Assignment2

Sara Salimzadeh,

Tung Nguyên Hũu Tùng,

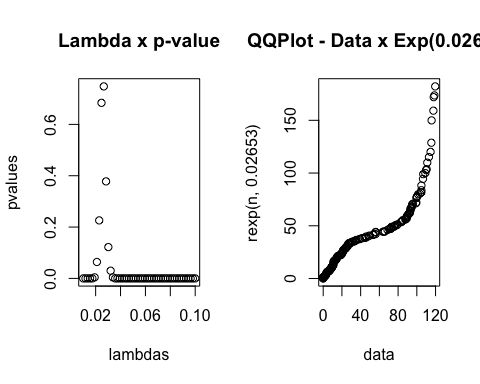
Felipe Santos Batista de Souza

Group 18

February 27

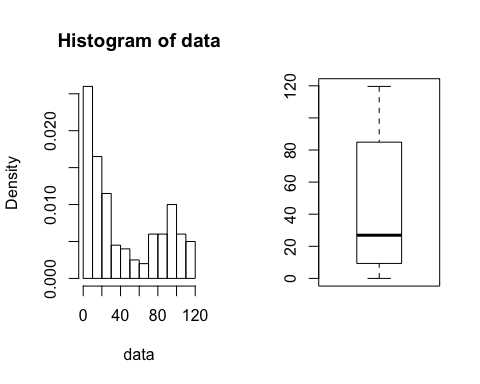
### Exercise 1  
  
# Question 1.1  
data <- read.table(file = "telephone.txt", header = TRUE)  
data = unlist(data, use.names = FALSE)   
  
# The statistic to be used is the median - t = median for the given population's sample  
t=median(data)   
  
numLambdas=50  
B=1000;   
tstar=numeric(B)  
n=length(data)  
  
#Initially the interval for lambda will be split into equidistant values for lambda. Afterwards, the p-value  
#will be calculated for each of them.  
  
#Spliting the lambda interval in 50 different points:  
lambdas=numeric(numLambdas)  
pvalues=numeric(numLambdas)  
  
lambdas=seq(0.01,0.1,length=50)  
  
for (j in 1:numLambdas) {  
 # Creating Xstars and surrogate Tstars  
 tstar=numeric(B) #Creating a Tstar vector to be used in the Bootstrap test  
 for (i in 1:B){  
 xstar=rexp(n,lambdas[j]) # generating simulated samples  
 tstar[i]=median(xstar) #generating surrogated ts  
 }  
  
 pl=sum(tstar<t)/B  
 pr=sum(tstar>t)/B  
 pvalues[j]=2\*min(pl,pr)  
}

par(mfrow=c(1,2))



# Question 1.2:  
# We generate the bootstrap interval, using the median as location estimator.  
  
TstarBootstrapInterval = numeric(B)  
for(i in 1:B){  
 Xstar = sample(data,replace=TRUE) #Generates a ramdon permutation of 'size of data' elements  
 TstarBootstrapInterval[i]=mean(Xstar) # Computes the statistic for the sample.  
}  
Tstar25=quantile(TstarBootstrapInterval,0.025)  
Tstar975=quantile(TstarBootstrapInterval,0.975)  
  
T1 = mean(data)  
c(2\*T1-Tstar975,2\*T1-Tstar25) # ==> mean is in this interval

## 97.5% 2.5%   
## 38.34240 48.80156.  
  
par(mfrow=c(1,2))



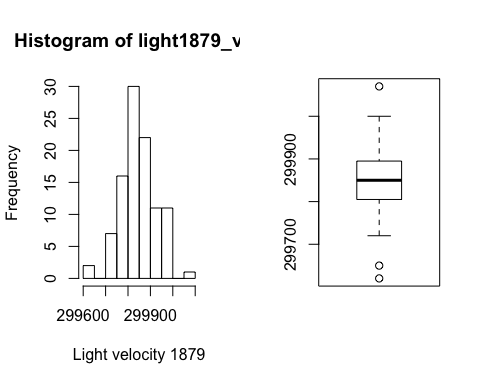
**Exercise 1**

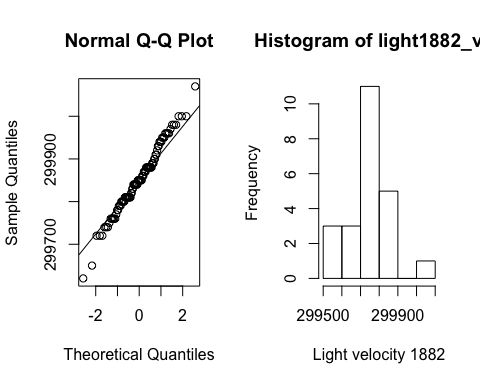
1. Considering the given interval for lambda, the approach to understand whether the data follows exponential distribution with a lambda within this interval, will be to compute the p-value for different values of lambda. The intention is to identify the behavior of p-value depending on the lambda. As can be seen from the graph, p-value is the highest (0.78) when lambda=0.02653. At this point we would fail to reject the null hypothesis with highest p-value, so it would be plausible to say that the data stems from an exponential distribution with lambda close to 0.02653. In fact, from the test, p-value is greater than 0.05 in the lambda interval between 0.021 and 0.032 (please note) that these values may slightly change every time the test is executed (since there are random variables involved). By plotting a QQ-plot between the data and exponential (0.02653) is also possible to see that a graph that resembles the 0-1 line which reinforces that the data could stem from exponential distribution with lambda=0.02653.
2. We generate the bootstrap interval, using the median as location estimator Conclusion: Considering the bootstrap confidence interval, the mean consumption for this population falls in the interval between 38.12714 and 48.77792. The boxplot shows that the distribution of consumption is not symmetric around the median, with a considerable scattered consumption between the median and the third quartile. On the other hand, the consumption figures between Q1 and median (so representing 25% of the data) are concentrated which could represent an opportunity for target market campions. There is also an opportunity for high-end consumers (above 80) since 25% of the consumers are located at this region of the boxplot.

