

# Assignment 2-Group11

Name: Jiamian Liu VU student number: 2632301

Name: Xiaoyu Yang VU student number: 2640948

Name: Fangzheng Lyu VU student number: 2644757

## Exercise 1

1. Because the given  $\lambda$  from an interval  $[0.01, 0.1]$ , we need to test each possible value from this interval. Therefore, except for the standard bootstrap test code, we need to apply a loop from 0.01 to 0.1 with the step of 0.0001 to calculate the p-value for each value.

The code is shown as Fig.1. The final output “p” means the biggest p-value of the test and the “lamadaflag” means the corresponding value of  $\lambda$ .

```
> bill <- scan("telephone.txt", skip=1)
Read 200 items
> B=1000
> t=median(bill)
> tstar=numeric(B)
> n=length(bill)
> lamada=0.01
> p=0
> lamadaflag=0
>
> for(count in 1:900){
+   lamada=lamada+0.0001;
+   for (i in 1:B){
+     xstar=rexp(n, lamada)
+     tstar[i]=median(xstar)
+   }
+   p1=sum(tstar<t)/B
+   pr=sum(tstar>t)/B
+   ptemp=2*min(p1, pr)
+   if(ptemp>p){
+     p=ptemp
+     lamadaflag=lamada
+   }
+ }
> p
[1] 0.996
> lamadaflag
[1] 0.0258
```

**Figure1: Bootstrap test**

We could know that when  $\lambda$  is 0.0258, the p-value reaches the max value of 0.996, which is not less than 5%. So we will not reject the  $H_0$ , thus the data in telephone.txt stems from the exponential distribution  $\text{Exp}(\lambda)$  with some  $\lambda$  from  $[0.01; 0.1]$

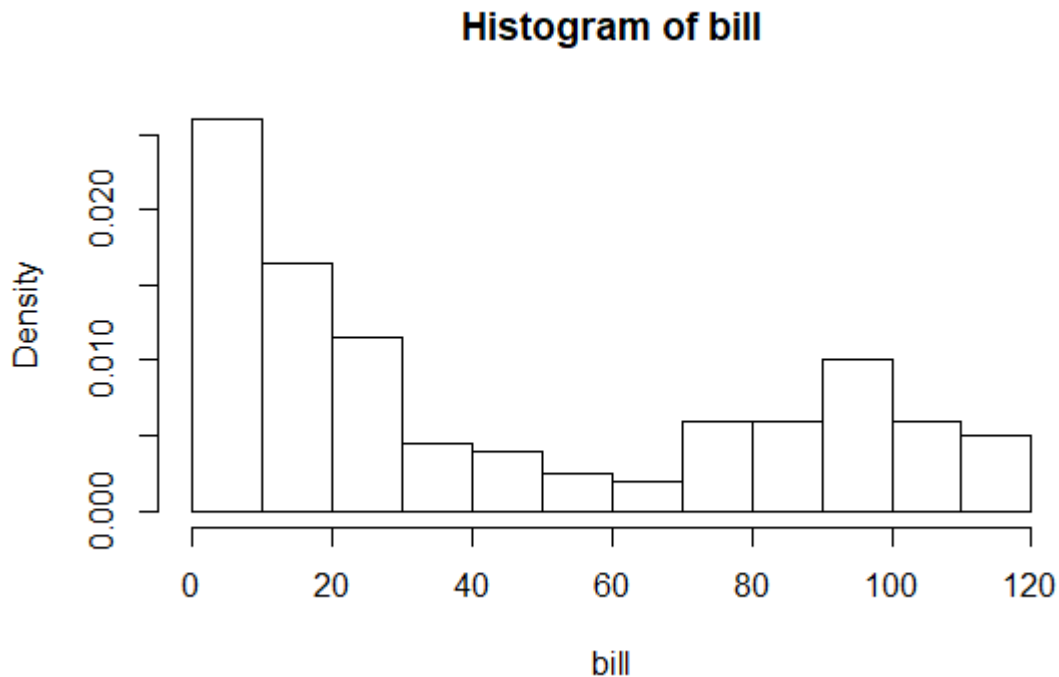
2. We plot the histogram of bill to see the distribution of the original data set. The code is shown in Fig.2 and the Histogram is shown in Fig.3.

```

> bill <- scan("telephone.txt",skip=1)
Read 200 items
> B=1000
> hist(bill,prob=T)

```

**Figure2: Histogram code**



**Figure3: Histogram of bill**

From Fig.3 we could know there is many bills around 0 so we need to check it. We use the summary function to check the statistical results as shown in Fig.4.

```

> summary(bill)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.000   9.385   26.905   43.588   84.828  119.630

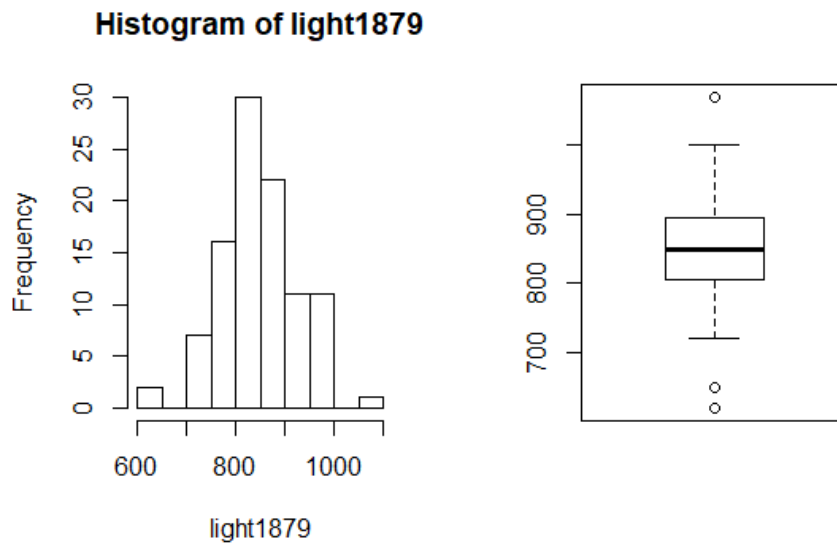
```

**Figure4: Statistical results**

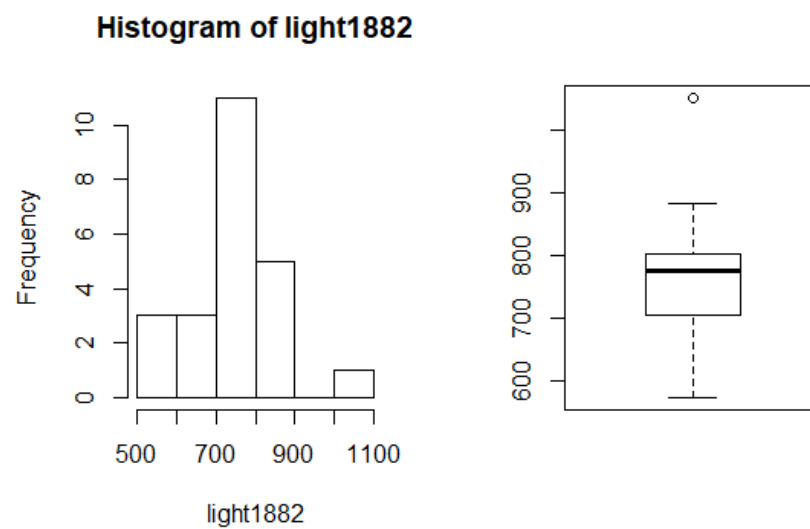
It shows the min value is 0 so the company need to check if all the survey is useful. From the histogram we could know that there may be two main group of customers, one is lower than 30 and the other is around 100 (larger than 70). The company should reconsider the different bills' plan for different groups for better market competitiveness.

## Exercise 2

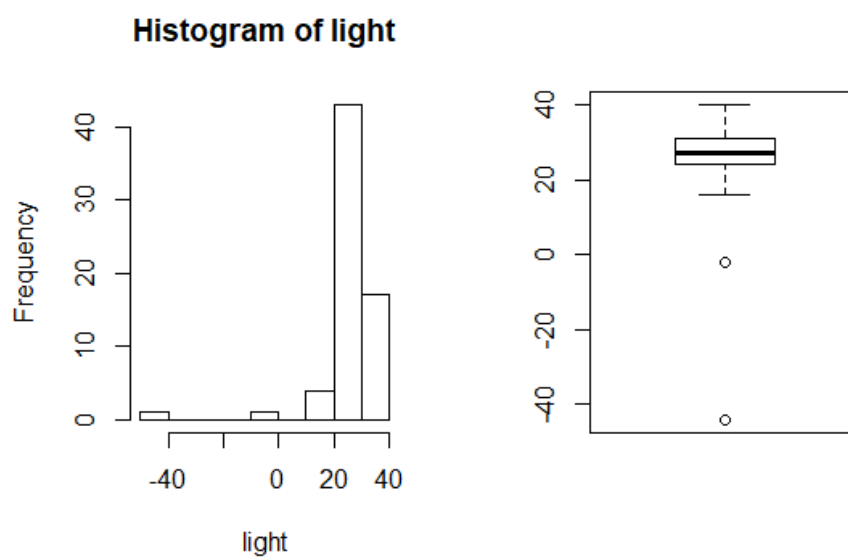
1. The histograms and box plots of each dataset is shown in Fig.5, Fig.6 and Fig.7.



