# Assign2

Frank Guo

2024-03-26

## Answer Sheet for Stats 707 Assignment 2

### Question 1

It is well known that many people try to improve their chances of winning a lottery by using lucky numbers or buying their tickets from certain lucky stores. To estimate what proportion of population would buy from a lucky store, a question was asked in a class survey. Participants were asked to respond Yes/No to the following question:
"If you learned that a certain store has sold the winning lotto ticket a few times in the past year, would you be tempted to buy a ticket from this store?"
In the survey, 12 out of 43 respondents said 'Yes'. These responses are stored in an Excel file luckylotto.CSV on Canvas.

Task 1:

a. Based on this data, what will be your estimate for the proportion of people in the whole population that would consider buying their tickets this way? [5 Marks].

b. Based on the key concepts of Statistical Inference that you have learned so far, Explain why this estimate can be considered to be a good estimate? [5 Marks].

Task 2:

a. Find the confidence intervals for your estimate at the 95% and the 99% levels and interpret each. Specify what the 'confidence' actually means and what it doesn't. [6 Marks].
b. Explain why the 99% confidence interval is wider than the 95% one? [4 Marks].

Task 3:

It is of interest to see if the proportion of people in the population who would consider buying their ticket from a lucky store is 0.5, or if it is less than that. Use a t-test to answer this question.
a. Clearly define the null and alternative hypotheses that are most appropriate for this question.[4 Marks].

b. Perform the t-test and paste your output in your report.[2 Marks].

c. Interpret the findings of the test. [ 4 Marks].

### Answer for Question 1:

#### Answer for Task 1:

```
df <-  read.csv('data/luckylotto.csv',header=TRUE,sep=',')
sample <- df %>% pull('response_code')
estimator <- prop <-  mean(sample)
estimator
```

```
## [1] 0.2790698
```

**a:**

Use mean of response_code as an estimator for the proportion of people in the whole population. the mean of response_code equals the proportion of sample. Hence, use ***Sample mean*** as the estimator of ***population proportion***. in this case, the proportion equels to the mean of response_code.

**b:**.
Use the mean of response_code as an estimator of proportion of whole population because of:
the sample is random and independent. the sample size is 45, larger than 30. the Sample Mean is an unbiased estimator of population mean.

**Answer for Task2:**

```
t1 <- t.test(sample,conf.level = 0.95)
print('confidence intervals for 95%:')
```

```
## [1] "confidence intervals for 95%:"
```

```
t1$conf.int
```

```
## [1] 0.1393953 0.4187443
## attr(,"conf.level")
## [1] 0.95
```

```
t2 <- t.test(sample, conf.level = 0.99)
print('confidence interval for 99%:')
```

```
## [1] "confidence interval for 99%:"
```

```
t2$conf.int
```

```
## [1] 0.09233251 0.46580702
## attr(,"conf.level")
## [1] 0.99
```

**a.**.
The confidence interval at 95% confident level is : 0.1393953 0.4187443.
I have 95% confident that the confidence interval that we get from our sample
contains the true proportion of population.
The confidence interval at 99% confident level is : 0.09233251 0.46580702.
I have 99% confident that the true proportion of population is inside this interval. The "confidence" here is applicable to the process by which confident interval are found,not the probability that the true Proportion lies in the interval.

**b.**.
Asking for more confidence,the range should be expended to allow for a larger range of potential values for the population proportion to fall within. That is why the range of interval of 99% confidence level is wider than 95% confidence level.

**Answer for Task3:**

**a.**.

Null hypotheses H0: the proportion of population = 0.5.

Alternative hypotheses HA: the proportion of population $< 0.5$

**b.**.

```
tr <- t.test(sample,alternative='less',mu=0.5,conf.level = 0.95)
print(tr)
```

```
##
##  One Sample t-test
##
## data:  sample
## t = -3.1921, df = 42, p-value = 0.001337
## alternative hypothesis: true mean is less than 0.5
## 95 percent confidence interval:
##       -Inf 0.3954802
## sample estimates:
## mean of x
## 0.2790698
```

The P value is 0.001337461 at 95% confidence level. which is less than 0.05.So we ***reject*** the Null hypotheses H0.

One Sample t-test.

data: sample
t = -3.1921, df = 42, p-value = 0.001337.
alternative hypothesis: the proportion of population is less than 0.5.
95 percent confidence interval:
-Inf 0.3954802.
sample estimates:
mean of x
0.2790698

**c.**.

One Sample t-test. what we do is an One sample t-test.

data: sample.

t = -3.1921, df = 42, p-value = 0.001337.

