**Assignment 2-Group11**

**Name:** Jiamian Liu **VU student number:** 2632301

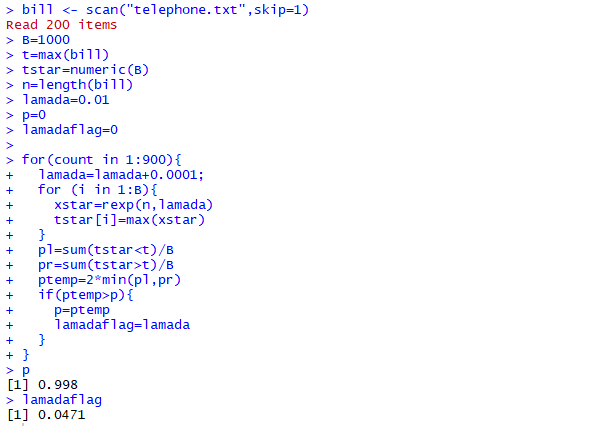
**Name:** Xiaoyu Yang **VU student number:** 2640948

**Name:** Fangzheng Lyu **VU student number:** 2644757

# Exercise 1

1. Because the given λ 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 λ.



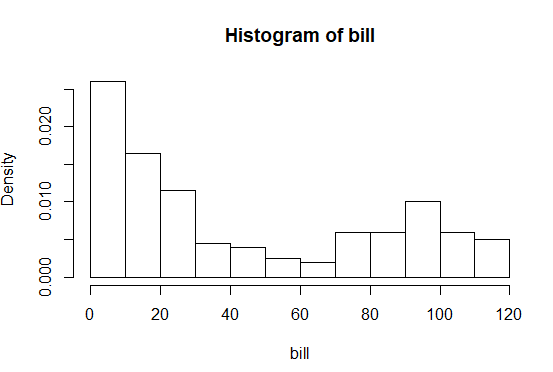
**Figure1: Bootstrap test**

We could know that when λ is 0.0471, the p-value reaches the max value of 0.998, which is not less than 5%. So we will not reject the H0, thus the data in telephone.txt stems from the exponential distribution Exp(λ) with some λ from [0:01; 0:1]

1. 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.



**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.



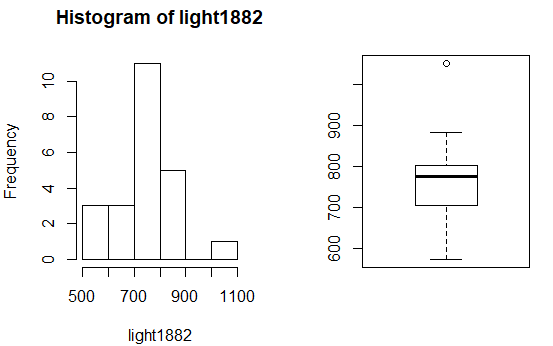
**Figure4: Statistical results**

It shows the min value is 0 so the company need to check if all the survey is useful.

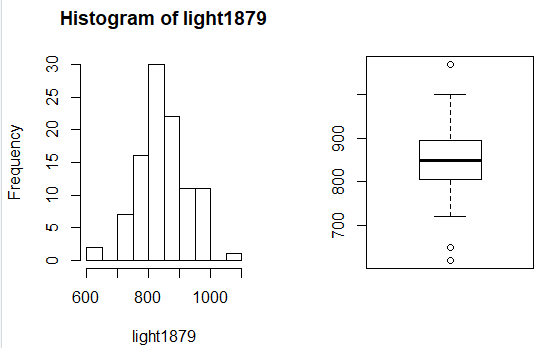
From the histogram we could know that there may be two main group of customers, one is lower than 30 and the other is around 100. The company should reconsider the different bills’ plan for different groups for better market competitiveness.

# Exercise 2

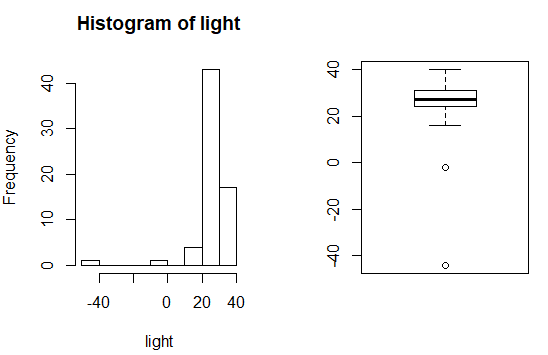
1. The histograms and box plots of each dataset is shown in Fig.5, Fig.6 and Fig.7.