**Figure5: Histogram and boxplot of “light1879.txt”**



**Figure6: Histogram and boxplot of “light1882.txt”**



**Figure7: Histogram and boxplot of “light.txt”**

From the histograms and box plots we could know that there is always a peak value of frequency in the given data set and there always are some too high/low values (outliers) in the given dataset. In Michelson's data the peak value is around 800 and the Newcomb's dataset is around 20, which all of them show the measurement of velocity of light.

2. We first pre-preprocess the given data to make the data represent the speed of light as first 5 lines. Then we use the processed data to calculate the confidence interval with mean and median separately. The code is shown as follows.

```
light1879 <- scan("light1879.txt")
light1882 <- scan("light1882.txt")
light <- scan("light.txt")
value1=light1879+299000
value2=light1882+299000
value3=744200000/(light+24800)

M1=mean(value1)
T1=median(value1)
M2=mean(value2)
T2=median(value2)
M3=mean(value3)
T3=median(value3)

B=1000
Tstar11=numeric(B)
Tstar12=numeric(B)
Tstar21=numeric(B)
Tstar22=numeric(B)
Tstar31=numeric(B)
Tstar32=numeric(B)
for(i in 1:B)
{
  Xstar1=sample(value1,replace=TRUE)
  Xstar2=sample(value2,replace=TRUE)
  Xstar3=sample(value3,replace=TRUE)
  Tstar11[i]=mean(Xstar1)
  Tstar12[i]=median(Xstar1)
  Tstar21[i]=mean(Xstar2)
  Tstar22[i]=median(Xstar2)
  Tstar31[i]=mean(Xstar3)
  Tstar32[i]=median(Xstar3)
}
Tstar1125=quantile(Tstar11,0.025)
Tstar11975=quantile(Tstar11,0.975)
Tstar1225=quantile(Tstar12,0.025)
Tstar12975=quantile(Tstar12,0.975)
Tstar2125=quantile(Tstar21,0.025)
Tstar21975=quantile(Tstar21,0.975)
Tstar2225=quantile(Tstar22,0.025)
Tstar22975=quantile(Tstar22,0.975)
Tstar3125=quantile(Tstar31,0.025)
Tstar31975=quantile(Tstar31,0.975)
Tstar3225=quantile(Tstar32,0.025)
Tstar32975=quantile(Tstar32,0.975)

c11=c(2*M1-Tstar11975,2*M1-Tstar1125)
c12=c(2*T1-Tstar12975,2*T1-Tstar1225)
c21=c(2*M2-Tstar21975,2*M2-Tstar2125)
c22=c(2*T2-Tstar22975,2*T2-Tstar2225)
c31=c(2*M3-Tstar31975,2*M3-Tstar3125)
c32=c(2*T3-Tstar32975,2*T3-Tstar3225)
```

**Figure8: Code for three data set**

The confidence interval results are shown in Fig.9.

```

> c11
  97.5%      2.5%
299836.5 299866.8
> c12
  97.5%      2.5%
299830 299860
> c21
  97.5%      2.5%
299710.7 299796.0
> c22
  97.5%      2.5%
299752 299825
> c31
  97.5%      2.5%
299733.2 299790.4
> c32
  97.5%      2.5%
299742.2 299772.4

```

**Figure9: Confidence interval results**

We could know for Michelson in 1879, when mean is used, the confidence interval is [299836.5, 299866.8] as c11. When median is used, the confidence interval is [299830, 299860] as c12.

We could know for Michelson in 1882, when mean is used, the confidence interval is [299710.7, 299796.0] as c21. When median is used, the confidence interval is [299752, 299825] as c22.

We could know for Newcomb, when mean is used, the confidence interval is [299733.2, 299790.4] as c31. When median is used, the confidence interval is [299742.2, 299772.4] as c32.

3. From the results we could know that we get shorter confidence intervals when using popular medians other than popular means, which means it tends to be more accurate confidence intervals when the median is applied. Because the mean is highly influenced by the high/low values (outliers) while the median is more robust and stable in such cases.
4. From the internet, the speed of light is 299792458m / s, which is 299792.458km/s. It is only consistent with the measurements of Michelson in 1882 among the 3 dataset. Because the actual value exists in the confidence intervals of the dataset.

## Exercise 3

1. Make the following assumptions based on the question:

$H_0$ : the median duration  $\mu$  is smaller or equal to 31 days

$H_1$ : median is greater than 31 days

Because of the small sample size, we are not sure about normality and whether it is from a symmetric population with a certain median. So we choose to use sign test to check whether we should reject  $H_0$  or not. The code and results are shown below as figure 10:

```
> binom.test(sum(klm<=31), length(klm), alternative = 'less', p = 0.5)

Exact binomial test

data:  sum(klm <= 31) and length(klm)
number of successes = 20, number of trials = 60, p-value = 0.006745
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.4464656
sample estimates:
probability of success
      0.3333333

> |
```

**Figure 10: Code and results of KLM median test**

Conclusion: As the p-value is  $0.006745 < 0.05$ . So we reject  $H_0$ , which means that the median duration  $\mu$  is greater than 31 days.

2. Make the following assumptions based on the question:

$H_0$ : the criterium is met.

$H_1$ : the criterium is not met.

```
> binom.test(sum(klm>72),length(klm),p=0.1,alternative = "greater")

Exact binomial test

data:  sum(klm > 72) and length(klm)
number of successes = 13, number of trials = 60, p-value = 0.005681
alternative hypothesis: true probability of success is greater than 0.1
95 percent confidence interval:
 0.1331878 1.0000000
sample estimates:
probability of success
      0.2166667
```

**Figure 11: Code and results of criterium sign test**

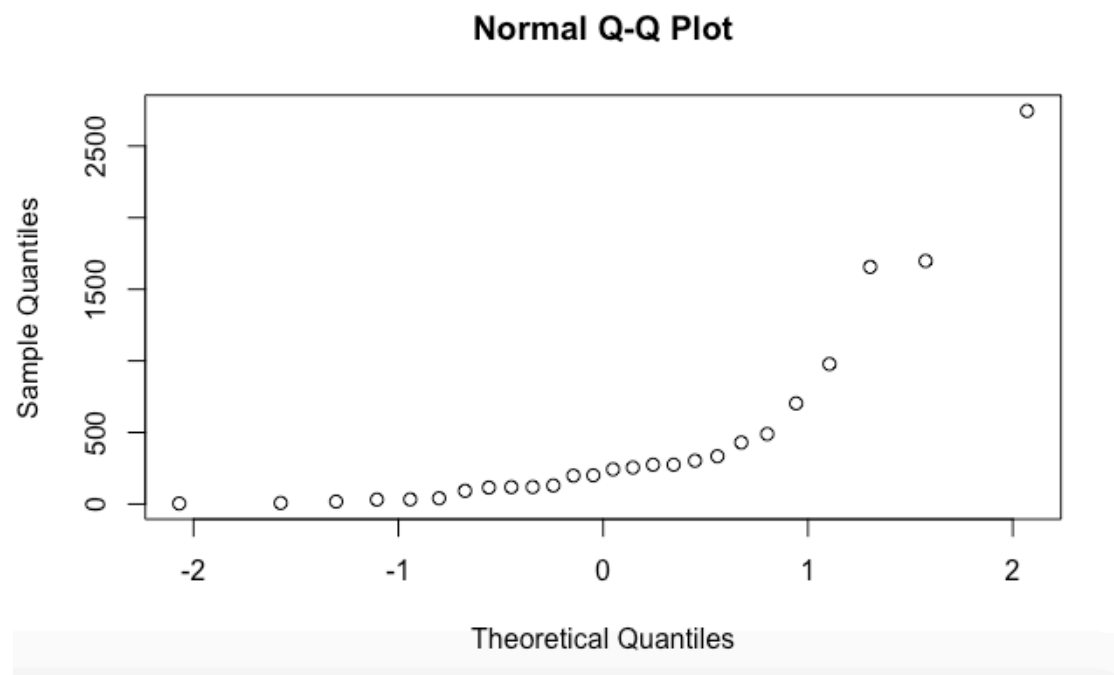
As the p-value of this sign test is  $0.005681 < 0.05$ . So we reject  $H_0$ , which means the criterium is not met.