T value is -3.1921, degree of freedom is 42, which is n-1,n=43.
p-value = 0.001337 which is $< 0.05$. We reject the H0 that the proportion of population = 0.5.
alternative hypothesis: the proportion of population is less than 0.5. this is a one sample t test and we are not care about the $> 0.5$.

95 percent confidence interval:-Inf 0.3954802. We have 95% confident that the proportion of population lie in the interval from -infinit to 0.3954802.

sample estimates: mean of x 0.2790698, the sample mean is 0.2790698.

## Question 2

For this question, we will use the data in the Excel worksheet Treesize.CSV. A copy of this can be found on Canvas. This data was collected in the US state of Georgia to test if the tree species growth is superior in a warmer climate compared to a cooler one. The data comes from two regions – north and south. The northern region is elevated and hence hosts a much cooler climate compared to the southern region. 30 pine trees were randomly selected from each region. Sizes were determined by measuring the diameter at breast height (DBH) for each tree in the sample. The data contains the following variables:

**Task 1:**

Description DBH for trees in the northern region DBH for trees in the southern region Variable Type Numeric Numeric.

| Variable | Description | Variable Types |
|---|---|---|
| North | DBH for trees in the northern region | Numeric |
| South | DBH for trees in the southern region | Numeric |

a. Produce a descriptive summary of the two groups of data. Paste your output.[2 Marks].

b. Use box plots and histograms to examine the two groups of data graphically. Paste your output. [4 Marks].

c. Based on the descriptive summary and the plots, describe the patterns in the data.

**Task 2.**

What are the appropriate null and alternative hypotheses for comparing the two groups of data? Justify your choice. [5 Marks].

**Task3.**

a.Explain which t-test will be appropriate for this data and why? [4 Marks].

b.Perform this t-test to test the hypotheses you described in Task 2. Paste your output. [2 Marks].

c.What do you conclude at 1% level? Report your finding in the context of the question posed. [4 Marks].

## Answer for Question2:

```
library(tidyverse)


df2 <-  read.csv('data/Treesize.csv',header=TRUE,sep=',')

sampleNorth <- df2 %>% pull('North')

sampleSouth <- df2 %>% pull('South')
print(summary(sampleNorth))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.20   10.45   17.05   23.70   38.55   58.80
```

```
print(summary(sampleSouth))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.60   26.65   37.70   34.53   44.55   52.90
```

**Answer for Task1:**

**a.**.

For Northern region:

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.20 10.45 17.05 23.70 38.55 58.80.

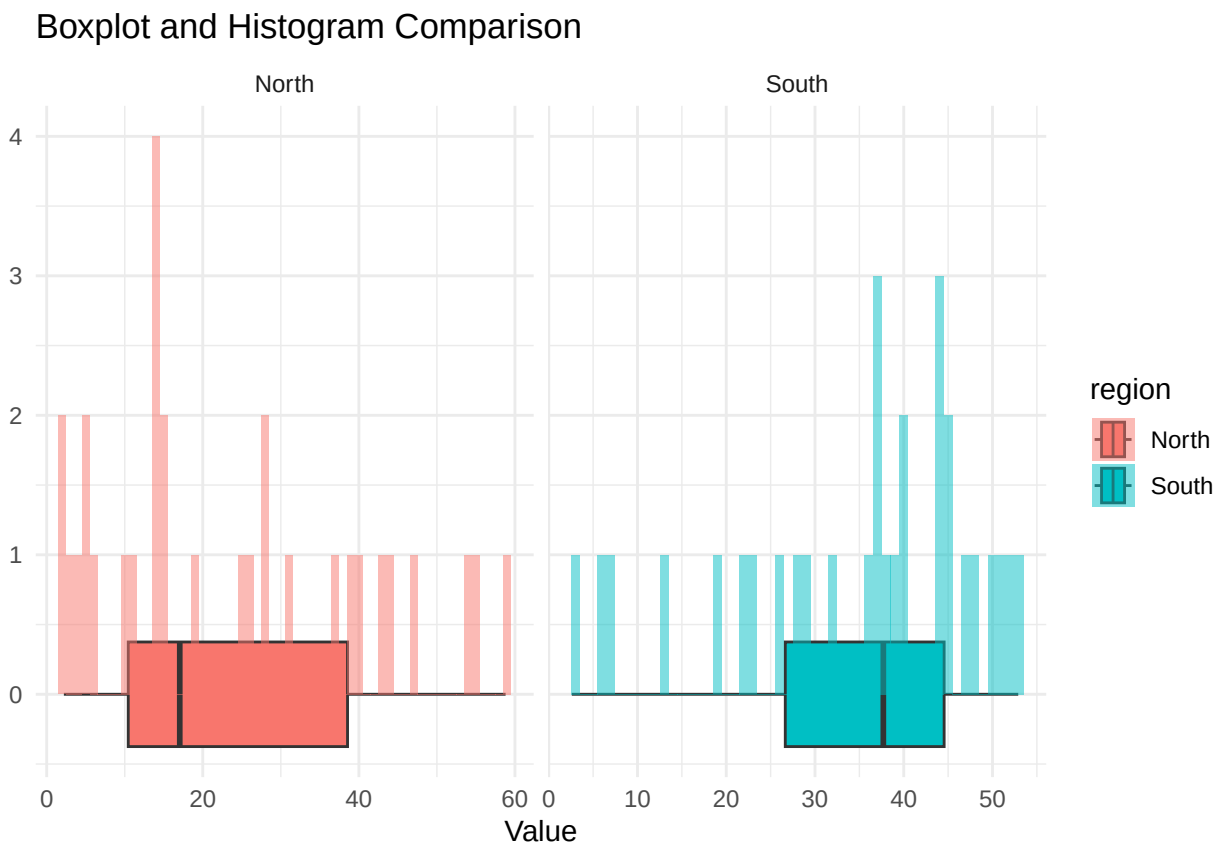
For Southern region:

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.60 26.65 37.70 34.53 44.55 52.90.

**b.**

```r
# Load required library
library(ggplot2)

# Combine data into a single data frame
combined_data <- data.frame(
  region = rep(c("North", "South"), each = length(sampleNorth)),
  value = c(sampleNorth, sampleSouth)
)

plot <- ggplot(combined_data, aes(x = value, fill = region)) +
  geom_boxplot() +
  geom_histogram(binwidth = 1, alpha = 0.5, position = "identity") +
  facet_grid(. ~ region, scales = "free") +
  labs(title = "Boxplot and Histogram Comparison",
       x = "Value", y = NULL) +
  theme_minimal()

# Display the plot
print(plot)
```




**c.**.

The data from North has the median 17.05,range from 2.20 to 58.8.It is Left Skewed.

The data from South has the median 37.70,range from 2.60 to 52.90. It is right skewed.

The Data from North has mean 23.70, while the data from South has mean 34.53.

The Data from North has one peak, and the Data from South has tow peak.

**Answer for Task2:**

Define

The Mean from the trees of South region u0. The Mean from the trees of North region u1;

Then

Null Hypothesis H0: u0 = u1.

Alternative Hypothesis H1:u0 > u1.

Because the samples are from two different population. And want to test if the growth of tree from south(u0) is superior in a warmer climate compared to a cooler one(u1).

**Answers for Task3:**

**a.** Two sample one side T test is appropriate for this case. Because the Sample data from ***Two Different Population*** and we only care about if the population of south is growth ***better*** than the North.

**b.**.
The t-test processes and gets result as following:

```
ts <- t.test(sampleSouth,sampleNorth,alternative = ("greater"),conf.level = 0.99)
print(ts)
```

```
##
##  Welch Two Sample t-test
##
## data:  sampleSouth and sampleNorth
## t = 2.6286, df = 55.725, p-value = 0.005529
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  0.9621957       Inf
## sample estimates:
## mean of x mean of y
##  34.53333  23.70000
```

**c.**.

according the test, p-value = 0.005529,is less than thesignificance level of 0.01 (associated with a 99% confidence level), so we will reject the null hypothesis.
That means the growth of tree from south(u0) is superior in a warmer climate compared to a cooler one(u1).

## Question 3

Beta-Blockers are a class of medications that block the effects of adrenaline on the heart, leading to a decrease in heart rate, thus lowering the blood pressure. An existing Beta- Blocker drug (Drug A) based on the molecule metoprolol is expected to reduce the diastolic blood pressure by 5mm Hg. A drug-maker has developed a new medication (Drug B) that is expected to be more effective and reduce the diastolic blood pressure by 6.5mm Hg. A clinical trial will need to be conducted to establish if Drug B is indeed more effective than Drug A. The FDA (Food and Drug Administration, USA) will only accept the findings if the level of significance is 0.05 (5%) and the power of the test is at least 0.8 (80%). You work as a statistician on this clinical trial. Based on past studies, you think that the s.d will range from between 3 and 10. Use the power.t.test command in R to determine the sample size needed to ensure at least 80% power at 5% level of significance for any s.d. within the given range.