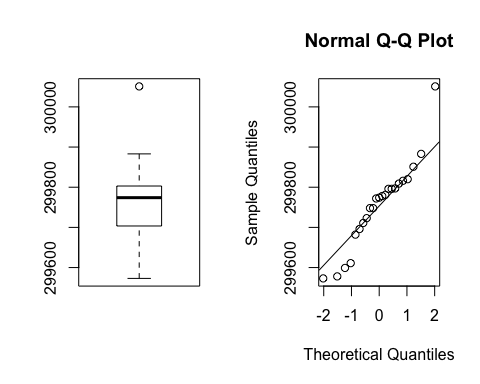
### Exercise 2

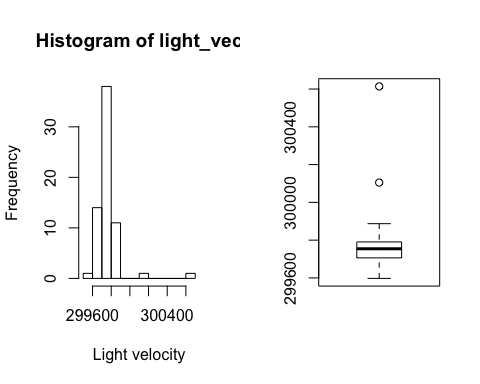
light1879\_dataframe=read.table("light1879.txt", header = FALSE)  
light1882\_dataframe=read.table("light1882.txt", fill = TRUE)  
light\_dataframe= read.table("light.txt", header = FALSE)  
  
light1879\_vec = unlist(light1879\_dataframe, use.names = FALSE)  
light1879\_vec = light1879\_vec + 299000  
  
light1882\_vec = unlist(light1882\_dataframe, use.names = FALSE)  
light1882\_vec <- light1882\_vec[!is.na(light1882\_vec)]  
light1882\_vec = light1882\_vec + 299000  
  
light\_vec = unlist(light\_dataframe, use.names = FALSE)  
light\_vec = 7.442 / (((light\_vec/1000) + 24.8)/1000000)  
  
# Question 2.1  
par(mfrow=c(1,2))  
  
median(light1879\_vec)

## [1] 299850









# Question 2.2  
# Dataset 1879  
#### MEAN   
B=1000  
Tstar=numeric(B)  
for(i in 1:B) {  
 Xstar=sample(light1879\_vec, replace=TRUE)  
 Tstar[i]=mean(Xstar)  
}  
Tstar25=quantile(Tstar,0.025)  
Tstar975=quantile(Tstar,0.975)  
Tmean= mean(light1879\_vec)  
c(2\*Tmean-Tstar975,2\*Tmean-Tstar25)

## 97.5% 2.5%   
## 299835.1 299867.9

#### Median  
Tstar=numeric(B)  
for(i in 1:B) {  
 Xstar=sample(light1879\_vec, replace=TRUE)  
 Tstar[i]=median(Xstar)  
}  
Tstar25=quantile(Tstar,0.025)  
Tstar975=quantile(Tstar,0.975)  
Tmedian = median(light1879\_vec)  
c(2\*Tmedian-Tstar975,2\*Tmedian-Tstar25)

## 97.5% 2.5%   
## 299830 299860

# Dataset 1882  
#### MEAN   
B=1000  
Tstar=numeric(B)  
for(i in 1:B) {  
 Xstar=sample(light1882\_vec, replace=TRUE)  
 Tstar[i]=mean(Xstar)  
}  
Tstar25=quantile(Tstar,0.025)  
Tstar975=quantile(Tstar,0.975)  
Tmean= mean(light1882\_vec)  
c(2\*Tmean-Tstar975,2\*Tmean-Tstar25)

## 97.5% 2.5%   
## 299711.1 299800.2

#### Median  
Tstar=numeric(B)  
for(i in 1:B) {  
 Xstar=sample(light1882\_vec, replace=TRUE)  
 Tstar[i]=median(Xstar)  
}  
Tstar25=quantile(Tstar,0.025)  
Tstar975=quantile(Tstar,0.975)  
Tmedian = median(light1882\_vec)  
c(2\*Tmedian-Tstar975,2\*Tmedian-Tstar25)

## 97.5% 2.5%   
## 299752 299825

# Dataset light.txt  
#### MEAN   
B=1000  
Tstar=numeric(B)  
for(i in 1:B) {  
 Xstar=sample(light\_vec, replace=TRUE)  
 Tstar[i]=mean(Xstar)  
}  
Tstar25=quantile(Tstar,0.025)  
Tstar975=quantile(Tstar,0.975)  
Tmean= mean(light\_vec)  
c(2\*Tmean-Tstar975,2\*Tmean-Tstar25)