## Exercise 4

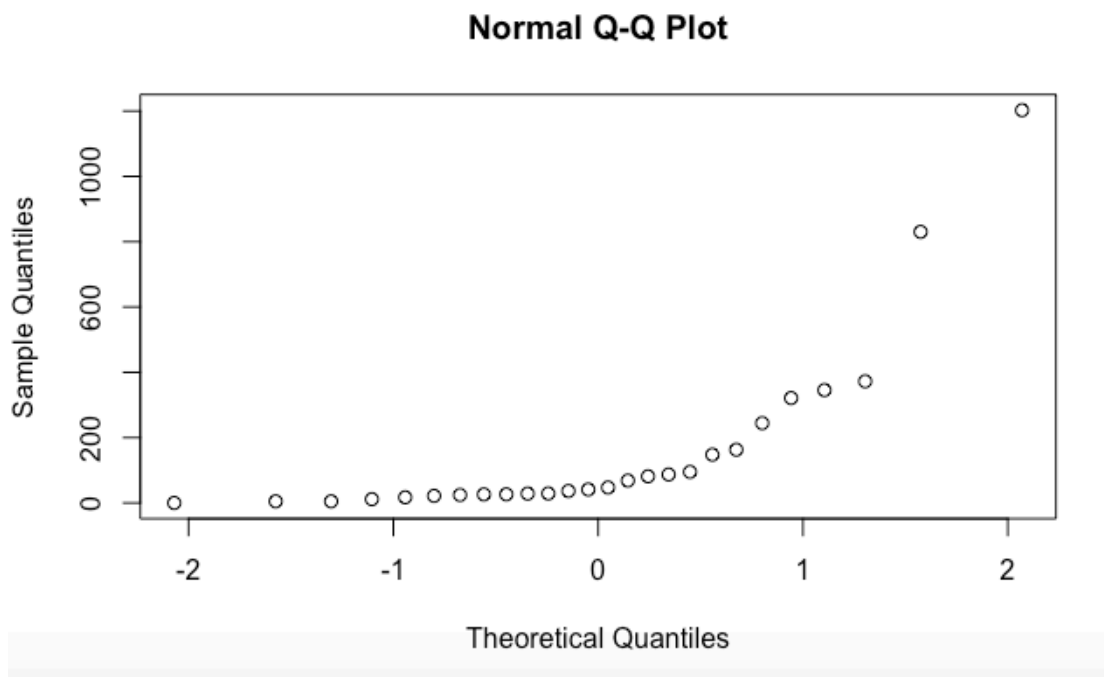
1. According to the question, this experiment has been carried out on different clouds and each cloud generate one numerical outcome and there are two groups of experimental units: the seeded and unseeded ones, so the data are not paired and we choose two sample t-test. We can make the following assumption:

$H_0$ : silver nitrate do not have effect on rain fall

$H_1$ : silver nitrate has an effect on rain fall



**Figure 12: Q-Q plot on clouds[,1]**



**Figure 13: Q-Q plot on clouds[,2]**

#### 1.1 two samples t-test

```
> t.test(clouds[,1],clouds[,2])
```

Welch Two Sample t-test

```
data: clouds[, 1] and clouds[, 2]
t = 1.9984, df = 33.856, p-value = 0.05375
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.740491 559.585876
sample estimates:
mean of x mean of y
441.9846 164.5619
```

**Figure 14: two samples t-test for clouds**

From the Q-Q plot we can see that the data “clouds” is not from a normal distribution. So the two sample t-test is not applicable here.

#### 1.2 Mann-Whitney test

```
> wilcox.test(clouds[,1],clouds[,2])
```

Wilcoxon rank sum test with continuity correction

```
data: clouds[, 1] and clouds[, 2]
W = 473, p-value = 0.01383
alternative hypothesis: true location shift is not equal to 0
```

**Figure 15: Mann-Whitney test for clouds**



The null hypothesis is that the rain fall data of seeded and unseeded are identical populations. To test the hypothesis, we apply the `wilcox.test` function to compare the two independent samples. As the p-value turns out to be 0.01383, and is less than the 0.05 significance level, we reject the null hypothesis. And T is large, which indicate that `clouds[,1]` is shifted towards the right from `clouds[,2]`, i.e. that seeded rain fall values are bigger than unseeded rain fall values.

### 1.3 Kolmogorov-Smirnov test

```
> ks.test(clouds[,1],clouds[,2])

Two-sample Kolmogorov-Smirnov test

data:  clouds[, 1] and clouds[, 2]
D = 0.42308, p-value = 0.01905
alternative hypothesis: two-sided
```

**Figure 16: Kolmogorov-Smirnov test for clouds**

As  $p\text{-value} = 0.01905 < 0.05$ ,  $H_0$  of equal means is rejected. The mean of `clouds[,1]` is larger, which indicates that the seeded clouds have more rain fall than the unseeded ones and silver nitrate has an positive influence on rain falls.

## 2. First we process the data to the square root values.

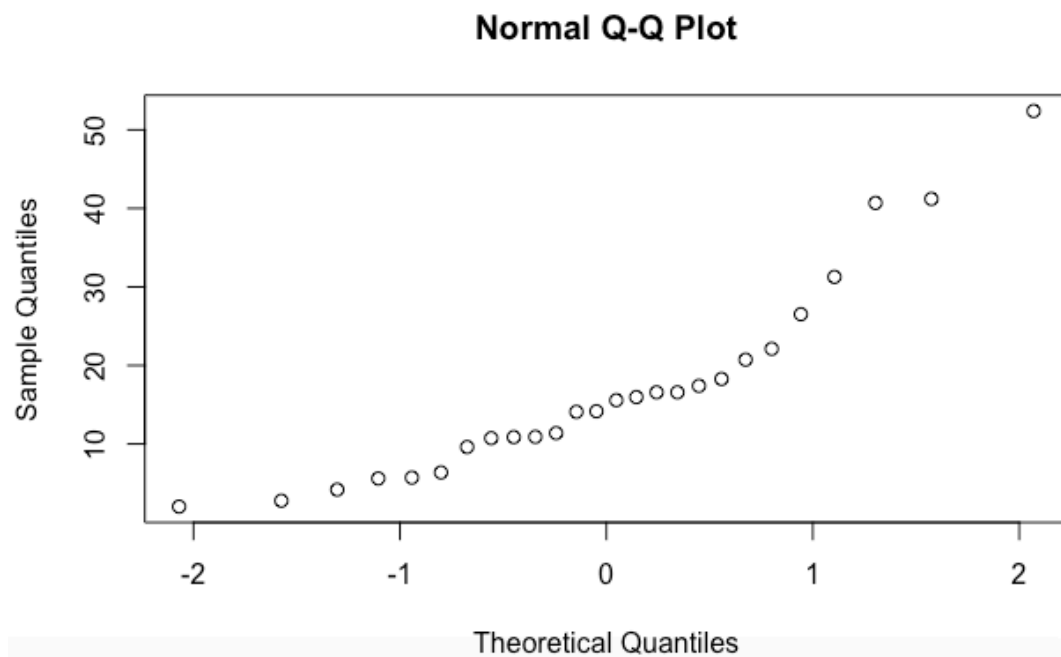
### 2.1 two samples t-test

```
> clouds2=sqrt(clouds)
> clouds2[,1]
 [1] 11.384200  5.603570 52.398473 22.115605 20.736441 17.401149
 [7] 10.908712  2.024846  9.612492  4.183300 14.166863 16.574076
[13] 16.574076  2.774887 40.693980 31.272992 14.092551 26.521689
[19] 41.204369 18.278403 10.876580 15.968719 10.737784 15.572412
[25]  5.718391  6.371813
> clouds2[,2]
 [1]  5.108816  5.128353  9.327379  9.746794 19.297668  0.100000
 [7]  4.159327  4.939636  3.391165 17.922053  8.276473  9.027735
[13]  6.877500  5.347897 28.811456 18.587630 34.678524  6.049793
[19]  2.213594  2.213594  6.410928  5.385165 12.767145 15.630099
[25] 12.157302  4.658326
> t.test(clouds2[,1],clouds2[,2])

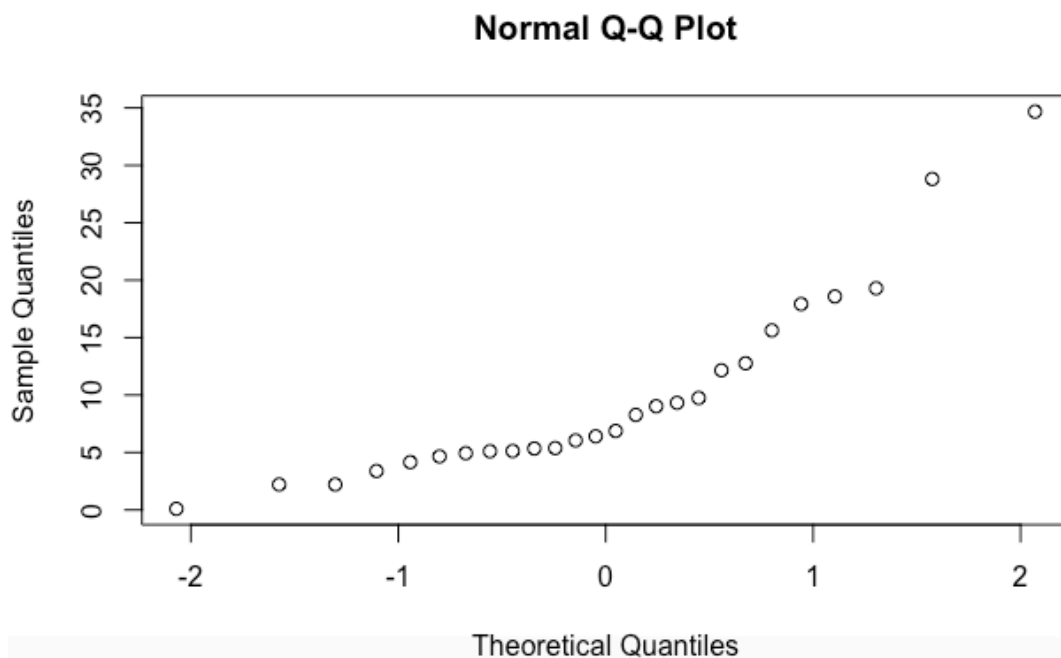
Welch Two Sample t-test

data:  clouds2[, 1] and clouds2[, 2]
t = 2.4246, df = 43.363, p-value = 0.01956
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.202087 13.071300
sample estimates:
mean of x mean of y
17.068014  9.931321
```

**Figure 17: Two sample t-test for  $\sqrt{\text{clouds}}$**



**Figure 18: Q-Q plot on clouds2[,1]**



**Figure 19: Q-Q plot on clouds2[,2]**

Same with the question (1.1). The data “clouds2” which the square root of data “clouds” is also not normal so it is not applicable for the question.