Paste your command and output. Clearly justify your choice of the sample size.[15 Marks].
## Answer for Question 3

```
power.t.test(delta=1.5,sd=3,sig.level = 0.05,power=0.8,type=c("two.sample"),alternative = c("one.sided"
```

```
##
##      Two-sample t test power calculation
##
##               n = 50.1508
##           delta = 1.5
##              sd = 3
##       sig.level = 0.05
##           power = 0.8
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

```
power.t.test(delta=1.5,sd=10,sig.level = 0.05,power=0.8,type=c("two.sample"),alternative = c("one.sided
```

```
##
##      Two-sample t test power calculation
##
##               n = 550.2383
##           delta = 1.5
##              sd = 10
##       sig.level = 0.05
##           power = 0.8
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

```
power.t.test(n=551,delta=1.5,sd=10,sig.level = 0.05,type=c("two.sample"),alternative = c("one.sided"))
```

```
##
##      Two-sample t test power calculation
##
##               n = 551
##           delta = 1.5
##              sd = 10
##       sig.level = 0.05
##           power = 0.8004819
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

There are three things matter to the power value: the sample size, bigger better. the difference of the parameter value between H0 and HA. the standard deviation of the data, smaller better. In these case, the standard deviation is the only thing varied. to get 80% power at 5% level of significance,when Sd = 10, the sample size should larger then 551. n = 550.2383 delta = 1.5 sd = 10 sig.level = 0.05 power = 0.8 alternative = one.sided

So the sample size should larger than 551 for Each group.

## Question 4:

THC (tetrahydrocannabinol) is the active ingredient in a cannabis plant. Legal cannabis products in the United States are required to report THC potency (total THC% by dry weight) on packaging. However,

THC potency can change based on the strain of cannabis being used in a particular product. The three main strains used are sativa, indica and hybrid. To investigate if the average THC potency changed based on the dominant strain used in a product, 86 samples of products available in the state of Colorado were tested for their THC potency. This data is stored in the THCdata_strain.csv on Canvas. Use an appropriate hypothesis testing model studied in class to test if the mean THC content is the same for all three strains or, if not, which strain(s) are different.

Task 1:
Which hypothesis testing model is the most appropriate here? And why? [5 Marks].

Task 2:

Write your null and alternative hypotheses for the model you selected in Task 1. Clearly write what these mean in terms of the study question. [5 Marks].

Task 3: Implement this hypothesis testing using R. Paste your code and outputs (including plots, if any). Are the assumptions for this model met? Explain why (or why not)! [10 Marks].

Task 4:

Conclude (in statistical terms) what your results indicate. Also, explain this in a non- technical manner that is easily understood by a person not educated in statistical methods.[5 Marks].

## Answer for Question 4:

**Answer for task1:**

There are 3 populations which is more than 2.

so AVONA is the most appropriate hypothesis testing if the average of THC potency(That is the mean of THC.measured by group) is the same in this case.

**Answer for Task2:**

Let:

u0:the average of THC potency of Hybird which is the mean of THC Potency of Hybird.

u1:the average of THC potency of Indice which is the mean of THC Potency of Indice.

u2:the average of THC potency of Sativa which is the mean of THC Potency of Sativa.

Define:

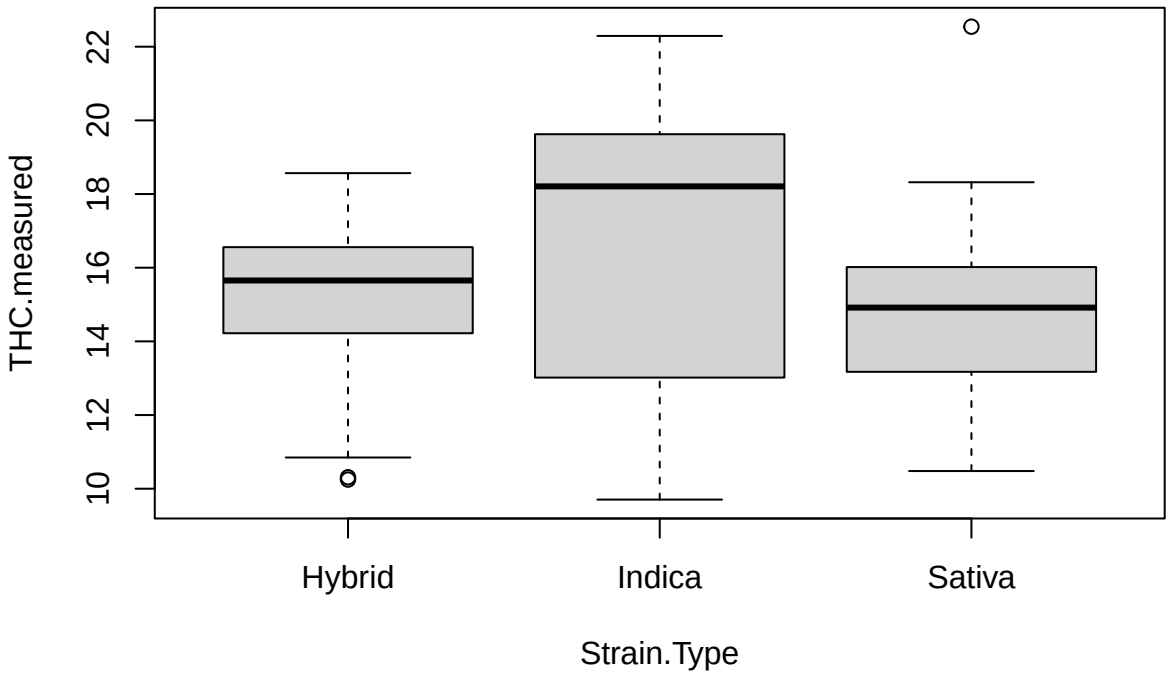
Null Hypothesis H0:

u0=u1=u2.

Alternative Hypothesis.
At lease one of the averages of THC potency is different.

**Answer for Task3:**