**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. 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.

1. We calculate the confidence intervals via sign test. The codes and results are shown in Fig.8.



**Figure8: Code and results of three data set**

We could know for light.txt, the confidence interval is [0.3599108,0.6111897].

For light1879.txt, the confidence interval is [0.3503202,0.5527198].

For light 1882.txt, the confidence interval is [0.2681962,0.6941220].

1. From the confidence interval, we could know that p-value is large enough that all the H0 is not rejected. Besides, the probability of success is high. We could think that all the data sets recorded are accurate and could be of use.
2. From the internet, the speed of light is 299 792 458 m / s.

It is consistent with the measurements of Michelson and Newcomb. Because the actual value exists in the confidence intervals of the dataset of measurements.

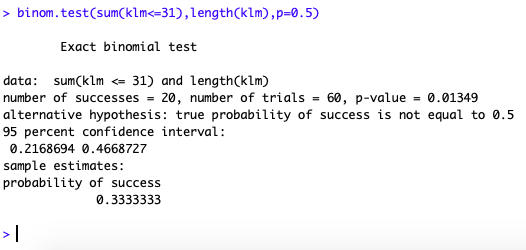
# Exercise 3

1. Make the following assumptions based on the question：

: the median duration μ is smaller or equal to 31 days

: 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 or not. The code and results are shown below as figure 9:



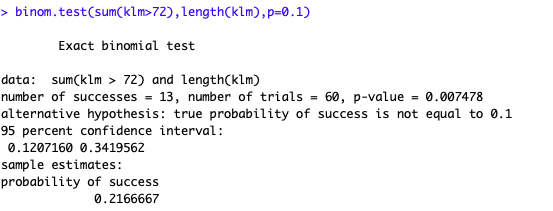
**Figure 9: Code and results of KLM median test**

Conclusion: As the p-value is 0.01349<0.05. So we reject , which means that the median duration μ is greater than 31 days.

1. Make the following assumptions based on the question：

: the criterium is met.

: the criterium is not met.



**Figure 10: Code and results of criterium sign test**

As the p-value of this sign test is 0.007478<0.05. So we reject , which means the criterium is not met.

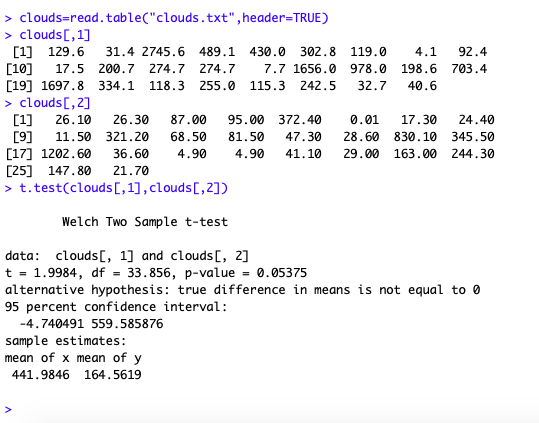
# Exercise 4

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:

: silver nitrate has an effect on rain fall

: silver nitrate do not have effect on rain fall

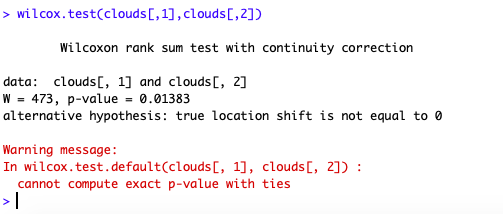
* 1. two samples t-test



**Figure 11: two samples t-test for clouds**

As p-value is 0.05375>0.05, of is not 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.

