February 28, 2018

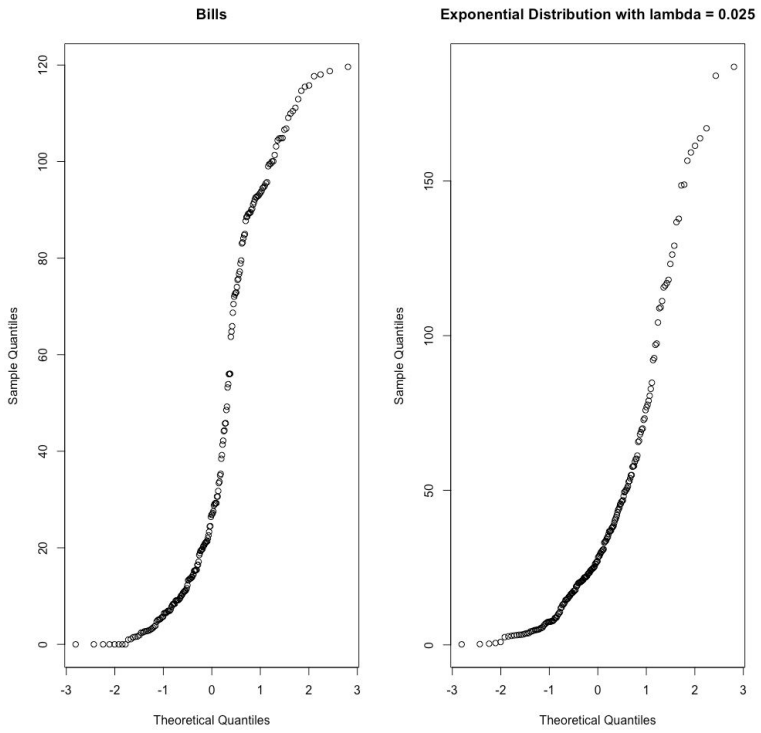
Assignment 2

Chao Zhang & Ibrahim Kanj
Group 13

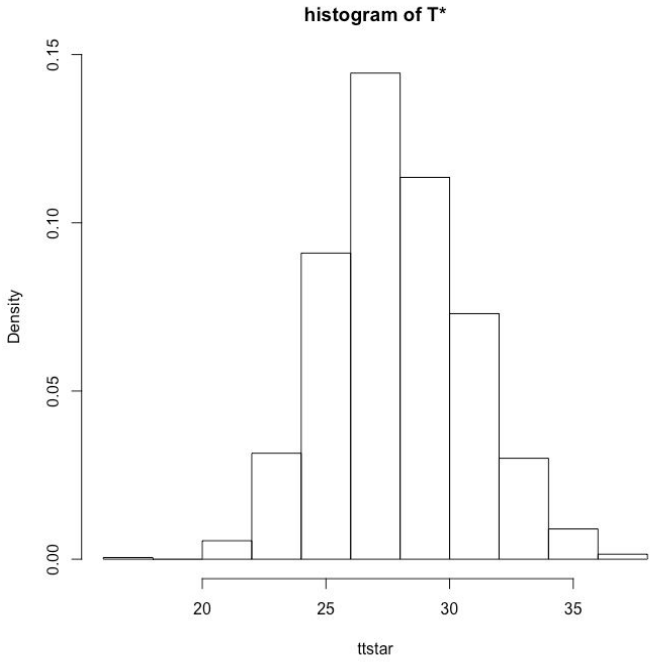
Exercise 1)

1) Checking λ from [0.01 , 0.1] using 0.005 steps. It seems that only for λ= 0.025 & 0.03 the data could stem from an exponential distribution where the bootstrap test had a p-value of (0.790 and 0.136 respectively).

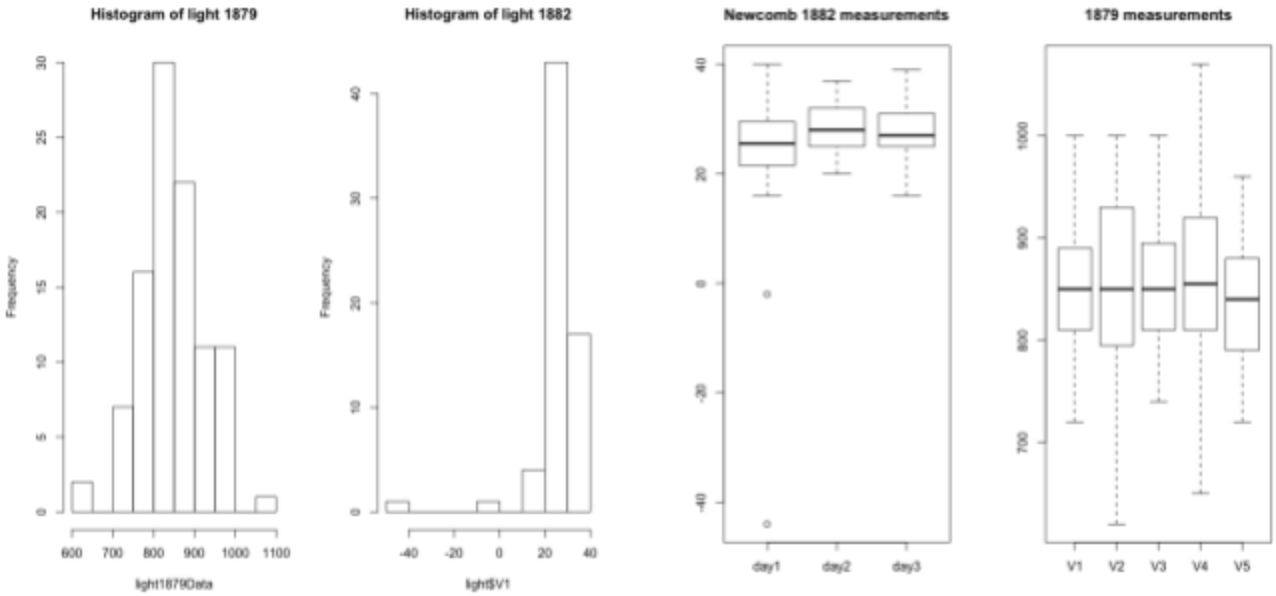
Plot of bills & Plot of an exponential distribution at λ = 0.025