## 2.2 Mann- Whitney test

```
> wilcox.test(clouds2[,1],clouds2[,2])

Wilcoxon rank sum test with continuity correction

data:  clouds2[, 1] and clouds2[, 2]
W = 473, p-value = 0.01383
alternative hypothesis: true location shift is not equal to 0
```

### Figure 20: Mann- Whitney test for $\sqrt{\text{clouds}}$

We can see that the p-value here is exactly same as the result of Mann-Whitney test for clouds, because MW test works by bringing the data of two independent samples into a single space, in which data is globally ranked. Because the distribution of samples after square root has not changed, so the p-value is same as the former result.

## 2.3 Kolmogorov-Smirnov test

```
> ks.test(clouds2[,1],clouds2[,2])

Two-sample Kolmogorov-Smirnov test

data:  clouds2[, 1] and clouds2[, 2]
D = 0.42308, p-value = 0.01905
alternative hypothesis: two-sided
```

### Figure 21: Kolmogorov-Smirnov test for $\sqrt{\text{clouds}}$

We can see that the p-value here is exactly same as the result of Kolmogorov-Smirnov test on the sample “clouds”. Because this test the distribution of samples after square root has not changed, so the p-value is same as the former result.

3. First we process the data to square root of the square root values.

### 3.1 two samples t-test

```

> clouds3 = sqrt(clouds2)
> clouds3[,1]
[1] 3.374048 2.367186 7.238679 4.702723 4.553728 4.171468 3.302834
[8] 1.422971 3.100402 2.045312 3.763889 4.071127 4.071127 1.665799
[15] 6.379183 5.592226 3.754005 5.149921 6.419063 4.275325 3.297966
[22] 3.996088 3.276856 3.946189 2.391316 2.524245
> clouds3[,2]
[1] 2.2602690 2.2645866 3.0540758 3.1219856 4.3929111 0.3162278
[7] 2.0394428 2.2225291 1.8415116 4.2334446 2.8768859 3.0046190
[13] 2.6224987 2.3125520 5.3676304 4.3113374 5.888474 2.4596328
[19] 1.4878153 1.4878153 2.5319811 2.3205958 3.5731142 3.9534920
[25] 3.4867323 2.1583155
> t.test(clouds3[,1],clouds3[,2])

Welch Two Sample t-test

data: clouds3[, 1] and clouds3[, 2]
t = 2.5968, df = 48.826, p-value = 0.0124
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2196477 1.7236468
sample estimates:
mean of x mean of y
 3.878988  2.907340

```

Figure 22: Two sample t-test for  $\sqrt{\sqrt{clouds}}$

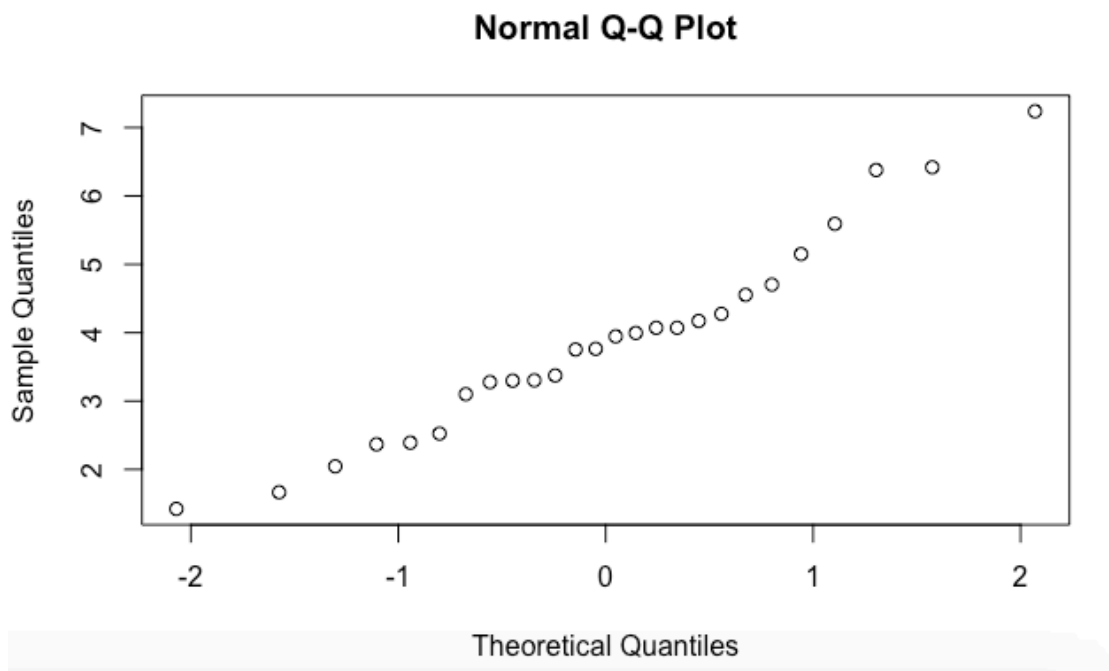
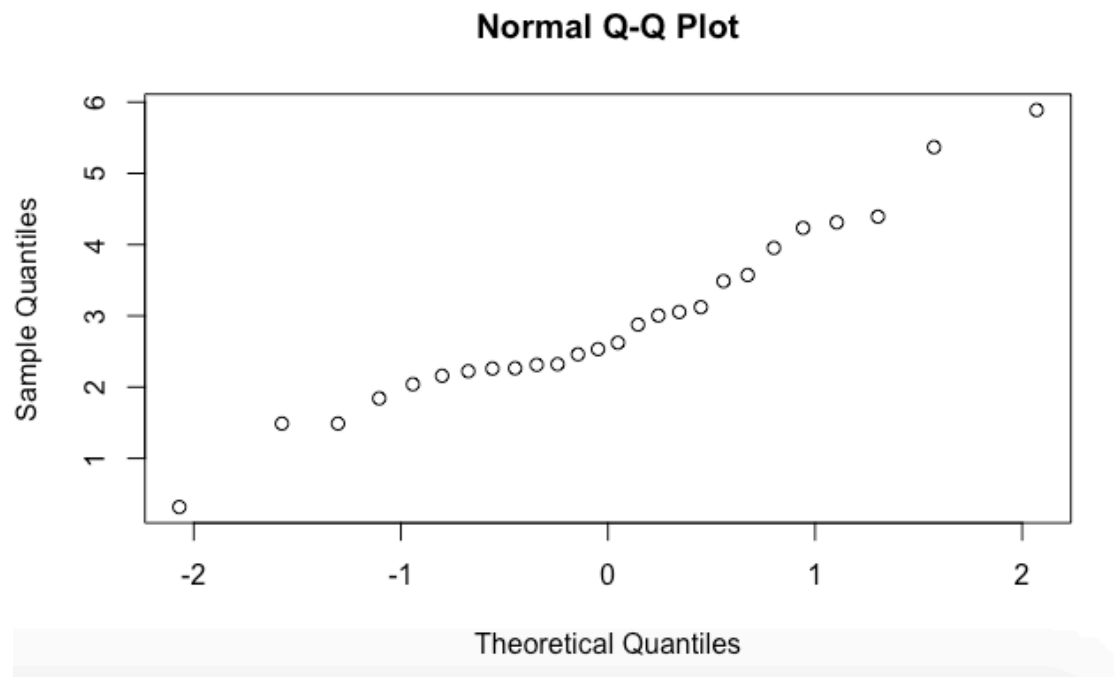


Figure 23: Q-Q plot on clouds3[,1]



**Figure 24: Q-Q plot on clouds3[,2]**

After square root twice, the data “clouds3” have followed a normal distribution. As p-value is  $0.0124 < 0.05$  and less than the former ones,  $H_0$  is rejected. Which means that silver nitrate has an effect on rain falls. And from the mean of seeded and unseeded clouds we can see that the seeded clouds have more rain fall than the unseeded ones.

### 3.2 Mann-Whitney test

```
> wilcox.test(clouds3[,1],clouds3[,2])

Wilcoxon rank sum test with continuity correction

data: clouds3[, 1] and clouds3[, 2]
W = 473, p-value = 0.01383
alternative hypothesis: true location shift is not equal to 0
```

**Figure 25: Mann-Whitney test for  $\sqrt{\sqrt{clouds}}$**

The reason of p-value did not change is as same as 2.2 had shown.