## 97.5% 2.5%   
## 299730.5 299790.5

#### Median  
Tstar=numeric(B)  
for(i in 1:B) {  
 Xstar=sample(light\_vec, replace=TRUE)  
 Tstar[i]=median(Xstar)  
}  
Tstar25=quantile(Tstar,0.025)  
Tstar975=quantile(Tstar,0.975)  
Tmedian = median(light\_vec)  
c(2\*Tmedian-Tstar975,2\*Tmedian-Tstar25)

## 97.5% 2.5%   
## 299742.2 299766.4

# The confidence interval for μ with 95% confidence is measured as below  
sd\_light=sd(light\_vec)  
len\_sample=length(light\_vec)  
c(Tmean-2\*sd\_light/sqrt(len\_sample), Tmean+2\*sd\_light/sqrt(len\_sample))

## [1] 299731.9 299795.9

# Question 2.4  
# 1879 dataset is in normal distribution  
accurate\_light\_velocity = 299792.458  
t.test(light1879\_vec, mu=accurate\_light\_velocity, conf.level=0.95)

##   
## One Sample t-test  
##   
## data: light1879\_vec  
## t = 7.5866, df = 99, p-value = 1.824e-11  
## alternative hypothesis: true mean is not equal to 299792.5  
## 95 percent confidence interval:  
## 299836.7 299868.1  
## sample estimates:  
## mean of x   
## 299852.4

# 1882 and light.txt is not in normal distribution  
wilcox.test(light1882\_vec, mu= accurate\_light\_velocity)

## Warning in wilcox.test.default(light1882\_vec, mu =  
## accurate\_light\_velocity): cannot compute exact p-value with ties

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: light1882\_vec  
## V = 83, p-value = 0.09736  
## alternative hypothesis: true location is not equal to 299792.5

wilcox.test(light\_vec, mu= accurate\_light\_velocity)

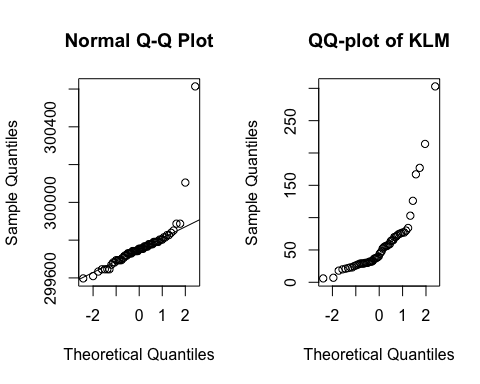
##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: light\_vec  
## V = 387, p-value = 4.451e-06  
## alternative hypothesis: true location is not equal to 299792.5

**Exercise 2**

1. Given the boxplot of the light1879.txt dataset, we can observe small number of outliers and the histogram is quite symmetric. We can presume that this dataset follows a normal distribution. In contrast, 1882 dataset might be not sampled normal distribution as we can observe some outliers in the boxplot diagram and the histogram is right skewed. The boxplot and histogram of last dataset - light.txt, seems to be symmetric but we can see some significant outliers in the first diagram. To evaluate the normality, we additionally used Shapiro-Wilk normality test. With p- value = 2.724e-12, we can reject the null hypothesis So the dataset doesn’t follow normal distribution.

|  |  |  |
| --- | --- | --- |
| Dataset | Mean confidence interval | Median confidence interval |
| light1879.txt | [299836.6, 299869.3] | [299830, 299860] |
| light1882.txt | [299713.2, 299800.9] | [299752, 299837] |
| light.txt | [299725.9, 299790.6] | [299742.2, 299772.4] |

1. As seen from the table above, the median confidence intervals of all datasets are smaller the mean confidence intervals respectively, which means that median confidence intervals can have more accurate estimation. This result is fairly reasonable as the estimating statistic x̄ is not robust against outliers that are located in all boxplots.
2. As mentioned above, the light1879.txt dataset can be sampled from normal distribution population. To test whether the result of this measurement is consistent with the currently most accurate value of the speed of light (299792.458), we use t-test with µ= 299792.458 (km/s). With p-value = 4.451e-06 < 0.05, we can reject the null hypothesis that the true mean of speed of light in this experiment is equal to 299792.458(most accurate value). Both light1882.txt and light.txt dataset do not follow the normal distribution so we apply Wilcoxon signed rank test those data. In the former test, the p-value = 0.09736 which means that we can’t reject the null hypothesis that the true mean is equal 299792.458, while the later test with p-value = 4.451e-06 indicates that we can reject the same null hypothesis. To sum up, the light1879 and light measurements are not equal to precise light speed while light1882 measurement can be assumed it is same as the most accurate speed of light (299792.458).