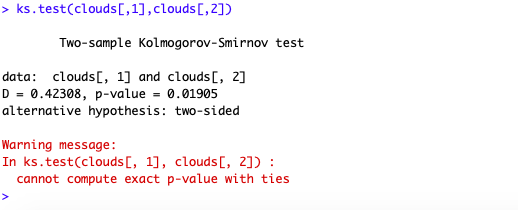
* 1. Mann-Whitney test



**Figure 12: 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.

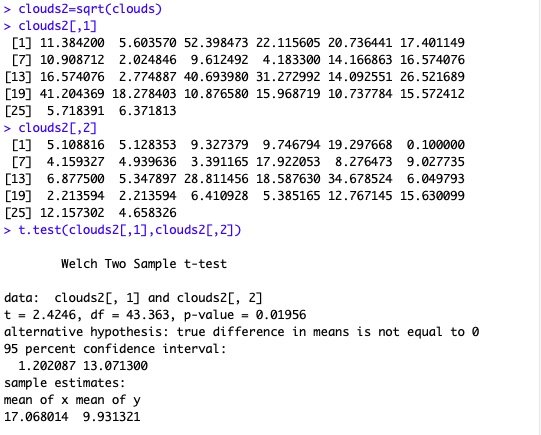
* 1. Kolmogorov-Smirnov test



**Figure 13: Kolmogorov-Smirnov test for clouds**

As p-value = 0.01905<0.05, 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.

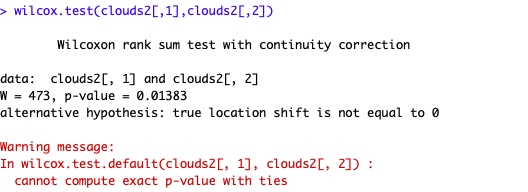
1. First we process the data to the square root values.
   1. two samples t-test



**Figure 14: Two sample t-test for**

As p-value is 0.01956<0.05 and less than the former one (0.05375) which is not square rooted, is rejected. Because after the square root calculation, the differences of data between the two samples have been narrowed, which indicates that the square root values of seeded and unseeded clouds have no significant difference.

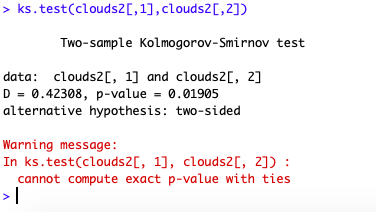
* 1. Mann- Whitney test



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

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.

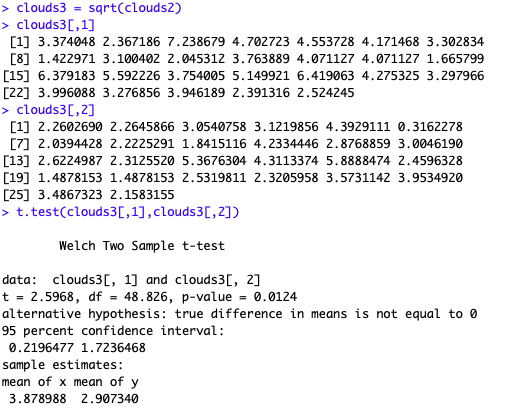
* 1. Kolmogorov-Smirnov test



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

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.

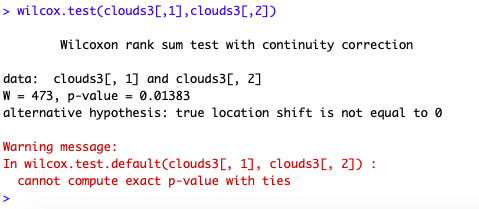
1. First we process the data to square root of the square root values.
   1. two samples t-test



**Figure 17: Two sample t-test for**

As p-value is 0.0124<0.05 and less than the former ones, is rejected. The reasons are same as the former one (2.1).