### 3.3 Kolmogorov-Smirnov test

```
> ks.test(clouds3[,1],clouds3[,2])
```

Two-sample Kolmogorov-Smirnov test

data: clouds3[, 1] and clouds3[, 2]

D = 0.42308, p-value = 0.01905

alternative hypothesis: two-sided

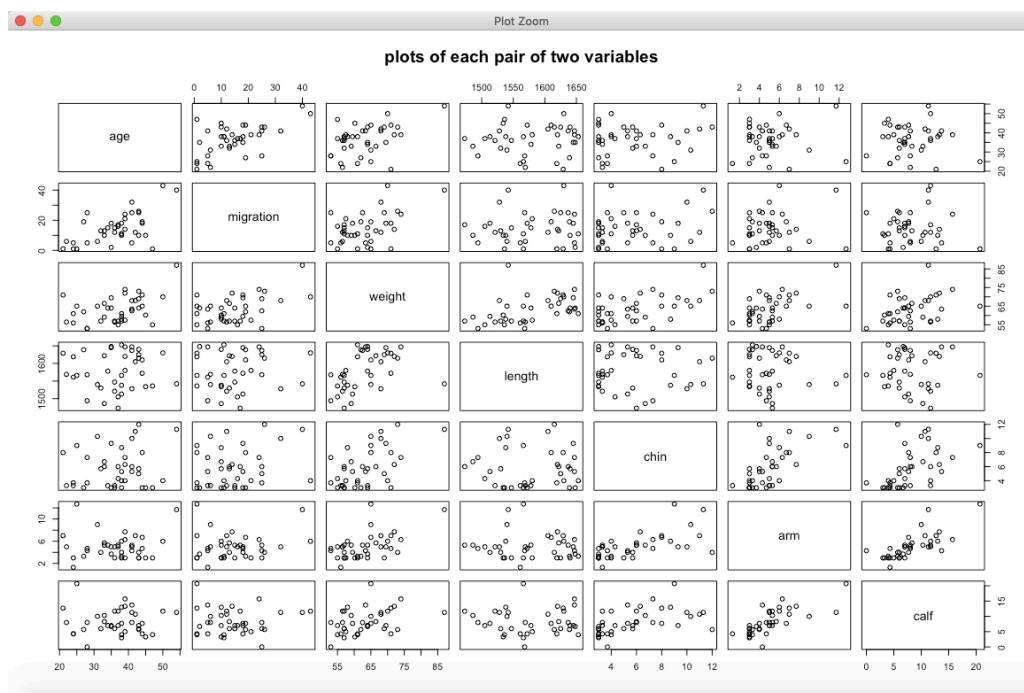
**Figure 26: Kolmogorov-Smirnov test for  $\sqrt{\sqrt{\text{clouds}}}$**

The reason of p-value did not change is as same as 2.3 had shown.

Conclusion: After getting the square root twice, the data sets are now follow the normal distribution, so all the three test method are applicable for the question.

## Exercise 5

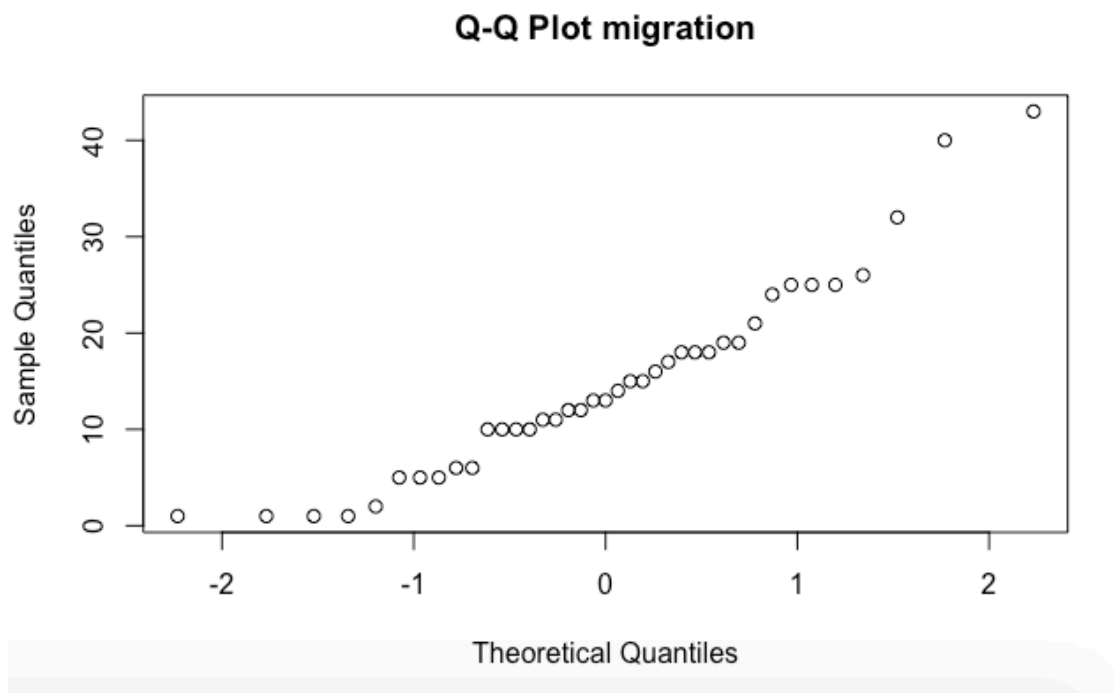
- Using the code `pairs(peruvians[1:7])` we get plots of each pair of two variables as below:



**Figure 27: plots of each pair of two variables**

- QQ-plots show that normality is not plausible for the migration sample. Hence, we choose to use the rank correlation test of Spearman to test the correlation between

migration and other variables .



**Figure 28: Q-Q plot migration**

2.1 The rank correlation test between migration and age

```
> cor.test(migration,age,method="spearman")
```

Spearman's rank correlation rho

```
data: migration and age
S = 5176.6, p-value = 0.002189
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.4760575
```

**Figure 29: rank correlation test of Spearman between migration and age**

We can see that p-value is  $0.002189 < 0.05$ . So there is correlation between migration and age, if normality is assumed.

2.2 The rank correlation test between migration and weight

```
> cor.test(migration,weight,method="spearman")
```

Spearman's rank correlation rho

```
data: migration and weight
S = 6415.1, p-value = 0.02861
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.3506956
```

### Figure 30: rank correlation test of Spearman between migration and weight

We can see that p-value is  $0.02861 < 0.05$ . So there is correlation between migration and weight. But the correlation is not too big. If normality is assumed.

#### 2.3 The rank correlation test between migration and length

```
> cor.test(migration,length,method="spearman")
```

Spearman's rank correlation rho

```
data: migration and length
S = 9044.3, p-value = 0.6087
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.08458432
```

### Figure 31: rank correlation test of Spearman between migration and length

We can see that p-value is  $0.6807 > 0.05$ . So there is no correlation between migration and length, if normality is assumed.

#### 2.4 The rank correlation test between migration and wrist

```
> cor.test(migration,wrist,method="spearman")
```

Spearman's rank correlation rho

```
data: migration and wrist
S = 7712.8, p-value = 0.1797
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2193498
```

### Figure 32: rank correlation test of Spearman between migration and wrist

We can see that p-value is  $0.1797 > 0.05$ . So there is no correlation between migration and wrist, if normality is assumed.

#### 2.5 The rank correlation test between migration and systolic



```
> cor.test(migration,systolic,method="spearman")

Spearman's rank correlation rho

data: migration and systolic
S = 11544, p-value = 0.3054
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
-0.1684286
```

**Figure 33: rank correlation test of Spearman between migration and systolic**

We can see that p-value is  $0.3054 > 0.05$ . So there is no correlation between migration and systolic, if normality is assumed.

## 2.6 The rank correlation test between migration and diastolic

```
> cor.test(migration,diastolic,method="spearman")

Spearman's rank correlation rho

data: migration and diastolic
S = 9137.6, p-value = 0.6494
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.07514098
```

**Figure 34: rank correlation test of Spearman between migration and diastolic**

We can see that p-value is  $0.6494 > 0.05$ . So there is no correlation between migration and diastolic, if normality is assumed.

# EXERCISE 6

## 1. data and graphical

### (1) Study the data:

As shown on the Table 1, we compute the mean and standard deviation of running time (lemo and energy drink). Specifically, the “lemo\_difference” = “running time before (lemo)” – “running time after (lemo)”.

We can find two basic results:

- Energy drink group seems run faster after drinking the energy drink, because the time of “energy\_after” is shorter than the time of “energy\_before”. Similarly, softdrink group seems run slower after drinking the softdrink, because the time of “lemo\_after” is longer than the time of “lemo\_before”.

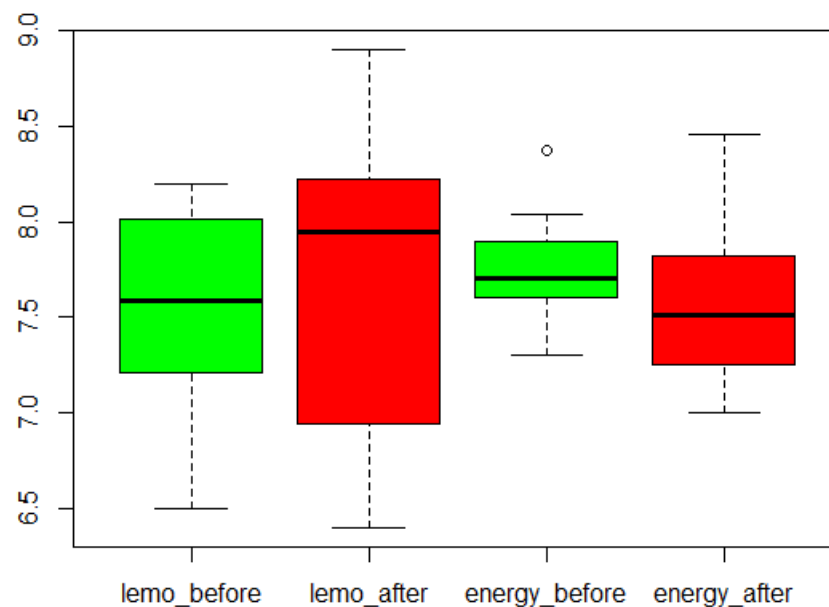
- b. From the standard deviation, we can find that energy drink group has lower standard deviation than softdrink group. Therefore, the energy drink group's data tend to be close to the mean of the set.