### Exercise 3  
  
# Question 3.1  
klm=scan("klm.txt")

m=sum(klm>31)  
n= length(klm)  
binom.test(m,n,p=0.5)

##   
## Exact binomial test  
##   
## data: m and n  
## number of successes = 40, number of trials = 60, p-value = 0.01349  
## alternative hypothesis: true probability of success is not equal to 0.5  
## 95 percent confidence interval:  
## 0.5331273 0.7831306  
## sample estimates:  
## probability of success   
## 0.6666667

# Question 3.2  
m=sum(klm>72)  
n= length(klm)  
binom.test(m,n,p=0.1)

##   
## Exact binomial test  
##   
## data: m and n  
## number of successes = 13, number of trials = 60, p-value =  
## 0.007478

## alternative hypothesis: true probability of success is not equal to 0.1  
## 95 percent confidence interval:  
## 0.1207160 0.3419562  
## sample estimates:  
## probability of success   
## 0.2166667

**Exercise 3**

1. According to the QQ-plot of KLM data, the sample distribution is not taken from normal population. For skewed distribution, the mean is highly influenced by the high/low values. In such cases, it is better to test location in terms of the median, instead of the mean.We can use the sign test or Wilcoxon test. However, the histogram of data shows that the distribution is not symmetric Therefore, the sign test is the best choice which prerequisites are all satisfied. The data are a random sample from a population with a certain median m. we test null hypothesis . The test statistics is which has the binary (N,0.5)- distribution under . Conclusion: As the p-value is 0.01, we can reject the null hypothesis. It means that the median of this population is bigger than 31.
2. This test is similar to sign test except the probability of success in binary distribution. we test null hypothesis . The test statistics is which has the binary (N,0.1)- distribution under . Conclusion: As the p-value is 0.007, we can reject the null hypothesis. It means that more than 10% of the parts arrives after the maximum delivery period of 72 days.

### Exercise 4   
  
# Question 4.1  
par(mfrow=c(1,4))  
cloud\_data = read.table("clouds.txt", header=TRUE)  
  
summary(cloud\_data$seeded)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.10 98.12 221.60 441.98 406.02 2745.60

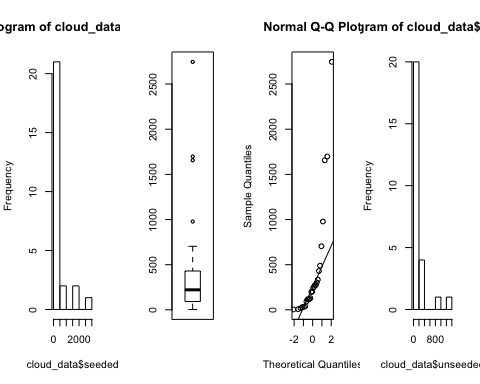
sd(cloud\_data$seeded)

## [1] 650.7872  
  
summary(cloud\_data$unseeded)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.01 24.82 44.20 164.56 159.20 1202.60

sd(cloud\_data$unseeded)

## [1] 278.4462



t.test(cloud\_data$seeded, cloud\_data$unseeded)

##   
## Welch Two Sample t-test  
##   
## data: cloud\_data$seeded and cloud\_data$unseeded  
## 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

wilcox.test(cloud\_data$seeded, cloud\_data$unseeded)

## Warning in wilcox.test.default(cloud\_data$seeded, cloud\_data$unseeded):  
## cannot compute exact p-value with ties

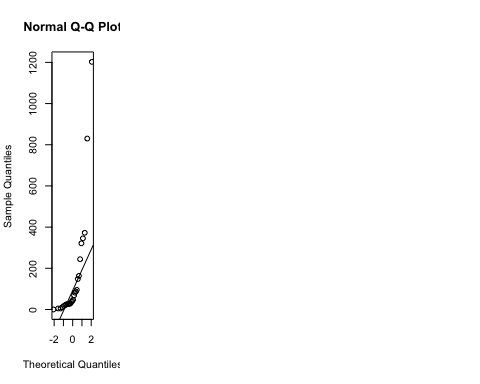
##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: cloud\_data$seeded and cloud\_data$unseeded  
## W = 473, p-value = 0.01383  
## alternative hypothesis: true location shift is not equal to 0

ks.test(cloud\_data$seeded, cloud\_data$unseeded)

## Warning in ks.test(cloud\_data$seeded, cloud\_data$unseeded): cannot compute  
## exact p-value with ties

##   
## Two-sample Kolmogorov-Smirnov test  
##   
## data: cloud\_data$seeded and cloud\_data$unseeded  
## D = 0.42308, p-value = 0.01905  
## alternative hypothesis: two-sided

# Question 4.2  
par(mfrow=c(1,4))



square\_root\_data = sqrt(cloud\_data)  
  
summary(square\_root\_data$seeded)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.025 9.894 14.870 17.068 20.122 52.398

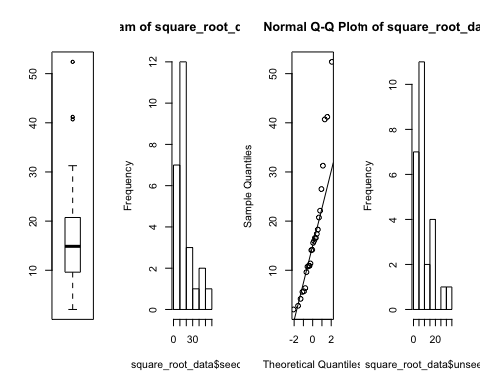
sd(square\_root\_data$seeded)

## [1] 12.51776  
  
summary(square\_root\_data$unseeded)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.100 4.982 6.644 9.931 12.615 34.679

sd(square\_root\_data$unseeded)

## [1] 8.28058



t.test(square\_root\_data$seeded, square\_root\_data$unseeded)

##   
## Welch Two Sample t-test  
##   
## data: square\_root\_data$seeded and square\_root\_data$unseeded  
## 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

wilcox.test(square\_root\_data$seeded, square\_root\_data$unseeded)