```
library(car)
df4 <- read.csv("data/THCdata_strain.csv",header = TRUE,sep=',')

bpo <- boxplot(THC.measured~Strain.Type,df4)
```




```
aggregate(THC.measured ~ Strain.Type, data = df4, FUN = length)
```

```
##   Strain.Type THC.measured
## 1      Hybrid           34
## 2      Indica           13
## 3      Sativa           39
```

From box plot,the data for Indice is right skewed, but other group looks symmetric.the range of Indice is larger than others. That means the variance of Indice will be bigger.

```
aggregate(THC.measured ~ Strain.Type, data = df4, FUN = sd)
```

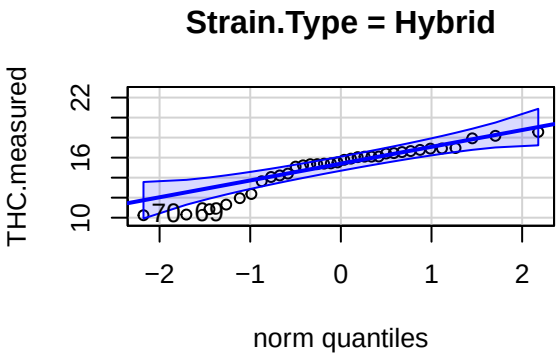
```
##   Strain.Type THC.measured
## 1      Hybrid     2.178725
## 2      Indica     3.960364
## 3      Sativa     2.292100
```

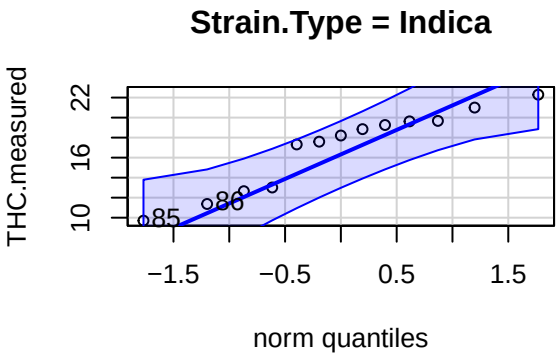
```
aggregate(THC.measured ~ Strain.Type, data = df4, FUN = mean)
```

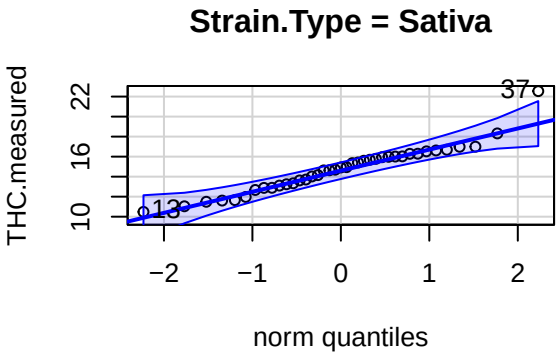
```
##   Strain.Type THC.measured
## 1      Hybrid     15.14949
## 2      Indica     16.96584
## 3      Sativa     14.73628
```

The standard deviation of Indice is almost the double of others.That is not satisfied the assumption that all populations have the same standard deviation.