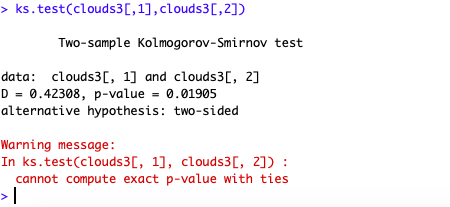
* 1. Mann-Whitney test



**Figure 18: Mann-Whitney test for**

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

* 1. Kolmogorov-Smirnov test

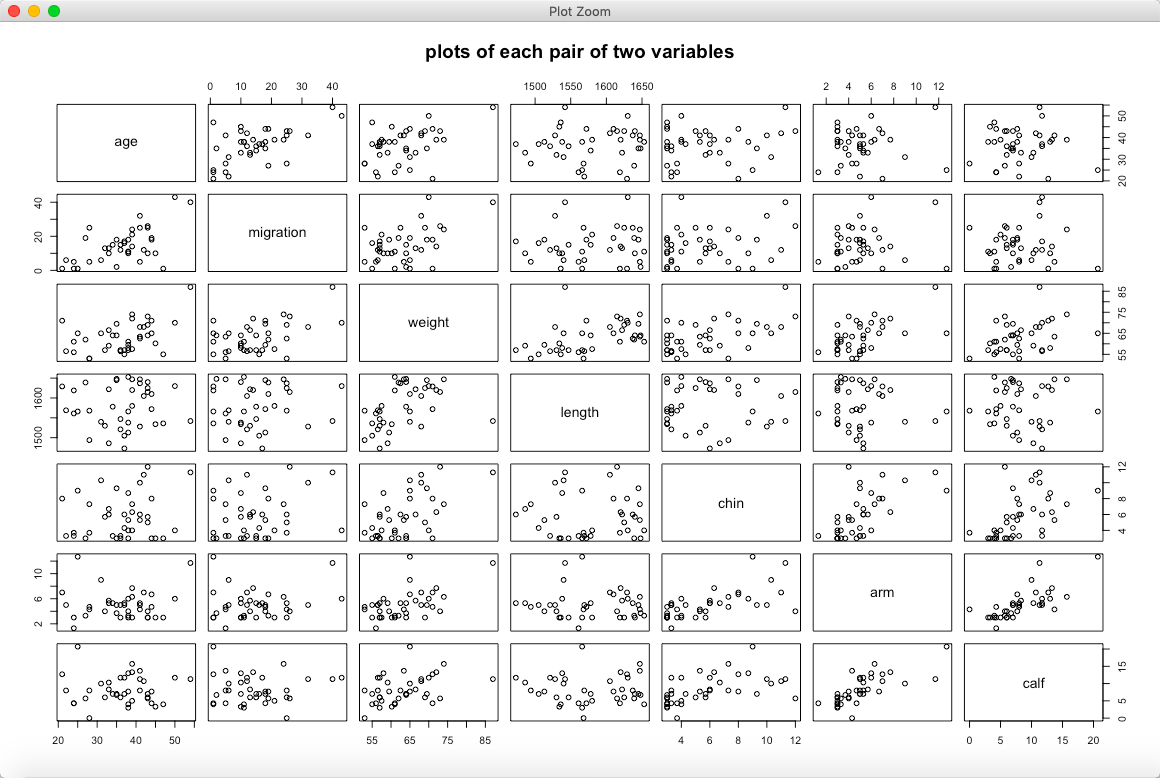


**Figure 19: Kolmogorov-Smirnov test for**

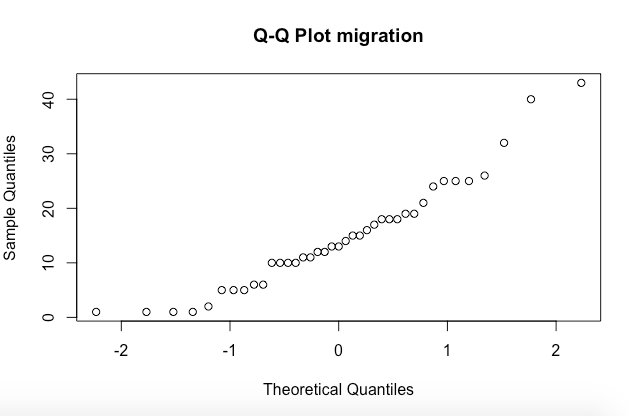
The reason of p-value did not change is as same as 2.3 had shown.

# Exercise 5

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

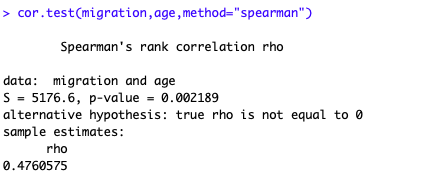


**Figure 20: plots of each pair of two variables**

1. 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 21: Q-Q plot migration**

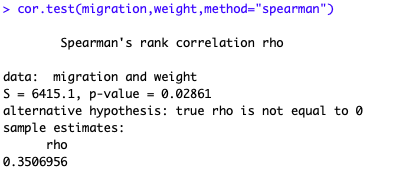
* 1. The rank correlation test between migration and age