## Warning in wilcox.test.default(square\_root\_data$seeded, square\_root\_data  
## $unseeded): cannot compute exact p-value with ties

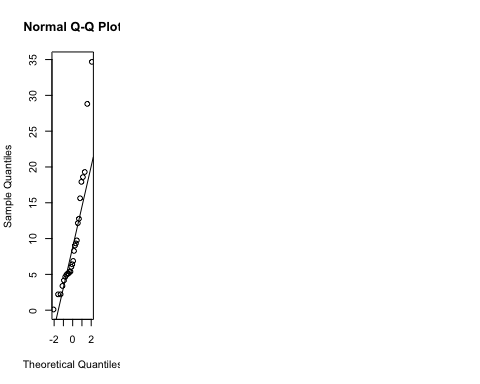
##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: square\_root\_data$seeded and square\_root\_data$unseeded  
## W = 473, p-value = 0.01383  
## alternative hypothesis: true location shift is not equal to 0

ks.test(square\_root\_data$seeded, square\_root\_data$unseeded)

## Warning in ks.test(square\_root\_data$seeded, square\_root\_data$unseeded):  
## cannot compute exact p-value with ties

##   
## Two-sample Kolmogorov-Smirnov test  
##   
## data: square\_root\_data$seeded and square\_root\_data$unseeded  
## D = 0.42308, p-value = 0.01905  
## alternative hypothesis: two-sided

# Question 4.3  
par(mfrow=c(1,4))



square\_root\_of\_square\_root\_data = sqrt(square\_root\_data)  
summary(square\_root\_of\_square\_root\_data$seeded)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.423 3.145 3.855 3.879 4.484 7.239

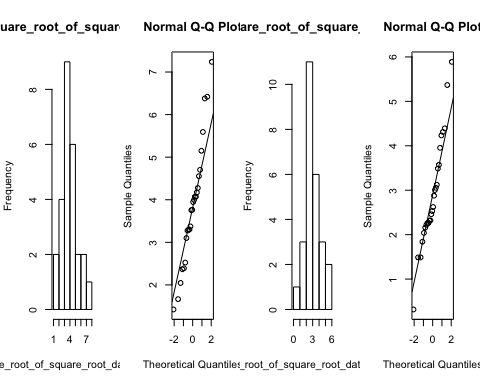
sd(square\_root\_of\_square\_root\_data$seeded)

## [1] 1.449941  
  
summary(square\_root\_of\_square\_root\_data$unseeded)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.3162 2.2320 2.5772 2.9073 3.5515 5.8888

sd(square\_root\_of\_square\_root\_data$unseeded)

## [1] 1.240097



t.test(square\_root\_of\_square\_root\_data$seeded, square\_root\_of\_square\_root\_data$unseeded)

##   
## Welch Two Sample t-test  
##   
## data: square\_root\_of\_square\_root\_data$seeded and square\_root\_of\_square\_root\_data$unseeded  
## 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

wilcox.test(square\_root\_of\_square\_root\_data$seeded, square\_root\_of\_square\_root\_data$unseeded)

## Warning in wilcox.test.default(square\_root\_of\_square\_root\_data$seeded,  
## square\_root\_of\_square\_root\_data$unseeded): cannot compute exact p-value  
## with ties

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: square\_root\_of\_square\_root\_data$seeded and square\_root\_of\_square\_root\_data$unseeded  
## W = 473, p-value = 0.01383  
## alternative hypothesis: true location shift is not equal to 0

ks.test(square\_root\_of\_square\_root\_data$seeded, square\_root\_of\_square\_root\_data$unseeded)

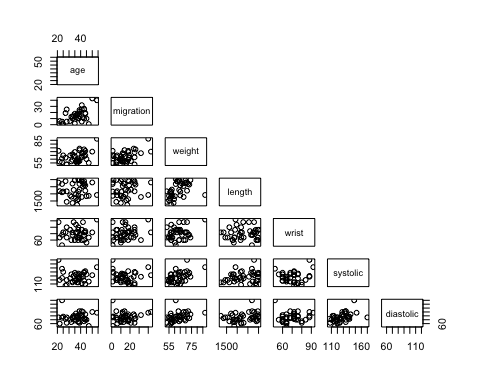
## Warning in ks.test(square\_root\_of\_square\_root\_data$seeded,  
## square\_root\_of\_square\_root\_data$unseeded): cannot compute exact p-value  
## with ties

##   
## Two-sample Kolmogorov-Smirnov test  
##   
## data: square\_root\_of\_square\_root\_data$seeded and square\_root\_of\_square\_root\_data$unseeded  
## D = 0.42308, p-value = 0.01905  
## alternative hypothesis: two-sided