```
qqPlot(THC.measured~Strain.Type,df4)
```



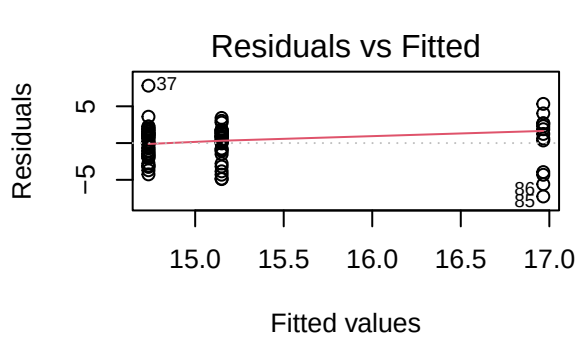



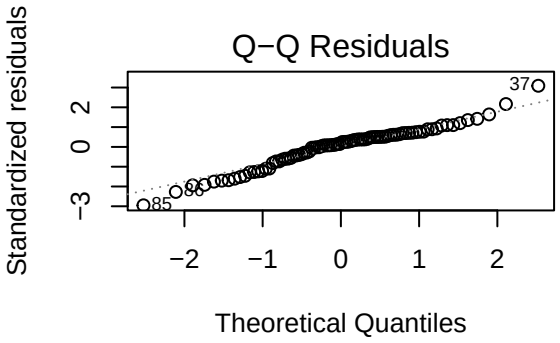

From qqPlot, the datas of Sativa and Indice looks Normal. Regarding Hybird, There are a few of data out of range, looks not good for ANOVA.It doesn't meet the assumption that all the populations are normally distributed.

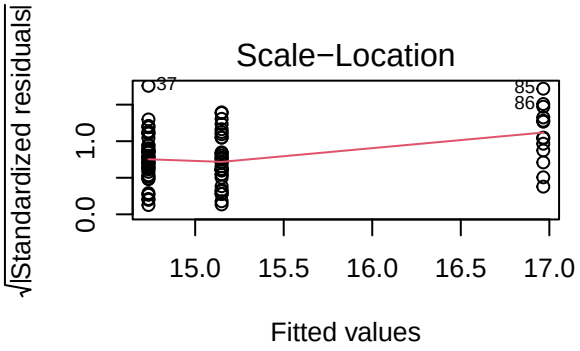
The original looks not met the assumption.

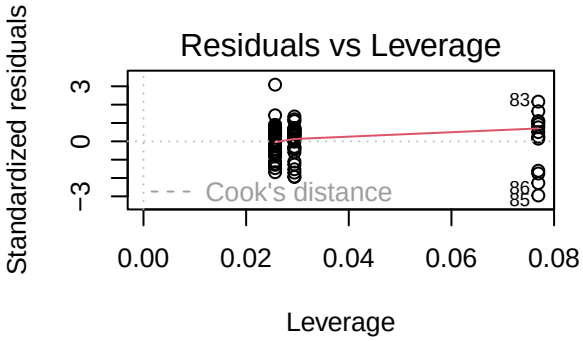
And if we run the one way ANOVA

```
anova0 = aov(THC.measured~Strain.Type,df4)
par(mfrow=c(2,2))
plot(anova0)
```



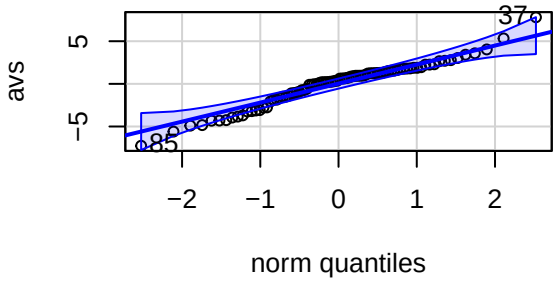

```
summary(anova0)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Strain.Type  2   48.9   24.45   3.727 0.0282 *
## Residuals   83  544.5    6.56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
avs <- residuals(anova0)
qqPlot(avs)
```