**Figure 22: 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.

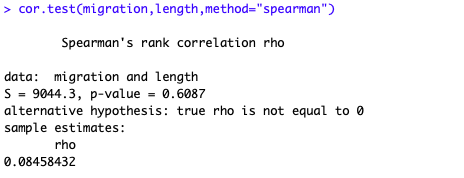
* 1. The rank correlation test between migration and weight



**Figure 23: 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.

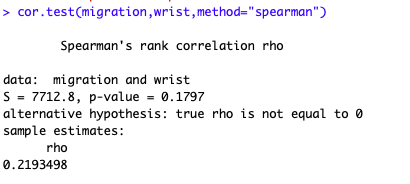
* 1. The rank correlation test between migration and length



**Figure 24: 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.

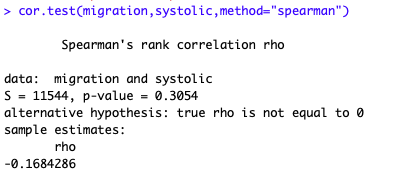
* 1. The rank correlation test between migration and wrist



**Figure 25: 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.

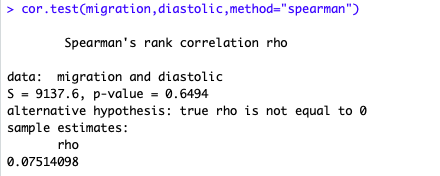
* 1. The rank correlation test between migration and systolic



**Figure 26: 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.

* 1. The rank correlation test between migration and diastolic



**Figure 27: 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**
2. **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:

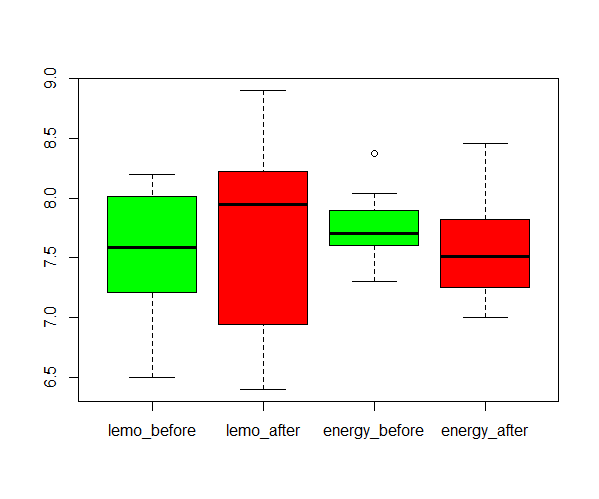
1. Energy drink group seems run faster after drinking the energy drink, because the time of “energy\_after” is shorter than the time of “energy\_before. Similarly, softdrink group seems run slower after drinking the softdrink, because the time of “lemo\_after” is longer than the time of “lemo\_before.
2. 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**

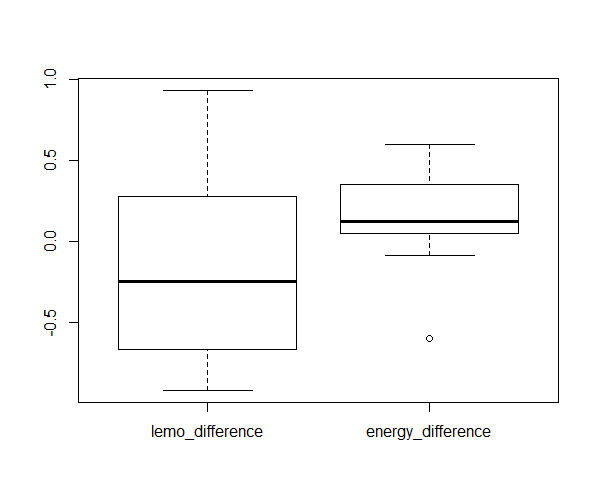
|  |  |  |
| --- | --- | --- |
|  | **mean** | **standard deviation** |
| **lemo\_before** | 7.554 | 0.5471157 |
| **lemo\_after** | 7.699 | 0.8339551 |
| **lemo\_difference** | -0.145 | 0.6232248 |
| **energy\_before** | 7.733 | 0.2971417 |
| **energy\_after** | 7.578 | 0.4585319 |
| **energy\_difference** | 0.154 | 0.3229258 |