2)



Exercise 2)



1) By looking at the histograms, it seems that 850 is the most frequent in 1879 and 26 in 1882 newcomb's measurements. Also the 1879 measurements could stem from a normal distribution.

If we change (850 and 26) to the speed of light we have 299,850 km/s (1879) and 299,766 km/s (1882).

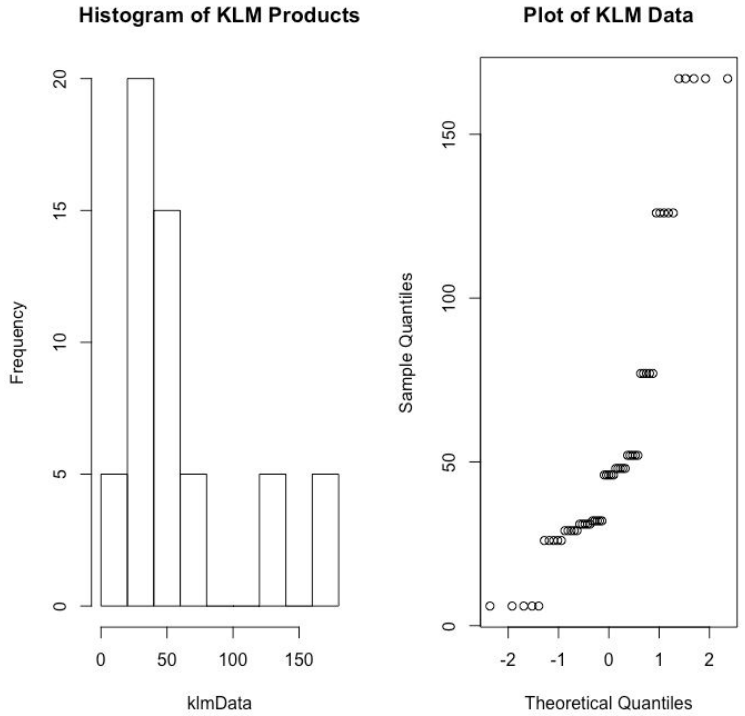
From the box plot we can see that the median is around 850 and 27 respectively.

2)The 95% confidence intervals for population means in 1879 are [299,836.6 , 299,868.2] and for 1882 are [299,730.27,299,790.52]
the 95% confidence intervals for population median in 1879 are [299,834.2 , 299,865.8] and for 1882 are [299,742.23,299,766.37]

3)The intervals show a huge difference between 1879 and 1882 where the difference is around 100 km/sec.

4)The speed of light is 299,792.458 km/sec which is different than the data back and even not included in the 95% confidence level.

Exercise 3)



1)Since the null hypothesis is median <= 31 days, then the alternate hypothesis is median > 31 thus we have a right tailed test. Since the data is not normal (from the histogram and qqnorm) and we have one sample then we can use the sign test.

the p-value is 0.058 thus the null hypothesis is not rejected (m = m0) and thus there are around N/2 items bigger/smaller than the median.

2) In the R code.
Summary: test failed because products that exceeds the 72 allowed days are more than 10% of the total products.

Exercise 4)

1) If paired t-test is used , the p-value is less than 0.05 then we could reject the null hypothesis and thus the true difference in means is different from 0. However t-test in our case is unpaired and it shows a p-value greater than 0.05 so we can not reject the null hypothesis.

We do not use paired t-test because our subjects are chosen from the start and they are independant (different 52 clouds) from which we decided to seed 26 (it is not the same as studying the same 26 clouds before and after seeding).