```
## [1] 37 85
```

```
# Perform Levene's test for homogeneity of variances
#levene_test <- leveneTest(avs ~ Strain.Type, data = df4)
#print(levene_test)
```



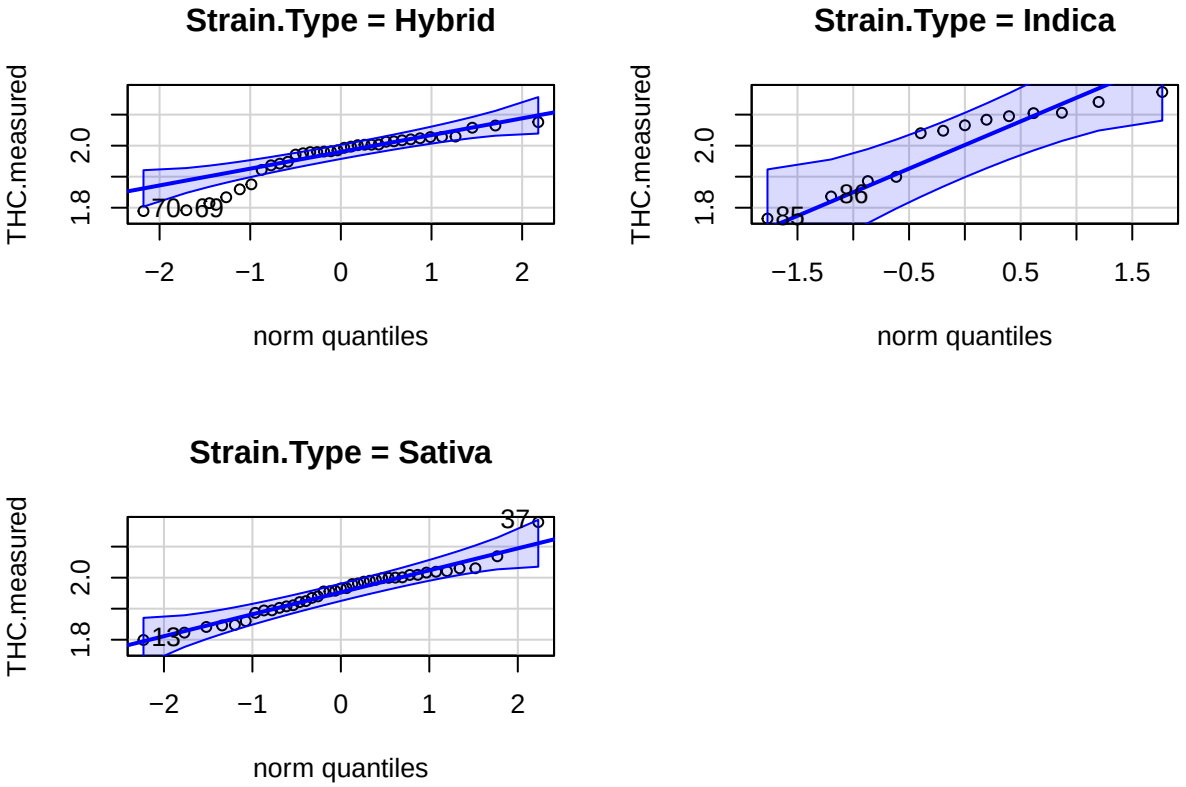

According the Diagnoise Plot:

Check the Q-Q plot for the normality, all point close to the line,the normality assumption is considered to be met.

Check residuals vs fitted and Scale-Location plots, the spread of the residuals of Indica obviously wider compared to the others.

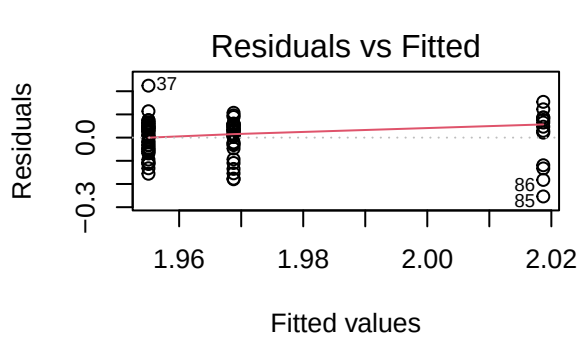
Try transform data,have another try.

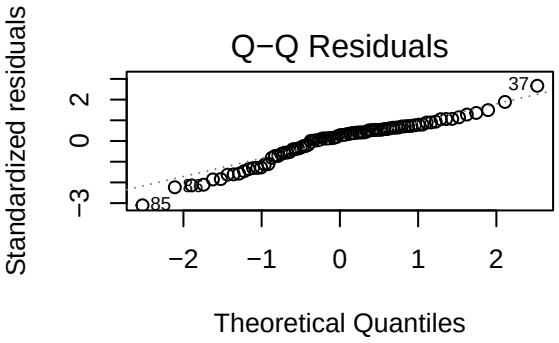
```
#remove outliers works perfect.
#outlier_indices <- which((df4$THC.measured %in% bpo$out)&(df4$Strain.Type == 'Sativa' ))