1. **Graphical representations:**

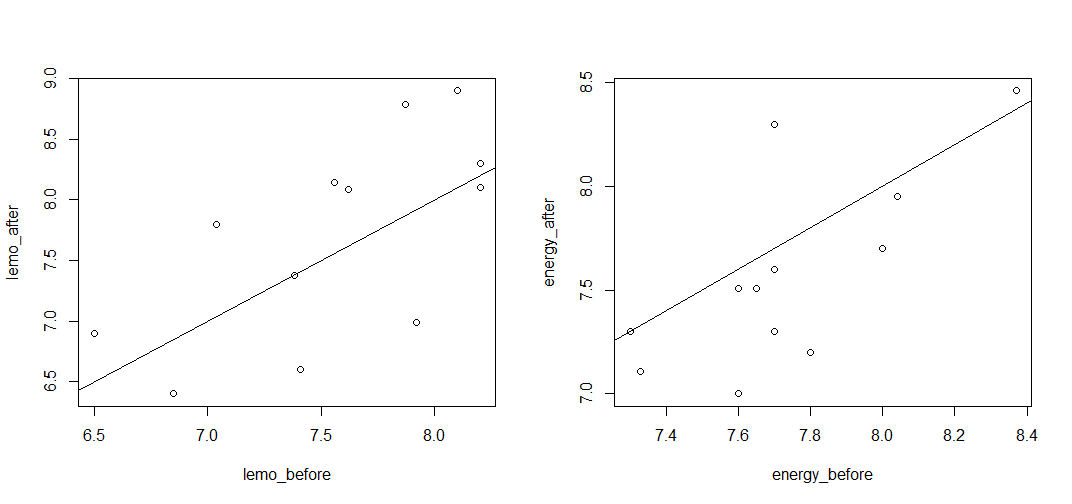
As shown on the figure28 and figure29, we can see the distribution of energy drink group is more concentrated.



**Figure28: boxplots of soft drink and energy drink (before and after)**



**Figure29: boxplots of soft drink and energy drink (difference)**



**Figure30 plot of before and after**

**code of (1):**

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

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

mydiff = c(lemo\_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\_difference"))

1. **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.

H0 = 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 H0, 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 H0, 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:

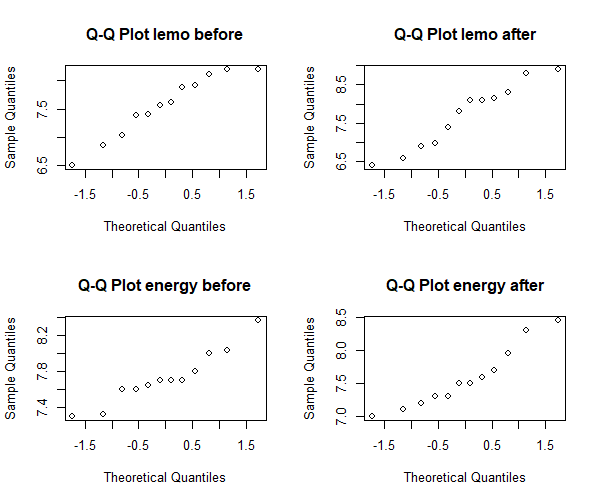
-0.05101059 0.35934392

sample estimates:

mean of the differences

0.1541667

Finally, we check the normality assumption on the differences.



**Figure 31, 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)

1. **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 anther 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