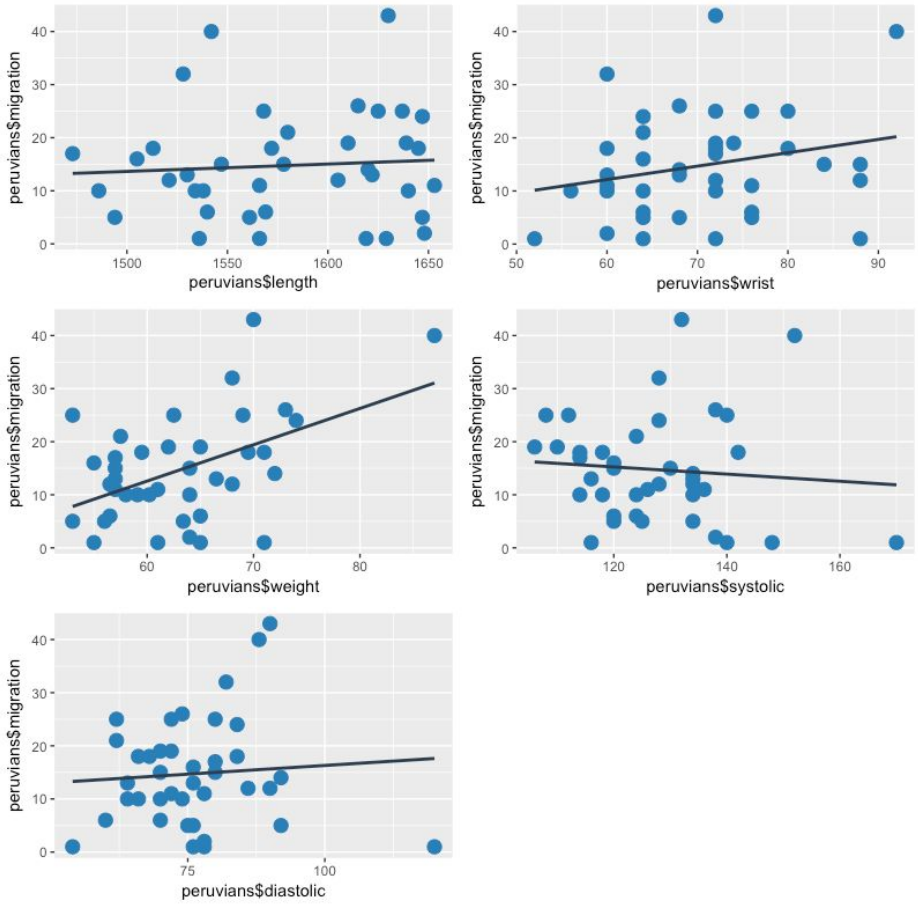
The Mann-Whitney test shows p-value less than 0.05 thus there is a true location shift of the unseeded values to the right from the seeded values.

The Kolmogorov-Smirnov test shows p-value less than 0.05 meaning the mean of unseeded is larger.

2)The p-value for the t-test became less than 0.05 thus we can say that there is a difference in the means between the seeded and unseeded clouds. However the Mann-Whitney and Kolmogorov-Smirnov show similar results as in number 1.

3)Similarly doing a square root of the square root, there have become a difference in the means between seeded and unseeded.

Exercise 5)



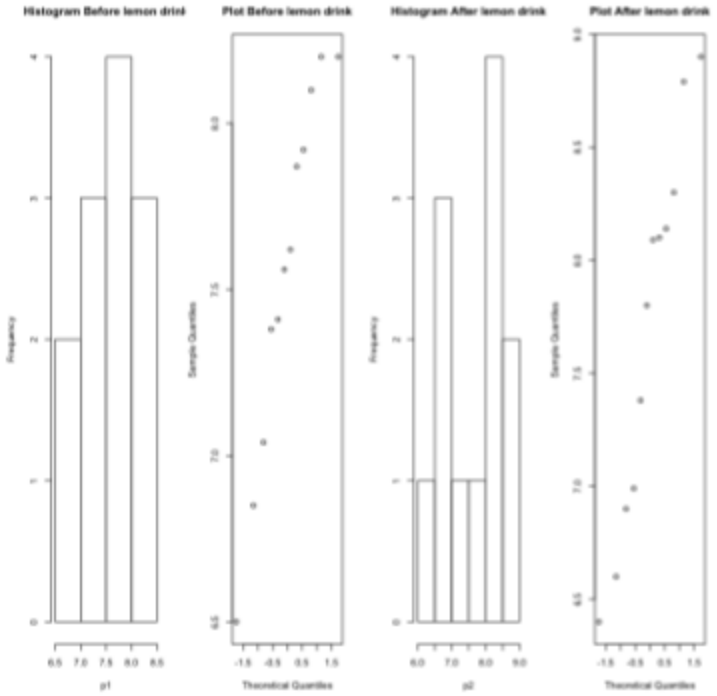
1)By having different variables plotted against migration, the only correlation seem to exist between weight and migration.

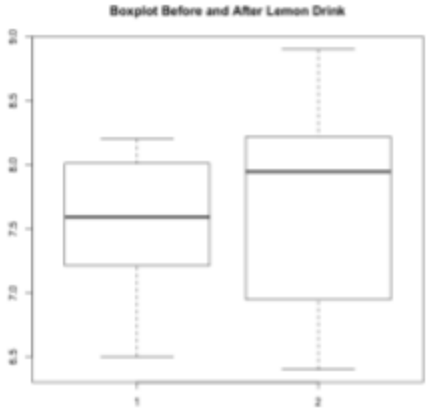
2) By looking at the p-value of the spearman rank correlation test we see that:

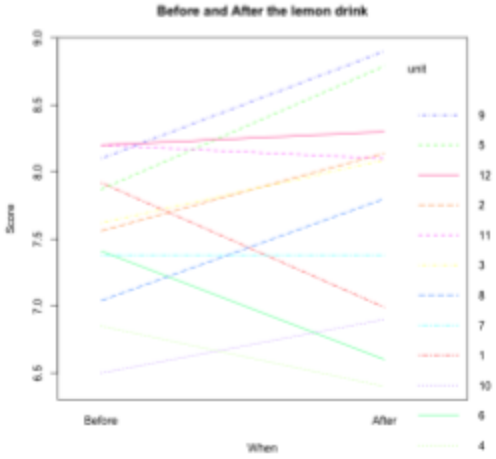
- migration ~ weight , p-value = 0.02861 thus null hypothesis is rejected and this means that there is a correlation between migration and weight.
- migration ~ length, p-value = 0.6087 thus we can not reject null hypothesis (correlation is zero)
- migration ~ wrist, p-value = 0.1797 thus we can not reject null hypothesis (correlation is zero)
- migration ~ systolic, p-value = 0.3054 thus we can not reject null hypothesis (correlation is zero)
- migration ~ diastolic, p-value = 0.6494 thus we can not reject null hypothesis (correlation is zero)

Exercise 6)

