the line or around the QQline, except some outliers. Thus, they can be considered to be normalized.

```
> boxplot(voltage.remote, voltage.local, boxwex=.3, names=c("Remote", "Local"),
+         col=c("grey", "white"), main="Voltage Distribution", ylab="voltage")
>
```



The boxplot shows that the distribution of voltage between remote and local is different. There are more outliers in voltage of remote. The distribution of remote is more left-skewed than that of local.

Besides, whole distribution of local fits normal distribution quite well in QQ plot. But the left part of the distribution of remote fits normal distribution not well.

All these prove that the distribution between remote and local seem different.

(b) In order to make the decision, which is a better choice for the manufacturing, we test whether there exists difference in the population means of voltage readings at the two locations. So, we set up the hypothesis as following:

Null hypothesis H0: difference= 0

Alternative hypothesis H1: difference≠ 0

Here, we need some assumptions:

(1) The two sample datasets are independent;

(2) The two distributions are normal.

Besides, the sample size is 30 which is also large enough, Thus, we can calculate the ci with qnorm(0.975).

```
> # (b)
> # Confidence Interval
> voltage.center = mean(voltage.remote - voltage.local)
> voltage.remote.se = var(voltage.remote) / length(voltage.remote)
> voltage.local.se = var(voltage.local) / length(voltage.local)
> voltage.margin = qnorm(0.975) * sqrt(voltage.remote.se + voltage.local.se)
> voltage.ci = voltage.center + c(-1, 1) * voltage.margin
> voltage.ci
[1] 0.1228182 0.6398484
>
```

Then, we can calculate the 95% CI for the difference is: [0.1228182, 0.6398484]. Because 0 is not belong to this interval, and all the values in this interval is bigger than 0, the remote location is a better choice. Thus, the manufacturing process cannot be established locally

```
> # T-test
> t.test(voltage.remote, voltage.local, alternative="two.sided",
+        paired=F, var.equal=F, conf.level=0.95)

        Welch Two Sample t-test

data:  voltage.remote and voltage.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

>
```

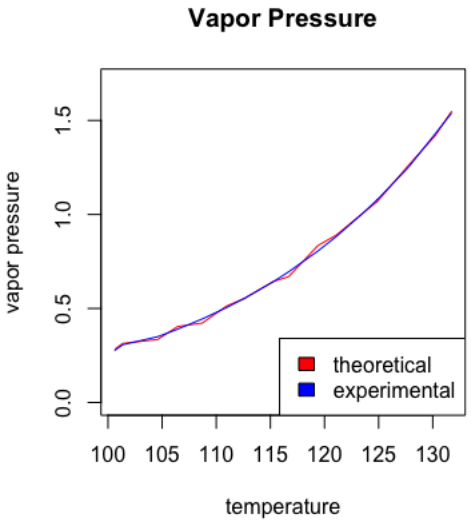
With t.test, we can check our result. Both $0 \notin [0.1172284, 0.654382]$ and p-value $= 0.005419 < 0.05$ prove that H0 is rejected which means the difference is not zero.

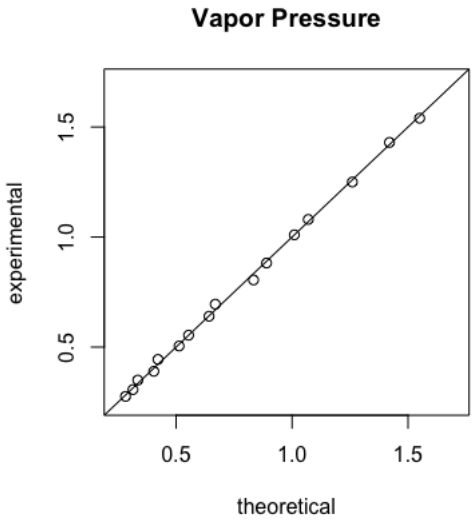
(c) From both (a) and (b), we can conclude that the expected voltage value of remote is higher than local. Thus, the manufacturing process should be established in remote location.