**Table 1 The mean and Standard Deviation of lemo and energy**

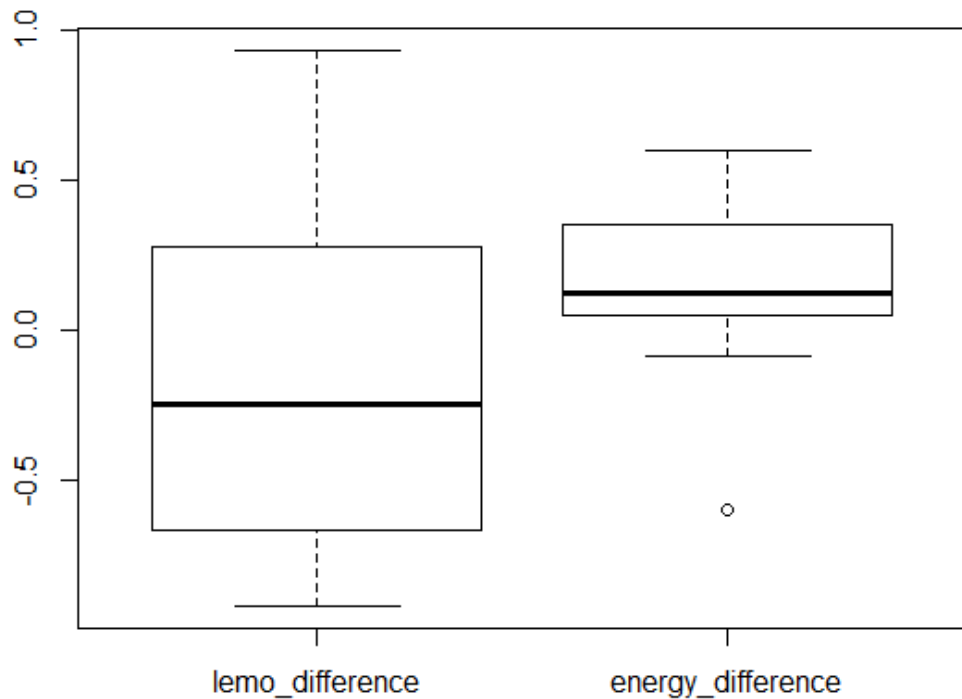
	<b>mean</b>	<b>standard deviation</b>
<b>lemo_before</b>	7.554	0.5471157
<b>lemo_after</b>	7.699	0.8339551
<b>lemo_difference</b>	-0.145	0.6232248
<b>energy_before</b>	7.733	0.2971417
<b>energy_after</b>	7.578	0.4585319
<b>energy_difference</b>	0.154	0.3229258

**(2) Graphical representations:**

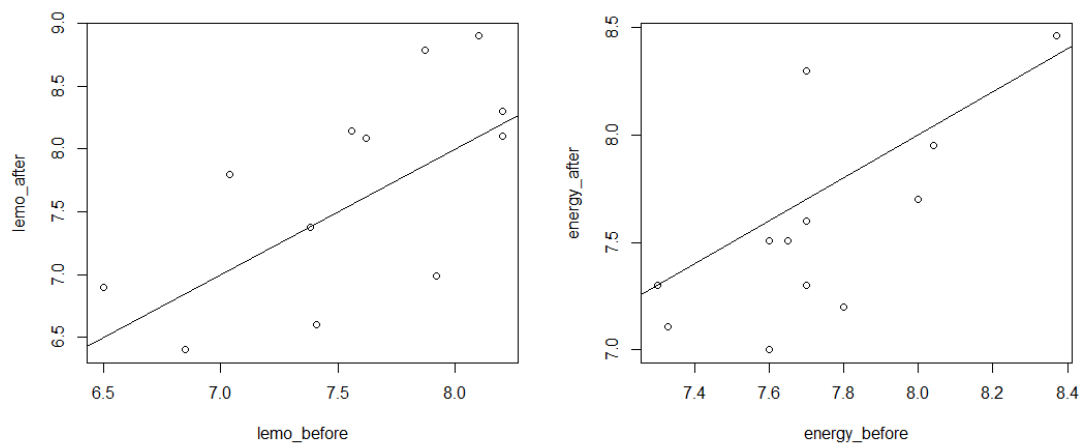
As shown on the figure35 and figure36, we can see the distribution of energy drink group is more concentrated.



**Figure35: boxplots of soft drink and energy drink (before and after)**



**Figure36: boxplots of soft drink and energy drink (difference)**



**Figure37 plot of before and after**

**code of (1):**

```
alldata = read.table("run.txt")

lemon_before = alldata[which(alldata$drink=="lemon"),"before"]
lemon_after = alldata[which(alldata$drink=="lemon"),"after"]
lemon_difference = lemon_before - lemon_after
energy_before = alldata[which(alldata$drink=="energy"),"before"]
energy_after = alldata[which(alldata$drink=="energy"),"after"]
energy_difference = energy_before - energy_after
mydiff = c(lemon_difference,energy_difference)
```

```

alldata$difference = mydiff

mean(lemo_before)
mean(lemo_after)
mean(lemo_difference)
mean(energy_before)
mean(energy_after)
mean(energy_difference)

sd(lemo_before)
sd(lemo_after)
sd(lemo_difference)
sd(energy_before)
sd(energy_after)
sd(energy_difference)

par(mfrow=c(1,2))
plot(lemo_before,lemo_after); abline(0,1)
plot(energy_before,energy_after); abline(0,1)

boxplot(lemo_before,lemo_after,energy_before,energy_after,
names=c("lemo_before","lemo_after","energy_before","energy_after"),col=c('green','
red',"green","red"))
boxplot(lemo_difference,energy_difference,names=c("lemo_difference","energy_diff
erence"))

```

## 2. Test separately

We can use paired t-test, because each group contains two numerical outcomes per experimental unit and our interest is in a possible difference between the two outcomes.

$H_0$  = there is not a significant difference in speed in the two running tasks.

For softdrink group, the p-value = 0.4373 > 0.05. Therefore, we cannot reject  $H_0$ , which means there is not a significant difference in speed in the two running tasks after drinking softdrink.

### result of paired t-test on softdrink:

```
> t.test(lemo_before,lemo_after,paired = TRUE)
```

Paired t-test

data: lemo\_before and lemo\_after

t = -0.80596, df = 11, p-value = 0.4373

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.5409781 0.2509781

sample estimates:

mean of the differences

-0.145

For energy drink group, the p-value = 0.1246 > 0.05. Therefore, we cannot reject  $H_0$ , which means there is not significant a difference in speed in the two running tasks after drinking energy drink.

#### **result of paired t-test on energy drink:**

```
> t.test(energy_before,energy_after, paired=TRUE)
```

Paired t-test

data: energy\_before and energy\_after

t = 1.6538, df = 11, p-value = 0.1264

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

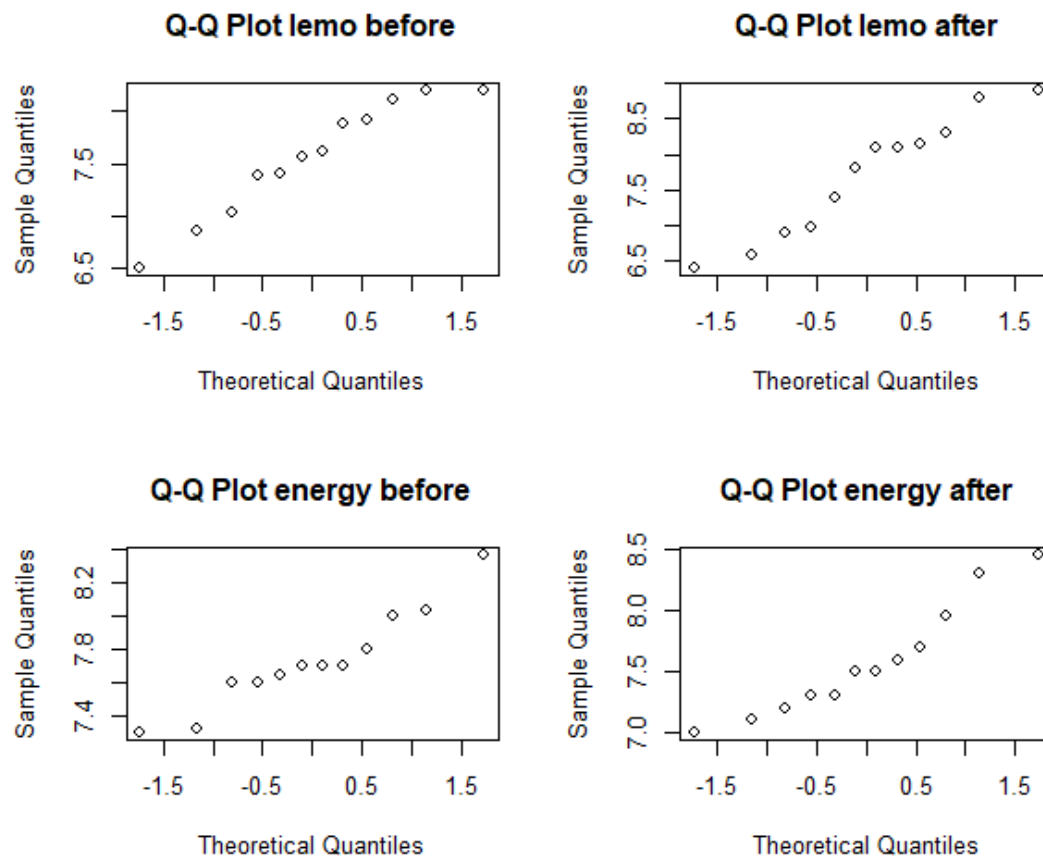
-0.05101059 0.35934392

sample estimates:

mean of the differences

0.1541667

Finally, we check the normality assumption on the differences.