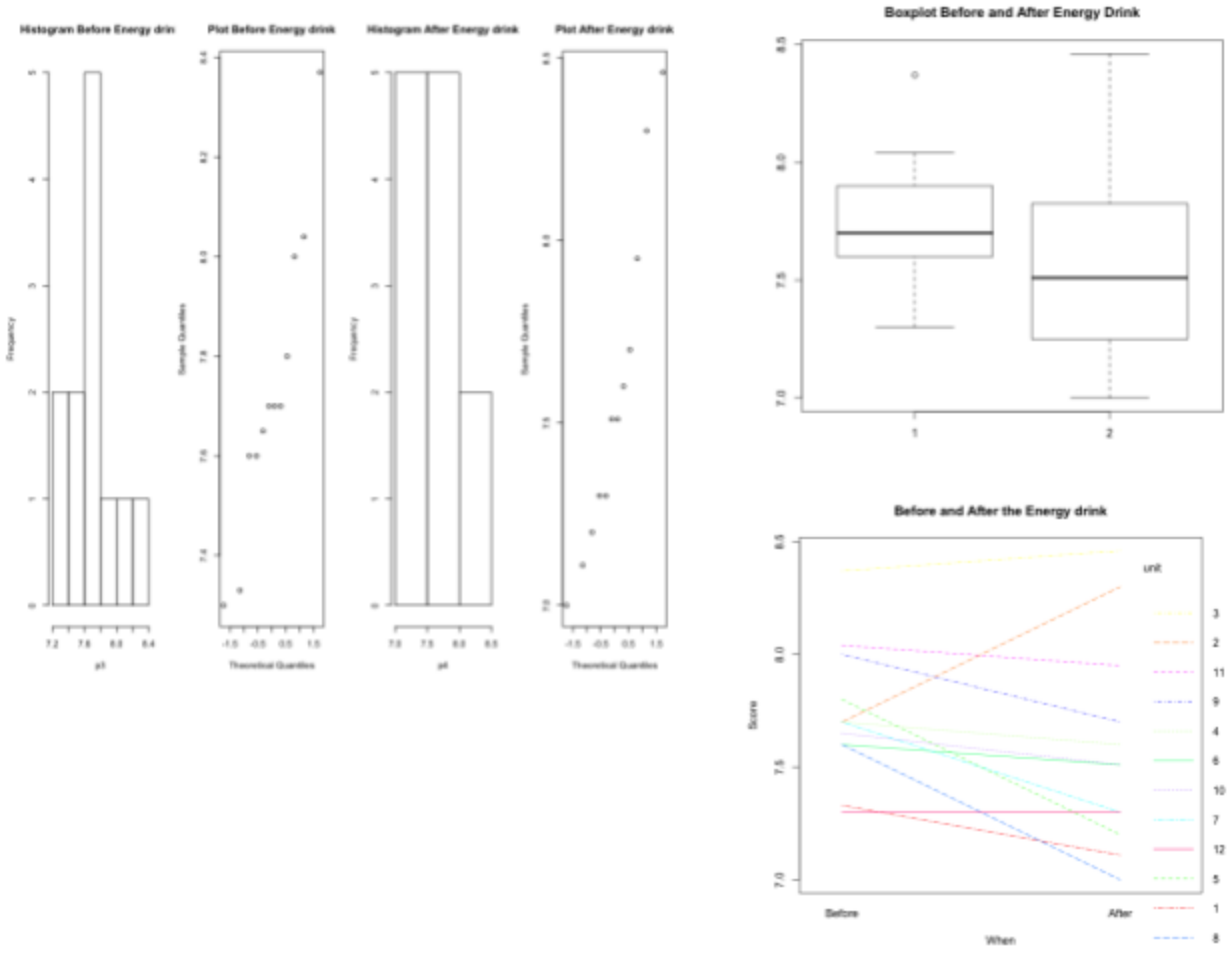
1) Lemon Drink







Energy Drink

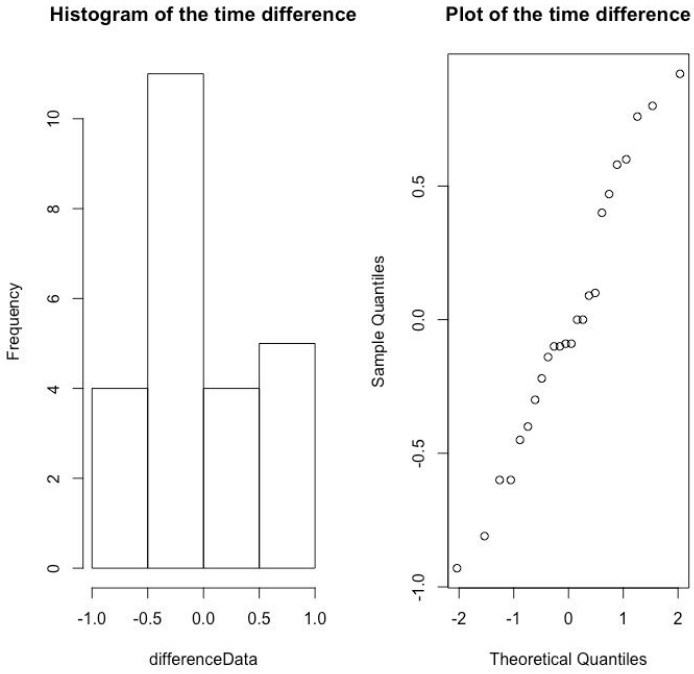


2)The histogram and plots show that we have a normal distribution so we can use the paired t-test.

The paired t-test on lemon shows that the p-value = 0.4373 thus we can not reject the null hypothesis => the difference in means is equal to 0 and thus lemon does not really enhance the running time.

Similarly the paired t-test on energy drink shows a p-value of 0.1264 => Energy drink does not enhance the running time.

3)



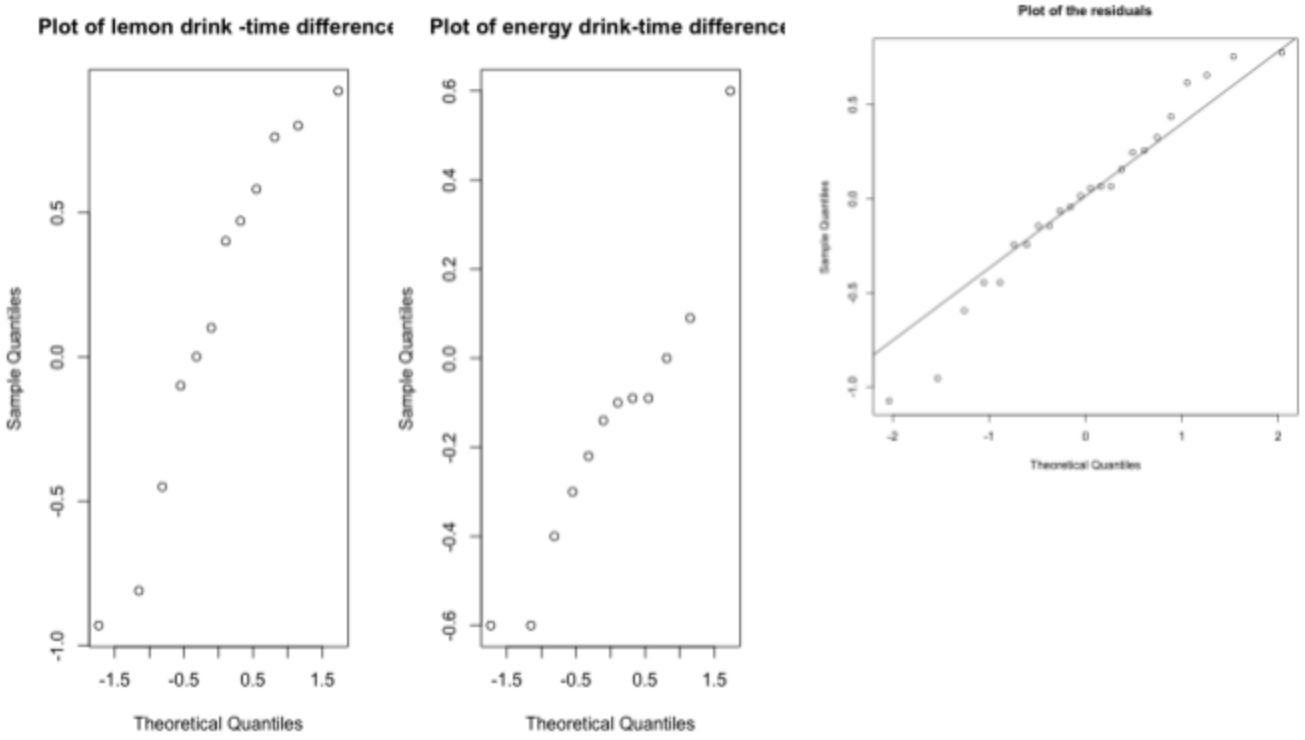
From the histogram and plot of the difference data we can see that the data is normal thus we can apply the 2 samples t-test. The resulting p-value is 0.1586 meaning that the null hypothesis is not rejected and that the difference in means is equal to 0. Thus there is not much difference between lemon and energy drink.

4)A better results could be obtained by having all the participants drinking energy drinks. If the purpose is to find whether the energy drink makes one faster or not then we should focus on increasing the number of observations. More time could also be helpful so that the run is not affected at all by any tiredness.