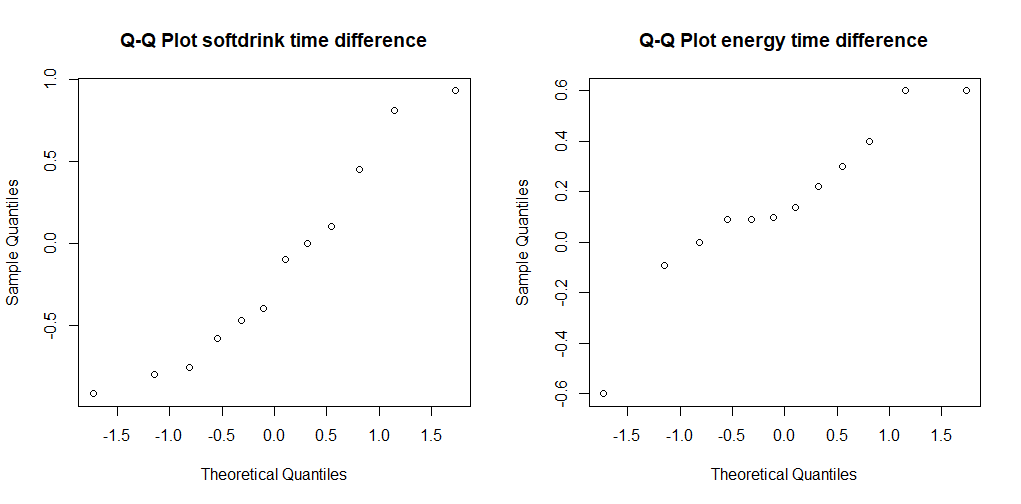
sample estimates:

mean of x mean of y

-0.1450000 0.1541667

The p-value=0.1586 > 0.05. Therefore, we cannot reject H0, which means there is not a time difference between soft drink and energy drink.

Finally, we check the normality assumption. Figure 32 shows the Q-Q plot of soft drink and energy drink time difference.



**Figure32. 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)

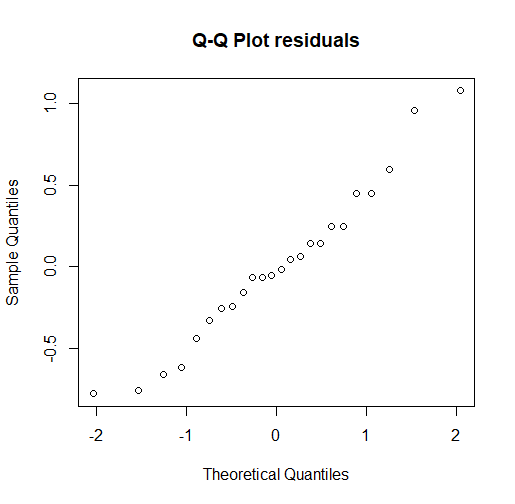
1. **plausible objection**

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 softdirnk and energy drink.
2. running on 500 meters
3. **Yes, because the t-test result of 3) also cannot reject H0.**
4. **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 34.



**Figure 34. 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,]

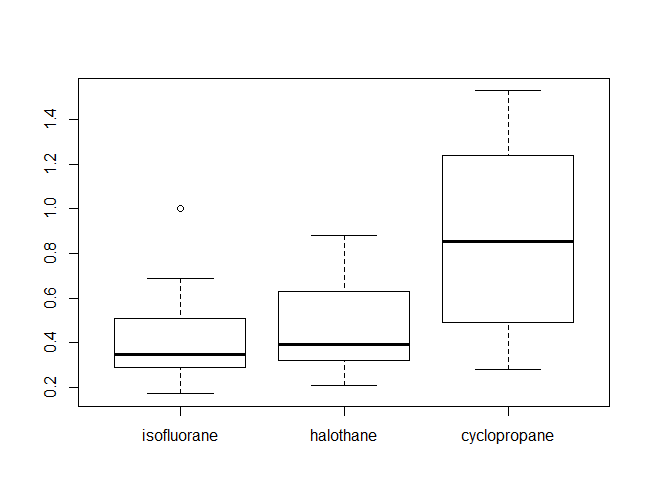
timeonaov= lm(yield~variety,data=diff\_frame)

anova(timeonaov)

par(mfrow=c(1,1)); qqnorm(residuals(timeonaov),main="Q-Q Plot residuals")