**Figure 38, Q-Q plots of energy drink and softdrink group**

After Shapiro-Wilk normality test, there is no reason to suspect that the running time sample are not taken from a normal population.

**result of Shapiro-Wilk normality test on soft drink:**

```
> shapiro.test(lemo_before)

Shapiro-Wilk normality test

data:  lemo_before
W = 0.93899, p-value = 0.485

> shapiro.test(lemo_after)

Shapiro-Wilk normality test

data:  lemo_after
W = 0.94117, p-value = 0.5133

> shapiro.test(energy_before)
```

Shapiro-Wilk normality test

```
data: energy_before  
W = 0.93334, p-value = 0.4169
```

```
> shapiro.test(energy_after)
```

Shapiro-Wilk normality test

```
data: energy_after  
W = 0.92444, p-value = 0.3249
```

### code of (2)

```
t.test(lemo_before, lemo_after, paired = TRUE)  
t.test(energy_before, energy_after, paired = TRUE)  
  
par(mfrow = c(2, 2))  
qqnorm(lemo_before, main = "Q-Q Plot lemo before")  
qqnorm(lemo_after, main = "Q-Q Plot lemo after")  
qqnorm(energy_before, main = "Q-Q Plot energy before")  
qqnorm(energy_after, main = "Q-Q Plot energy after")  
  
shapiro.test(lemo_before)  
shapiro.test(lemo_after)  
shapiro.test(energy_before)  
shapiro.test(energy_after)
```

### 3. Test whether these time differences

As the table 2 shows, we can find the time difference of soft drink and energy drink.

**Table 2. The time difference of soft drink and energy drink**

time difference of soft drink	time difference of energy drink
0.93	0.22
-0.58	-0.60
-0.47	-0.09
0.45	0.10
-0.92	0.60
0.81	0.09
0.00	0.40
-0.76	0.60
-0.80	0.30
-0.40	0.14
0.10	0.09
-0.10	0.00

Because there are two groups of student, one drinking softdrink and another one drinking energy drink respectively, they are two independent samples. We use two samples t-test.

#### results of t-test of time difference:

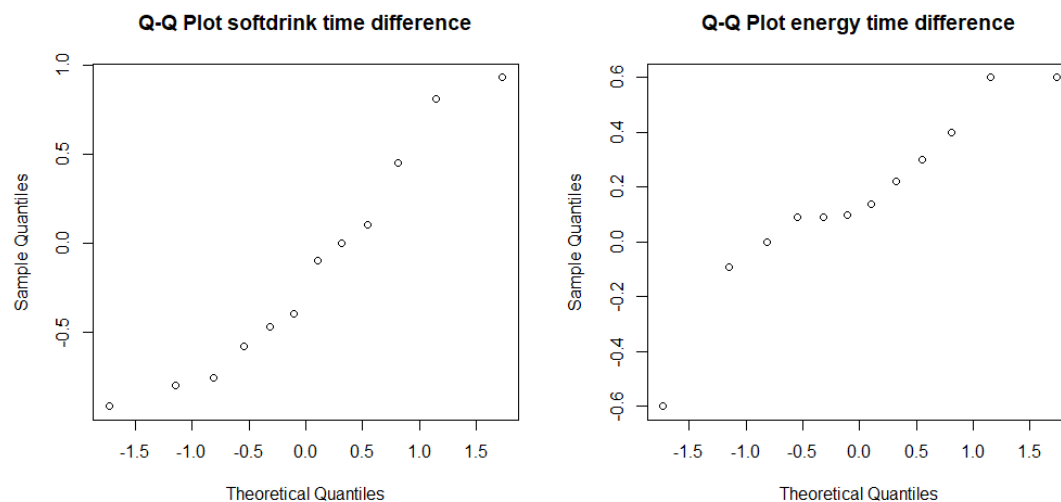
```
> t.test(lemo_difference,energy_difference)
```

Welch Two Sample t-test

```
data: lemo_difference and energy_difference
t = -1.4764, df = 16.509, p-value = 0.1586
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7276409  0.1293076
sample estimates:
 mean of x  mean of y
-0.1450000  0.1541667
```

The p-value=0.1586 > 0.05. Therefore, we cannot reject  $H_0$ , which means there is not a time difference between soft drink and energy drink.

Finally, we check the normality assumption. Figure 39 shows the Q-Q plot of soft drink and energy drink time difference.



**Figure39. Q-Q plot of soft drink and energy drink time difference**

After Shapiro-Wilk normality test, there is no reason to suspect that the differences are not taken from a normal population. The reason is that p-value of soft drink = 0.3723 > 0.05, and the p-value of energy drink = 0.2788 > 0.05.

#### result of Shapiro-Wilk normality test on soft drink:

```
> shapiro.test(lemo_before-lemo_after)
```

Shapiro-Wilk normality test



```
data: lemo_before - lemo_after  
W = 0.92927, p-value = 0.3725
```

#### result of Shapiro-Wilk normality test on energy drink:

```
> shapiro.test(energy_before-energy_after)  
  
Shapiro-Wilk normality test  
  
data: energy_before - energy_after  
W = 0.91913, p-value = 0.2788
```

Figure 33, Q-Q plots of energy drink and softdrink group

#### code of (3):

```
lemo_before = alldata[which(alldata$drink=="lemo"),"before"]  
lemo_after = alldata[which(alldata$drink=="lemo"),"after"]  
lemo_difference = lemo_before - lemo_after  
energy_before = alldata[which(alldata$drink=="energy"),"before"]  
energy_after = alldata[which(alldata$drink=="energy"),"after"]  
energy_difference = energy_before - energy_after  
  
t.test(lemo_difference,energy_difference)  
  
shapiro.test(lemo_difference-energy_difference)  
  
qqnorm(lemo_difference,main="Q-Q Plot softdrink time difference")  
qqnorm(energy_difference,main="Q-Q Plot energy time difference")  
shapiro.test(lemo_before-lemo_after)  
shapiro.test(energy_before-energy_after)
```

#### 4. plausible objection

The original design is testing on two different group of students, which is 12 pupils for each. However, we think different student may influence the results. Also, the 60 meters is too short to find the time difference.

We suggest the new design of the experiment:

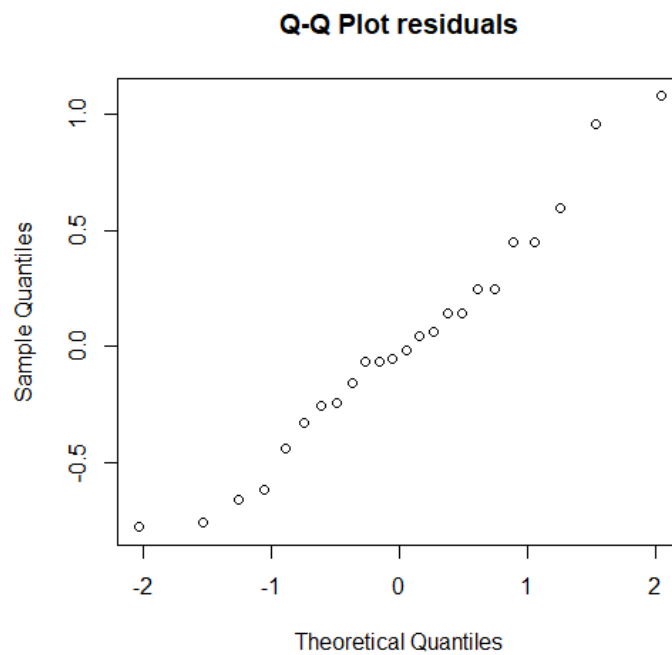
- (1) test on the same pupils, randomized and double-blind for drinking softdrink and energy drink.
- (2) running on 500 meters

#### 5. Yes

The t-test result of 3) also cannot reject  $H_0$ . The running distance is too short to compute a significant difference between energy drink and soft drink.

6. **normal distribution assumption** on these differences is needed for the analysis in 3).

The vector into 24 residuals to investigate this assumption in QQ-plot is shown on the figure 40.