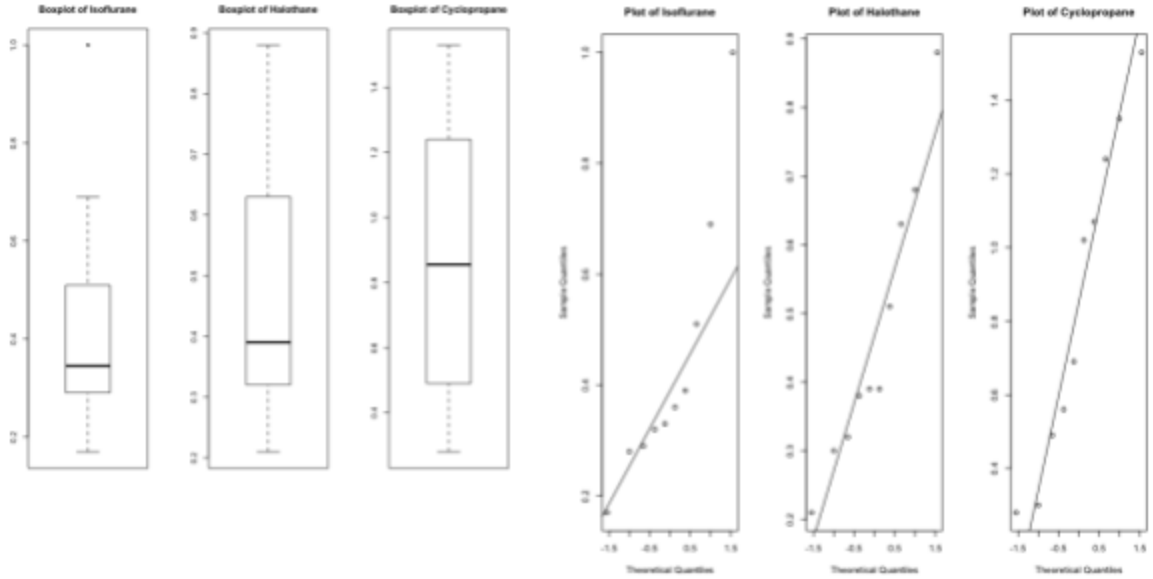
5)It could be more accurately tested by having the same person perform 3 runs.
One without drinking anything, then a run after drinking lemon and the last one after drinking energy drink. Thus one could test the two drinks while keeping the subject fixed, then repeat this scenario with 12 people.

6)The distributional assumptions are normal populations with population means and with equal population variances.

We can transform this vector into 24 residuals to investigate this assumption by performing anova.



Exercise 7)



1)From the plots it seems that isofluorane is not from a normal distribution, however both halothane and cyclopropane could be taken from a normal distribution.

2)By performing the 1-Anova test the p-value is 0.011 which is less than 0.05 thus we could reject the null hypothesis and say that the concentration differs according to different anesthesia. The estimated concentrations could be seen in anova's summary.

- concentration of plasma epinephrine for isofluorane is 0.434
- concentration of plasma epinephrine for halothane is 0.469
- concentration of plasma epinephrine for cyclopropane is 0.853

3)By performing Kruskal-Wallis test, we see that the p-value is 0.05948 then we can not reject the null hypothesis and so the concentrations is the same under the different drugs. The result here came contradicting to that of the 1-anova test, however this result is more accurate as the Anova has assumptions on the data and one of them is the data to be normal. Applying Anova on data which is not normal could lead to a wrong result.

APPENDIX

Full R Code

```
remove(list = ls())
par(mfrow=c(1,1))
#load(file="telephone.txt")

#Exercise 1)
bills = read.table("telephone.txt",header=TRUE)
boxplot(bills$Bills)
qqnorm(bills$Bills)

t=median(bills$Bills)
t
B=1000
tstar=numeric(B)
n=length(bills$Bills)

par(mfrow=c(1,2))
lambdaSeq = seq(0.01,0.1,by=0.005)
pValues = numeric(length(lambdaSeq))
for (aP in 1:length(lambdaSeq)){
  for (i in 1:B){
    xstar=rexp(n,lambdaSeq[aP])
    tstar[i]=median(xstar)
  }
  pl=sum(tstar<t)/B
  pr=sum(tstar>t)/B
  p=2*min(pl,pr)
  #pl;pr;p
  pValues[aP] = p
  #print(length(xstar))
}

#for lambda 0.025
ttstar=numeric(B)
for (i in 1:B){
  xxstar=rexp(n,0.025)
  ttstar[i]=median(xxstar)
}
par(mfrow=c(1,2))
qqnorm(bills$Bills,main = "Bills")
qqnorm(xxstar,main = "Exponential Distribution with lambda = 0.025")
```

```
#not sure if this is correct
hist(ttstar,prob=T, main="histogram of tstar & true density curve of T")
densmaxexp=function(x,n)  n*exp(-x)*(1-exp(-x))^(n-1)
lines(rep(t,2),seq(0,2*densmaxexp(t,n),length=2), type="l", col="red", lwd=3)
axis(1,t,expression(paste("t") ) )
u=seq(0,median(ttstar),length=1000)
lines(u,densmaxexp(u,n),type="l",col="blue")
#end not sure if this is correct

#Exercise 2)
remove(list = ls())
par(mfrow=c(1,1))

light = read.table("light.txt")
light1879 = read.table("light1879.txt")
light1882 = read.table("light1882.txt")

par(mfrow=c(1,2))
lightData = list()
lightData$day1 = light$V1[1:20]
lightData$day2 = light$V1[21:40]
lightData$day3 = light$V1[41:66]
boxplot(lightData) + title("Newcomb 1882 measurements")
boxplot(light1879) + title("1879 measurements")

light1879Data = numeric(100)
light1879Data[1:20] = light1879$V1[1:20]
light1879Data[21:40] = light1879$V2[1:20]
light1879Data[41:60] = light1879$V3[1:20]
light1879Data[61:80] = light1879$V4[1:20]
light1879Data[81:100] = light1879$V5[1:20]

hist(light1879Data,main = "Histogram of light 1879")
hist(light$V1, main = "Histogram of light 1882")

#2)

mean = mean(light1879Data)
standardDev = sd(light1879Data)
observations = length(light1879Data)
error <- 2*standardDev/sqrt(observations)
left <- mean-error
right <- mean+error
```

```
left
right

median = median(light1879Data)
standardDev = sd(light1879Data)
observations = length(light1879Data)
error <- 2*standardDev/sqrt(observations)
left <- median-error
right <- median+error
left
right

T1 = mean(light$V1)
B=1000
Tstar=numeric(B)
for(i in 1:B)
{
  Xstar=sample(light$V1,replace=TRUE)
  Tstar[i]=mean(Xstar)
}
Tstar25=quantile(Tstar,0.025)
Tstar975=quantile(Tstar,0.975)
sum(Tstar<Tstar25)
c(2*T1-Tstar975,2*T1-Tstar25)

TMed1 = median(light$V1)
B=1000
TMedstar=numeric(B)
for(i in 1:B)
{
  Xstar=sample(light$V1,replace=TRUE)
  TMedstar[i]=median(Xstar)
}
TMedstar25=quantile(TMedstar,0.025)
TMedstar975=quantile(TMedstar,0.975)
sum(TMedstar<TMedstar25)
c(2*TMed1-TMedstar975,2*TMed1-TMedstar25)
```