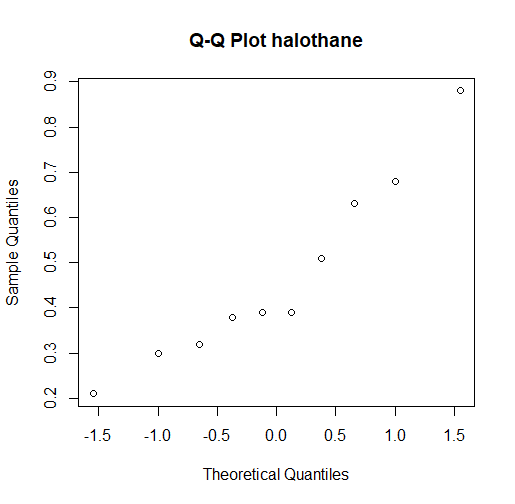
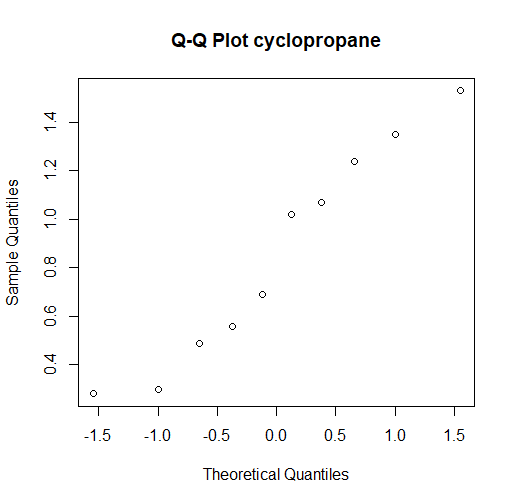
# EXERCISE 7

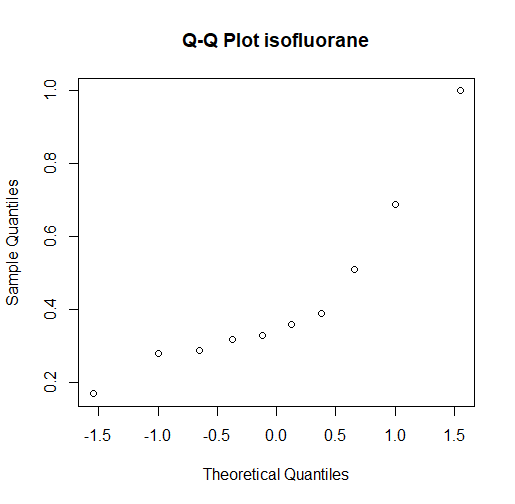
1. As the figure 35 shows, the boxplots of 3 samples.



**Figure 35. 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 36. 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")

1. **We use one-way ANOVA to test.**

The p-value=0.011< 0.05, so we reject H0. 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:

µ ˆ1 =0.4340; µ ˆ2 − µ ˆ1 = 0.0350; µ ˆ3 − µ ˆ1 = 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

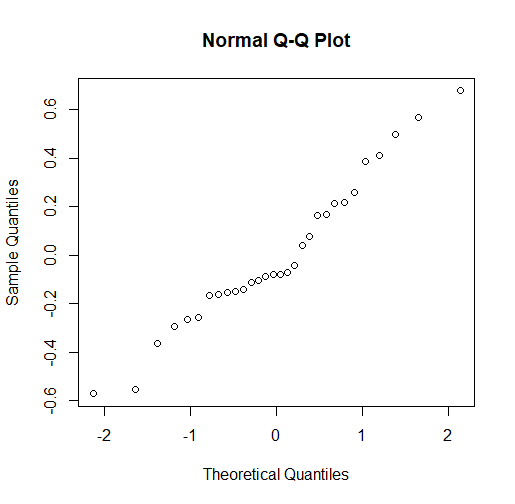
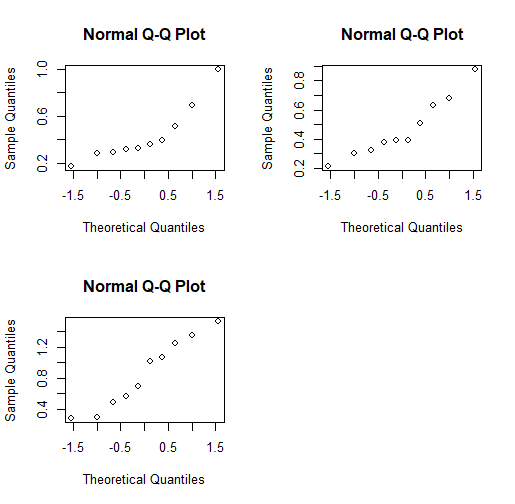
Residual standard error: 0.3177 on 27 degrees of freedom

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 37. 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))

1. **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,each=10)))

attach(dogframe)

kruskal.test(worms,group)

shapiro.test(isofl)

shapiro.test(haloe)

shapiro.test(cyclo)