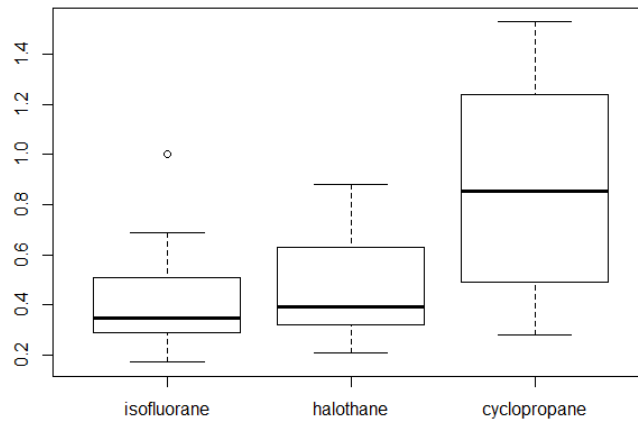
**Figure 40. Q-Q Plot residuals**

**code of (6):**

```
diff = data.frame(lemo_diff=lemo_difference,energy_diff=energy_difference)
diff_frame=
data.frame(yield=as.vector(as.matrix(diff)),variety=factor(rep(1:2,each=12)))
diff_frame[1:12,]
timeonaov= lm(yield~variety,data=diff_frame)
anova(timeonaov)
par(mfrow=c(1,1)); qqnorm(residuals(timeonaov),main="Q-Q Plot residuals")
```

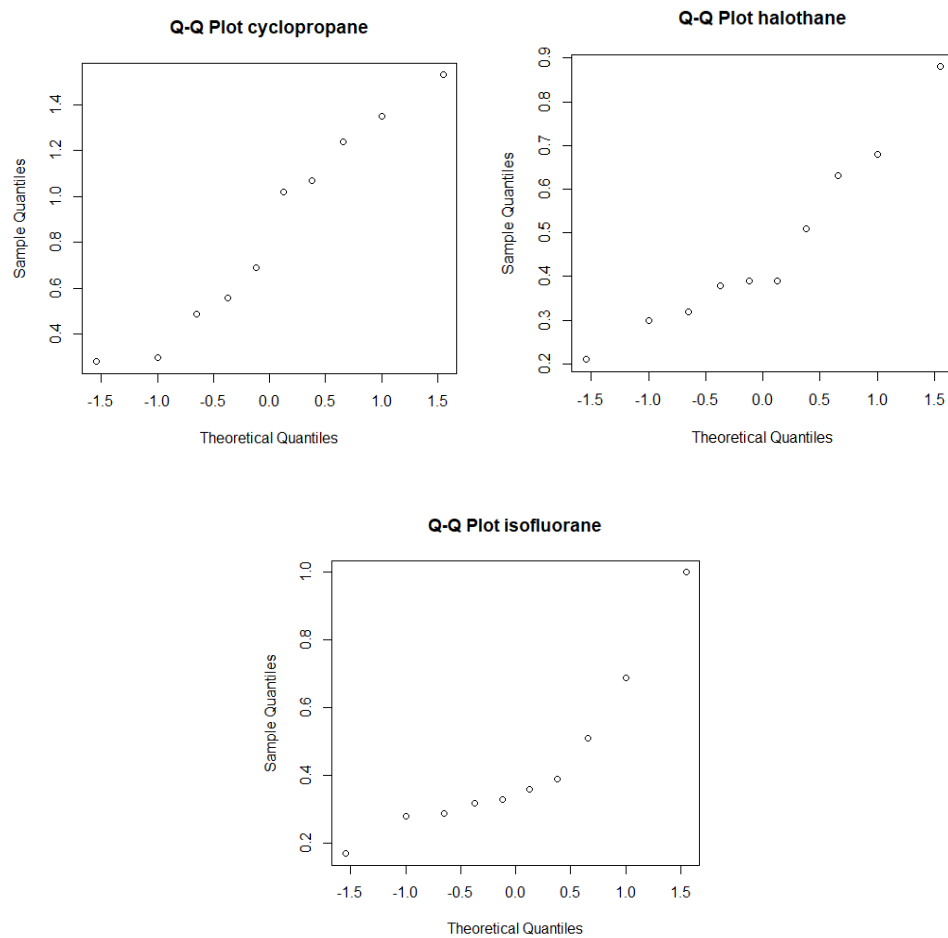
# EXERCISE 7

1. As the figure 41 shows, the boxplots of 3 samples.



**Figure 41. Boxplots of 3 samples**

As the figure shows, the Q-Q plots of 3 samples. There is no reason to suspect that the two samples are not taken from a normal population.



**Figure 42. Q-Q plots of 3 samples**

**code of (1):**

```
dogs = read.table("dogs.txt", header=TRUE)
isofl = dogs$isofluorane
haloe = dogs$halothane
cyclo = dogs$cyclopropane
boxplot(dogs)
qqnorm(isofl,main="Q-Q Plot isofluorane")
qqnorm(haloe,main="Q-Q Plot halothane")
qqnorm(cyclo,main="Q-Q Plot cyclopropane")
```

## 2. We use one-way ANOVA to test.

The p-value=0.011 < 0.05, so we reject H<sub>0</sub>. 1-way ANOVA yield significant differences among the three drugs.

**result of ANOVA:**

```
> anova(dogsonaov)
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value Pr(>F)
variety    2  1.0808  0.54040    5.355  0.011 *
Residuals 27  2.7247  0.10092
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We takes “isofluorane” as a base level and compares the other levels to it. The estimates:

$\mu^1 = 0.4340$ ;  $\mu^2 - \mu^1 = 0.0350$ ;  $\mu^3 - \mu^1 = 0.4190$ ;

Therefor, estimate of “isofluorane” is 0.4340; estimate of “halothane” is 0.4690; estimate of “cyclopropane” is 0.8530.

**result of summary:**

```
> summary(dogsonaov)

Call:
lm(formula = yield ~ variety, data = dogs_frame)

Residuals:
    Min       1Q   Median       3Q      Max
-0.5730 -0.1608 -0.0790  0.2000  0.6770

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4340     0.1005   4.320 0.000189 ***
variety2     0.0350     0.1421   0.246 0.807266
variety3     0.4190     0.1421   2.949 0.006504 **
---

```

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

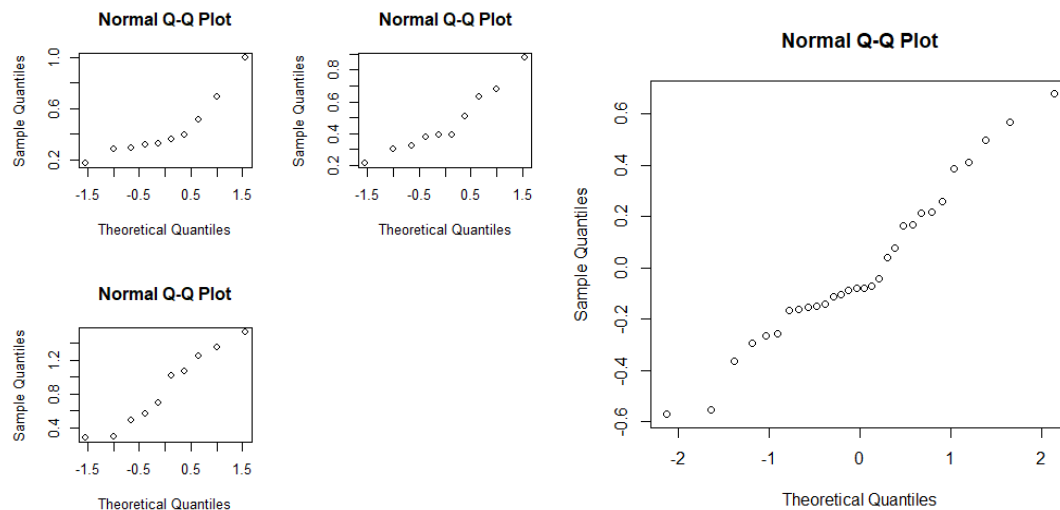
Residual standard error: 0.3177 on 27 degrees of freedom

Multiple R-squared: 0.284, Adjusted R-squared: 0.231

F-statistic: 5.355 on 2 and 27 DF, p-value: 0.011

### Diagnostics:

The Q-Q plot is shown on the figure 37.



**Figure 43. Q-Q plots of three drugs**

Because the 3 samples are small, separate QQ-plots are not so useful. The second plot, using residuals, uses all 30 points, but corrected for being sampled from different populations. The residuals don't seem to deviate significantly from normal.

### code of (2):

```
dogs_frame =  
data.frame(yield=as.vector(as.matrix(dogs)),variety=factor(rep(1:3,each=10)))  
dogsonaov=lm(yield~variety,data=dogs_frame)  
anova(dogsonaov)  
summary(dogsonaov)  
confint(dogsonaov)  
  
par(mfrow=c(2,2)); for (i in 1:3) qqnorm(dogs[,i])  
par(mfrow=c(1,1)); qqnorm(residuals(dogsonaov))
```

### 3. Kruskal-Wallis test

As the Kruskal-Wallis test shows, the p-value=0.05948>0.05, we cannot reject H0. There is not a significant difference among the three drugs.

```
> kruskal.test(worms,group)
```

Kruskal-Wallis rank sum test

data: worms and group

```
Kruskal-Wallis chi-squared = 5.6442, df = 2, p-value = 0.05948
```

The reason why there is a difference between the two test is that ANOVA is value-based, but Kruskal-Wallis is rank based. ANOVA can analyze normal distribution, but Kruskal-Wallis is suitable for non-normal distribution.

From Shapiro-Wilk normality test, "halothane" and "cyclopropane" samples are normal distribution. However, "isofluorane" is non-normal distribution.

```
> shapiro.test(isofl)

Shapiro-Wilk normality test

data:  isofl
W = 0.83093, p-value = 0.03434

> shapiro.test(haloe)

Shapiro-Wilk normality test

data:  haloe
W = 0.9234, p-value = 0.3862

> shapiro.test(cyclo)

Shapiro-Wilk normality test

data:  cyclo
W = 0.93334, p-value = 0.4815
```

**code of (3):**

```
dogframe=data.frame(worms=as.vector(as.matrix(dogs)),group=as.factor(rep(1:3,eac
h=10)))
attach(dogframe)
kruskal.test(worms,group)

shapiro.test(isofl)
shapiro.test(haloe)
shapiro.test(cyclo)
```