#Exercise 3)

```
remove(list = ls())
par(mfrow=c(1,1))
```

```
klm = read.table("klm.txt",header=TRUE)

klmData = numeric(55)
klmData[1:11] = klm$X70[1:11]
klmData[12:22] = klm$X70[1:11]
klmData[23:33] = klm$X70[1:11]
klmData[34:44] = klm$X70[1:11]
klmData[45:55] = klm$X70[1:11]

hist(klmData,main = "Histogram of KLM Products")
qqnorm(klmData, main = "Plot of KLM Data")

belowMed = sum(klmData<=31)
binom.test(belowMed,length(klmData),p=0.5)

#2)
aboveMed = sum(klmData>72)
newLength = length(klmData)/5 # IF p-value is not less than 0.05 then we can say than there
are (N/5)/2 products that exceed 72 days.
if (aboveMed < newLength)
{
  binom.test(aboveMed,newLength,p=0.5)
} else
{
  print("test failed where number of parts exceed the allowed 10%")
}
```

#Exercise 4)

```
remove(list = ls())
par(mfrow=c(1,2))
clouds = read.table("clouds.txt",header=TRUE)

boxplot(clouds$seeded)
boxplot(clouds$unseeded)

#p1 = hist(clouds$unseeded)
#p2 = hist(clouds$seeded)
#plot( p1, col=rgb(0,0,1,1/4))  # first histogram
#plot( p2, col=rgb(1,0,0,1/4),  add=T)

t.test(clouds$unseeded,clouds$seeded)
wilcox.test(clouds$unseeded,clouds$seeded)
```

```
ks.test(clouds$unseeded,clouds$seeded)

t.test(sqrt(clouds$unseeded),sqrt(clouds$seeded))
wilcox.test(sqrt(clouds$unseeded),sqrt(clouds$seeded))
ks.test(sqrt(clouds$unseeded),sqrt(clouds$seeded))

t.test(sqrt(sqrt(clouds$unseeded)),sqrt(sqrt(clouds$seeded)))
wilcox.test(sqrt(sqrt(clouds$unseeded)),sqrt(sqrt(clouds$seeded)))
ks.test(sqrt(sqrt(clouds$unseeded)),sqrt(sqrt(clouds$seeded)))
```