**Exercise 4**

1. Given the histograms and QQ-plots of seeded and unseeded cloud data, we can't assume that they are in a normal distribution. As a result, we can't trust the result of the two samples t-test in this case because the assumption of the normal distribution in t-test was violated. On the other hand, the Mann- Whitney and the Kolmogorov-Smirnov test can be adopted in this case for the reason that both don't assume observations are taken from normal distribution. Mann- Whitney test has p-value = 0.01383 < 0.05, then we can reject that the populations of two samples are the same. Similarly, Kolmogorov-Smirnov test has p-value = 0.01905 < 0.05 so we can draw the same conclusion.
2. As we can't assume that the square root of seeded and unseeded cloud data is normal distribution. The assumption of the normal distribution in the two samples t-test was violated so we shouldn't apply the two samples t-test to the data. Mann- Whitney test and the Kolmogorov-Smirnov test can be adopted in this case for the reason that both don't assume observations are taken from normal distribution. The former test, Mann-Whitney, has p-value = 0.01383 < 0.05, we can conclude that of same populations is rejected. The later test, Kolmogorov-Smirnov, has p-value = 0.01905 < 0.05 so we can also reject the null hypothesis that the populations of two samples are the same.
3. After samples are transformed by square root of the square, the data for both seeded and unseeded clouds seems to follow the normal distribution which leads to the validity of the two samples t-test in the case. With p-value = 0.0124 < 0.05 in the two samples t-test, we can reject the null hypothesis : μ = ν that the means of two populations are the same. The “p-value” values of Mann-Whitney and Kolmogorov-Smirnov test remain the same in comparison with two previous cases (Question 1 and 2)- since the rankings of samples are not modified after applying square root and square root of square root function - and therefore we have the identical conclusions in those tests.

###Exercise 5  
  
# Question 5.1  
peruvians=read.table("peruvians.txt",header=TRUE)  
peruvians = peruvians[,-c(5,6,7)]  
pairs(peruvians, upper.panel=NULL)



# Question 5.2  
attach(peruvians)   
  
# Test 5.2.1 (migration x age)

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

## Warning in cor.test.default(migration, age, method = "spearman"): Cannot  
## compute exact p-value with ties

##   
## 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

# Test 5.2.2 (migration x weight)  
cor.test(migration, weight,method="spearman")

## Warning in cor.test.default(migration, weight, method = "spearman"): Cannot  
## compute exact p-value with ties

##   
## 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

# Test 5.2.3 (migration x length)  
cor.test(migration, length,method="spearman")

## Warning in cor.test.default(migration, length, method = "spearman"): Cannot  
## compute exact p-value with ties

##   
## 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

# Test 5.2.4 (migration x wrist)  
cor.test(migration, wrist,method="spearman")

## Warning in cor.test.default(migration, wrist, method = "spearman"): Cannot  
## compute exact p-value with ties

##   
## 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

# Test 5.2.5 (migration x diastolic)  
cor.test(migration, diastolic,method="spearman")

## Warning in cor.test.default(migration, diastolic, method = "spearman"):  
## Cannot compute exact p-value with ties

##   
## 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

**Exercise 5**

1. Based on the pairs is possible to see a potential correlation in the following pairs: migration x age, migration x weight, migration x wrist. This conclusion is based on the fact that the plot from these pairs resembles a straight line passing through the origin.
2. Tests will be conducted using Spearman rank correlation test which doesn't assume normality between the two variables.

* Test 5.2.1 (migration x age): As can be seen from the test, p-value = 0.0021, which led us to reject the null hypothesis that rho is equal to 0. In fact, the calculated rho based on the samples is 0.4760.
* Test 5.2.2 (migration x weight): As can be seen from the test, p-value = 0.02861, which led us to reject the null hypothesis that rho is equal to 0. In fact, the calculated rho based on the samples is 0.3506.
* Test 5.2.3 (migration x length): As can be seen from the test, p-value = 0.6087, which led us to fail to reject the null hypothesis that rho is equal to 0 - considering a 0.05 confidence level.

The calculated rho based on the samples is 0.0845 which is very close to 0 indicating a weak correlation between the variables.

* Test 5.2.4 (migration x wrist): As can be seen from the test, p-value = 0.1797, which led us to fail to reject the null hypothesis that rho is equal to 0 - considering a 0.05 confidence.

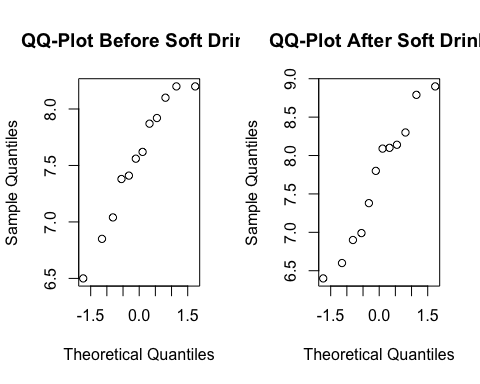
The calculated rho for the sample is rho = 0.2193.

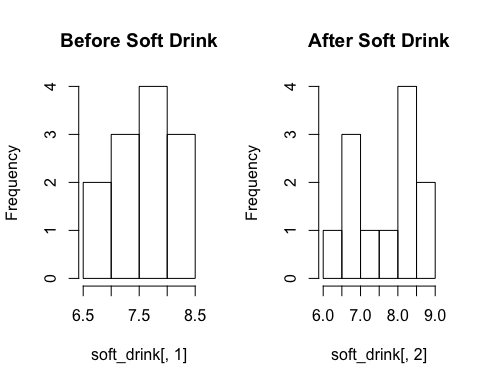
* Test 5.2.5 (migration x diastolic): As can be seen from the test, p-value = 0.6494, which led us to fail to reject the null hypothesis that rho is equal to 0 - considering a 0.05 confidence level.