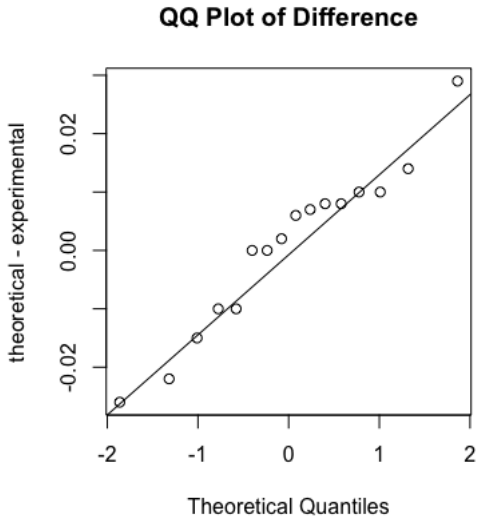
Question 3:

Firstly, let's do some analysis on VAPOR.csv by three diagrams. For the data represent the trend of vapor pressure by different temperature. I think we should focus on the comparison between theoretical and experimental data instead of discussing about the distribution. Hence, I drew these diagrams to show the difference.

```
> # Question 3
> # Analysis about the theoretical model for vapor pressure.
> vapor = read.csv(file=file.path("./Mini Project 4/VAPOR.csv"))
>
> # Diagram 1
> lim = c(0, max(max(vapor$theoretical), max(vapor$experimental)) * 1.1)
> plot(x=vapor$temperature, y=vapor$theoretical, type="l", col="red",
+      xlab="temperature", ylab="vapor pressure", ylim=lim,
+      main=paste("Vapor Pressure"))
> lines(x=vapor$temperature, y=vapor$experimental, type="l", col="blue")
> legend("bottomright", c("theoretical", "experimental"), fill=c("red", "blue"))
>
> # Diagram 2
> lim = c(min(min(vapor$theoretical), min(vapor$experimental)) * 0.9,
+         max(max(vapor$theoretical), max(vapor$experimental)) * 1.1)
> plot(x=vapor$theoretical, y=vapor$experimental, type="p", xlim=lim, ylim=lim,
+      xlab="theoretical", ylab="experimental", main=paste("Vapor Pressure"))
> abline(0, 1)
>
> vapor.difference = vapor$theoretical- vapor$experimental
>
> # Diagram 3
> qqnorm(vapor.difference, ylab="theoretical - experimental",
+        main="QQ Plot of Difference")
> qqline(vapor.difference)
>
```







From the diagrams, we can see that the theoretical results is quite similar to the experimental. The distribution of the difference is around zero. Hence, we could conclude that the difference between the experimental and calculated values tends to be zero. The theoretical model looks like a good one.

This is a paired two-sample test.

Assume:

X: difference between experimental and theoretical values of the vapor pressure for dibenzothiophene.

$\boldsymbol{\mu} = \mathrm{E}(\mathrm{X})$

H_0: $\boldsymbol{\mu} = 0$, good model.

H_1: $\boldsymbol{\mu} \neq 0$, bad model.

Although the question did not say that the X is follow a normal distribution. From the QQ plot, we can assume that:

X is normal and we do not know the variance.

```
> # Confidence Interval
> vapor.n = length(vapor$temperature)
> vapor.mean = mean(vapor.difference)
> vapor.sd = sd(vapor.difference)
> vapor.margin = qt(0.975, vapor.n -1) * vapor.sd / sqrt(vapor.n)
> vapor.ci = vapor.mean + c(-1, 1) * vapor.margin
> vapor.ci
[1] -0.006887694  0.008262694
>
> # T-test
> t.test(vapor$theoretical, vapor$experimental, alternative="two.sided",
+        paired=T, var.equal=F, 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

> qt(0.975, vapor.n -1)
[1] 2.13145
>
```

1): CI: $0 \in [-0.006887694, 0.008262694]$;

2) $|t_{obs}| = |0.19344| < 2.13145 = qt(0.975, 15)$;

3) p-value = 0.8492 > 0.05.

1), 2) and 3) prove that H_0 is accepted. Hence, it is a good model.