#df4 <- df4[-outlier_indices, ]

#bpo2 <- boxplot(THC.measured~Strain.Type,df4)
#outlier_indices2 <- which(df4$THC.measured %in% bpo2$out)
#df4 <- df4[-outlier_indices2, ]

#bpo3 <- boxplot(THC.measured~Strain.Type,df4)
#outlier_indices3 <- which(df4$THC.measured %in% bpo3$out)
#df4 <- df4[-outlier_indices3, ]
#boxplot(THC.measured~Strain.Type,df4)
#count_by_group <- aggregate(THC.measured ~ Strain.Type, data = df4, FUN = length)

#print(count_by_group)
df4[,'THC.measured'] <- sqrt(sqrt(df4[,'THC.measured']))
qqPlot(THC.measured~Strain.Type,df4)
```

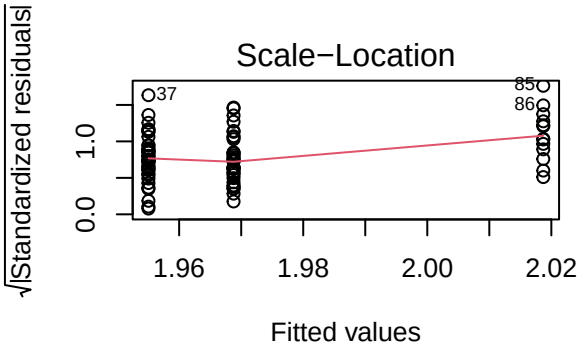


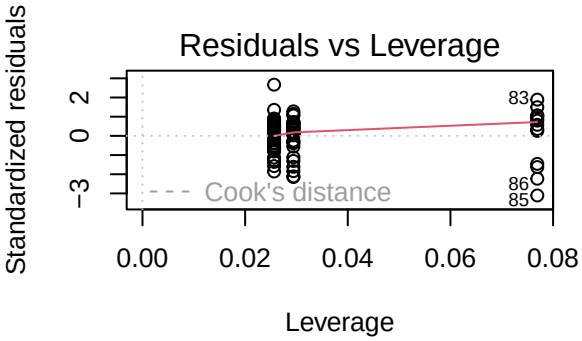

Run One way ANOVA.

```
anova1 = aov(THC.measured~Strain.Type,df4)
par(mfrow=c(2,2))
plot(anova1)
```



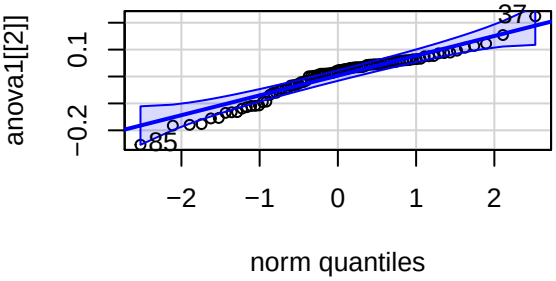

```
summary(anova1)
```

```
##             Df Sum Sq  Mean Sq F value Pr(>F)
## Strain.Type  2 0.0396 0.019799   2.739 0.0705 .
## Residuals   83 0.6000 0.007228
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#aus <- residuals(anova0)
qqPlot(anova1[[2]])
```

```
## [1] 85 37
```

```
# Perform Levene's test for homogeneity of variances
#levene_test <- leveneTest(aus ~ Strain.Type, data = df4)
#print(levene_test)
```




According the residual plots, the residual spread of the three dataset are similiar,much better than the ANOVA test from original data.

***So the assumptions are considered to be met after transforming data***

**Answer for Task 4:**

```
summary(anova1)
```

```
##             Df Sum Sq  Mean Sq F value Pr(>F)
## Strain.Type  2 0.0396 0.019799   2.739 0.0705 .
## Residuals   83 0.6000 0.007228
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The P-value=0.0705 $>$ 0.05, So we will not reject the H0(u1=u2=u3),there is no significant envidence to prove the average THC potency changed based on the dominant strain used in a product. The mean THC content is the same for all three strains sativa, indica and hybrid.