#Exercise 5)

```
remove(list = ls())
par(mfrow=c(1,5))
peruvians = read.table("peruvians.txt",header=TRUE)
#peruvians[,-c(5,6,7)]
library(ggplot2)
install.packages("gridExtra")
library(gridExtra)

plotLength = ggplot(peruvians, aes(x=peruvians$length, y=peruvians$migration)) +
  geom_point(color='#2980B9', size = 4) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE, color='#2C3E50' )
plotWrist = ggplot(peruvians, aes(x=peruvians$wrist, y=peruvians$migration)) +
  geom_point(color='#2980B9', size = 4) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE, color='#2C3E50')
plotWeight = ggplot(peruvians, aes(x=peruvians$weight, y=peruvians$migration)) +
  geom_point(color='#2980B9', size = 4) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE, color='#2C3E50')
plotSystolic = ggplot(peruvians, aes(x=peruvians$systolic, y=peruvians$migration)) +
  geom_point(color='#2980B9', size = 4) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE, color='#2C3E50')
plotDiastolic = ggplot(peruvians, aes(x=peruvians$diastolic, y=peruvians$migration)) +
  geom_point(color='#2980B9', size = 4) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE, color='#2C3E50')
grid.arrange(plotLength,plotWrist,plotWeight,plotSystolic,plotDiastolic)

par(mfrow=c(1,5))
qqnorm(peruvians$migration)
qqnorm(peruvians$length,main = "Plot of Length")
qqnorm(peruvians$wrist,main = "Plot of Wrist")
qqnorm(peruvians$weight,main = "Plot of Weight")
qqnorm(peruvians$systolic,main = "Plot of Systolic")
```

```
qqnorm(peruvians$diastolic,main = "Plot of Diastolic")

cor.test(peruvians$migration,peruvians$length,method="spearman")
cor.test(peruvians$migration,peruvians$wrist,method="spearman")
cor.test(peruvians$migration,peruvians$weight,method="spearman")
cor.test(peruvians$migration,peruvians$systolic,method="spearman")
cor.test(peruvians$migration,peruvians$diastolic,method="spearman")

#Exercise 6)
remove(list = ls())
par(mfrow=c(1,1))
run = read.table("run.txt",header=TRUE)

runLemon = run[1:12,]
runEnergy = run[13:24,]

p1 = runLemon$before
p2 = runLemon$after

par(mfrow=c(1,4))
hist(p1,main = "Histogram Before lemon drink")
qqnorm(p1,main = "Plot Before lemon drink")
hist(p2,main = "Histogram After lemon drink")
qqnorm(p2,main = "Plot After lemon drink")

par(mfrow=c(1,1))
boxplot(p1,p2)+title("Boxplot Before and After Lemon Drink")
before.new <- data.frame(score = p1, when = "Before",
                         unit = factor(1:12))
after.new <- data.frame(score = p2, when = "After",
                        unit = factor(1:12))
df.new <- rbind(before.new, after.new)
#display df.new to see the structure
df.new
# load nlme
library(nlme)
#attach the dataframe to call variables directly
attach(df.new)
#create plot
interaction.plot(when, unit, score, ylab = "Score", xlab = "When",
                 col = rainbow(12))+title("Before and After the lemon drink")

pEnergyBefore = runEnergy$before
```

```
pEnergyAfter = runEnergy$after
par(mfrow=c(1,4))
hist(pEnergyBefore,main = "Histogram Before Energy drink")
qqnorm(pEnergyBefore,main = "Plot Before Energy drink")
hist(pEnergyAfter,main = "Histogram After Energy drink")
qqnorm(pEnergyAfter,main = "Plot After Energy drink")

par(mfrow=c(1,1))
boxplot(pEnergyBefore,pEnergyAfter)+title("Boxplot Before and After Energy Drink")

# detach df.new to clean up
detach(df.new)

beforeEnergy.new <- data.frame(score = pEnergyBefore, when = "Before",
                     unit = factor(1:12))
afterEnergy.new <- data.frame(score = pEnergyAfter, when = "After",
                    unit = factor(1:12))
df.new <- rbind(beforeEnergy.new, afterEnergy.new)
#display df.new to see the structure
df.new
#attach the dataframe to call variables directly
attach(df.new)
#create plot
interaction.plot(when, unit, score, ylab = "Score", xlab = "When",
              col = rainbow(12)) + title("Before and After the Energy drink")
detach(df.new)

#2)

t.test(p1,p2,paired=TRUE)
t.test(pEnergyBefore,pEnergyAfter,paired=TRUE)

#3)

differenceData = numeric(24)
differenceData = run$after[1:24]-run$before[1:24]
par(mfrow=c(1,2))
hist(differenceData,main = "Histogram of the time difference")
qqnorm(differenceData,main = "Plot of the time difference")
t.test(differenceData[1:12],differenceData[13:24])

#6)
```

```
theDrinks = run$drink
theValues = differenceData
difDataByDrink = data.frame(drinks = theDrinks, values = theValues)
anovaDrink=lm(difDataByDrink$values~difDataByDrink$drinks,data=difDataByDrink)
anova(anovaDrink)
summary(anovaDrink)

par(mfrow=c(1,2))
qqnorm(difDataByDrink$values[1:12],main = "Plot of lemon drink -time difference")
qqnorm(difDataByDrink$values[13:24],main = "Plot of energy drink-time difference")

qqnorm(residuals(anovaDrink),main = "Plot of the residuals")
qqline(residuals(anovaDrink))
#residuals belong to normal distribution

#Exercise 7)
remove(list = ls())
par(mfrow=c(1,1))
concentrations = read.table("dogs.txt",header=TRUE)

par(mfrow=c(1,3))
boxplot(concentrations$isofluorane) + title("Boxplot of Isoflurane")
boxplot(concentrations$halothane) + title("Boxplot of Halothane")
boxplot(concentrations$cyclopropane) + title("Boxplot of Cyclopropane")

par(mfrow=c(1,3))
qqnorm(concentrations$isofluorane,main = "Plot of Isoflurane")
qqline(concentrations$isofluorane)
qqnorm(concentrations$halothane,main="Plot of Halothane")
qqline(concentrations$halothane)
qqnorm(concentrations$cyclopropane,main="Plot of Cyclopropane")
qqline(concentrations$cyclopropane)

#2)

concFrame
=data.frame(plasmaConc=as.vector(as.matrix(concentrations)),anesthesia=factor(rep(1:3,each=
10)))
anovaConc=lm(concFrame$plasmaConc~concFrame$anesthesia,data=concFrame)
anova(anovaConc)
summary(anovaConc)
confint(anovaConc)
```

#3)

```
kruskal.test(concFrame$plasmaConc,concFrame$anesthesia)
```