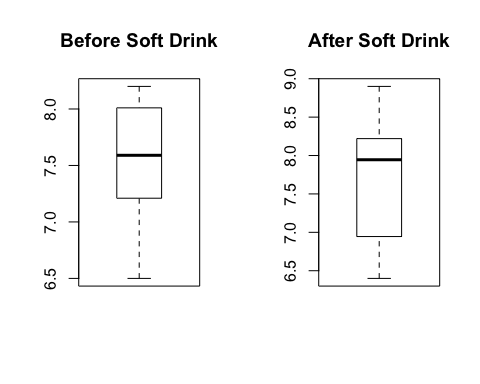
The calculated rho based on the samples is 0.0751 which is very close to 0 indicating a weak correlation between the variables.

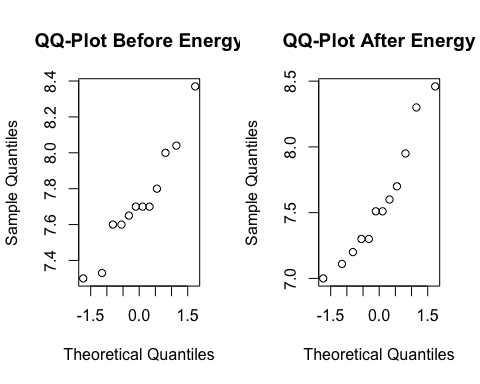
### Exercise 6  
  
# Question 6.1  
run=read.table("run.txt")

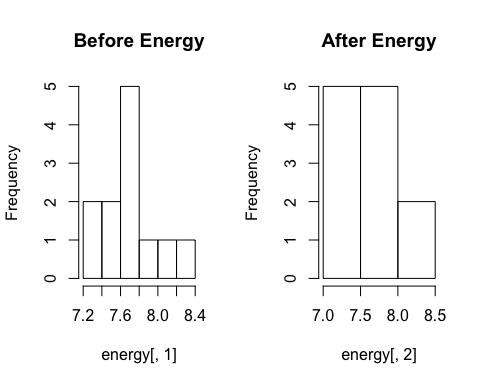
soft\_drink=run[run$drink=='lemo', 1:2]  
energy=run[run$drink=='energy', 1:2]  
  
par(mfrow=c(1,2))

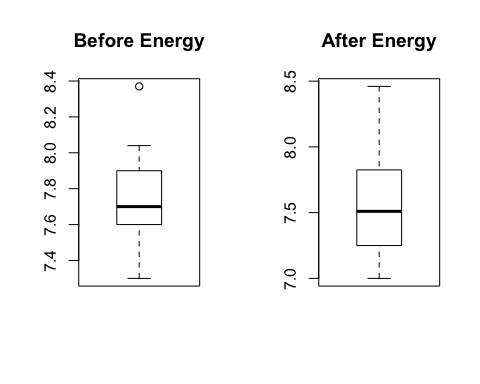








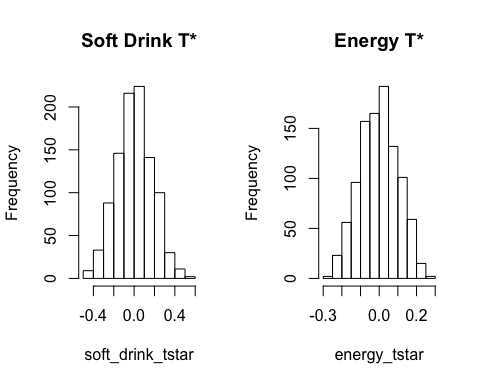




# Question 6.2  
soft\_drink=run[run$drink=='lemo', 1:2]  
energy=run[run$drink=='energy', 1:2]  
mystat=function(x,y) {mean(x-y)}  
B=1000  
  
#soft drink  
soft\_drink\_tstar=numeric(B)  
for (i in 1:B){  
 soft\_drink\_star=t(apply(cbind(soft\_drink[,1],soft\_drink[,2]),1,sample))  
 soft\_drink\_tstar[i]=mystat(soft\_drink\_star[,1],soft\_drink\_star[,2])  
}  
soft\_drink\_myt=mystat(soft\_drink\_star[,1],soft\_drink\_star[,2])  
pl=sum(soft\_drink\_tstar<soft\_drink\_myt)/B  
pr=sum(soft\_drink\_tstar>soft\_drink\_myt)/B  
p\_soft\_drink=2\*min(pl,pr)  
p\_soft\_drink

## [1] 0.52

#energy drink  
energy\_tstar=numeric(B)  
for (i in 1:B){  
 energy\_star=t(apply(cbind(energy[,1],energy[,2]),1,sample))  
 energy\_tstar[i]=mystat(energy\_star[,1],energy\_star[,2])  
}  
energy\_myt=mystat(energy\_star[,1],energy\_star[,2])



pl=sum(energy\_tstar<energy\_myt)/B  
pr=sum(energy\_tstar>energy\_myt)/B  
p\_energy=2\*min(pl,pr)  
p\_energy

## [1] 0.24

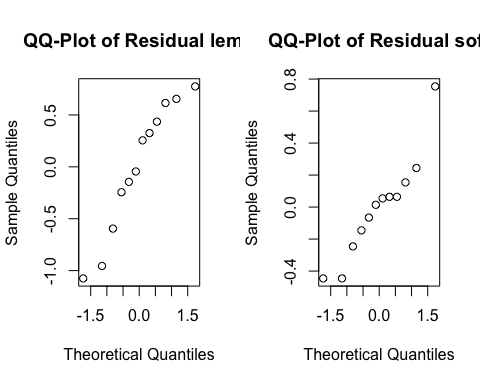
# Question 6.3  
n=nrow(run)  
diff\_time <- data.frame(numeric(n), character(n))  
diff\_time[,1]= run[,2]- run[,1]  
diff\_time[,2]= run[,3]  
wilcox.test(diff\_time[1:12,1],diff\_time[13:24,1])

## Warning in wilcox.test.default(diff\_time[1:12, 1], diff\_time[13:24, 1]):  
## cannot compute exact p-value with ties

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: diff\_time[1:12, 1] and diff\_time[13:24, 1]  
## W = 98.5, p-value = 0.1332  
## alternative hypothesis: true location shift is not equal to 0

# Question 6.6  
par(mfrow=c(1,2))

diff\_frame= data.frame(time\_diff=as.vector(diff\_time[,1]), drink\_type=factor(rep(1:2,each=12)))  
diff\_time\_aov= lm(time\_diff~drink\_type,data = diff\_frame)  
time\_residual=residuals(diff\_time\_aov)

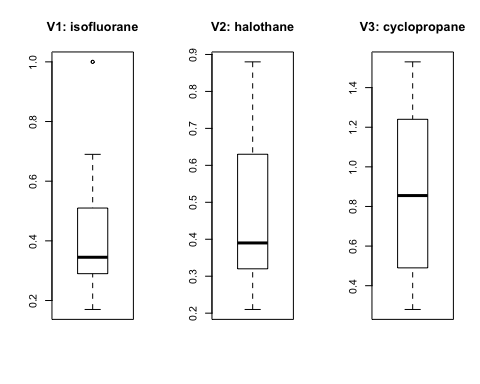


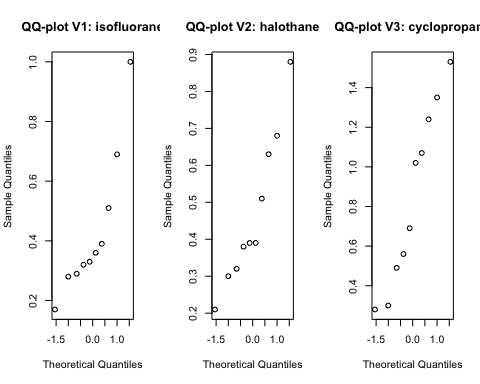
**Exercise 6**

1. To study the data, we plot the running time before and after both soft and energy drink separately. It includes QQ-plot, histogram and box plot of all of them. The median of running time increased slightly after soft drink. Moreover, the running time after drinking spread in larger range rather than before drinking. As the sample size is really small, it is hard to conclude the normality of population. QQ-plot of before and after for soft drink and their histograms demonstrates the samples are not normal. Similarly, in case of energy drink, the histograms and QQ-plots prove that the sample is not taken from normal population for both before and after energy drink. The median of running time decreased after energy drink. Additionally, the running time after drinking spread in larger range than before.
2. Since we can’t assume normality for both soft drink and energy, we should apply permutation test for two cases separately. We generate 1000 randomly chosen permutations to estimate the distribution of our test statistic under . Conclusion: In case of soft drink, since the p-value is about 0.16 we can’t reject null hypothesis. In terms of energy, the p-value is 0.29 and still greater than critical value and similarly we fail to reject null hypothesis here.
3. In such case, we have two different groups without normal population. Therefore, we apply Man-Whitney test. The sample stems from soft drink population (S) and similarly originates from energy drink (E). We test null hypothesis that the populations are the same. Conclusion: of equal means is not rejected. The underlying distribution of time difference in soft drink is similar to energy drink.
4. Time differences (after - before) are negative numbers for some cases. It may because of dependency of two measurement (before and after) in a short period of time. Additionally, the sample size for both cases seems small. The distribution of time difference is not normal in soft and energy drink and we can’t use t-samples t-test.
5. Yes, we have similar objections here.
6. In part 3, we took advantage of Man-Whitney test that it doesn’t require to presume on the distribution. To transform vector to residuals, we should assume normality of time difference in both drink types and apply ANOVA and residual function.

### Exercise 7  
  
# Question 7.1  
dogs=read.table("dogs.txt",header = TRUE)

par(mfrow=c(1,3))





# Question 7.2  
dogsframe= data.frame(plasma=as.vector(as.matrix(dogs)), drugs=factor(rep(1:3,each=10)))  
dogsaov= lm(plasma~drugs,data = dogsframe)  
anova(dogsaov)

## Analysis of Variance Table  
##   
## Response: plasma  
## Df Sum Sq Mean Sq F value Pr(>F)   
## drugs 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

summary(dogsaov)

##   
## Call:  
## lm(formula = plasma ~ drugs, data = dogsframe)  
##   
## 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 \*\*\*  
## drugs2 0.0350 0.1421 0.246 0.807266   
## drugs3 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

drug1 = 0.4340  
drug2 = drug1 + 0.0350  
drug3 = drug1 + 0.4190  
drug1; drug2; drug3

## [1] 0.434

## [1] 0.469

## [1] 0.853

# Question 7.3  
attach(dogsframe)  
kruskal.test(plasma,drugs)

##   
## Kruskal-Wallis rank sum test  
##   
## data: plasma and drugs  
## Kruskal-Wallis chi-squared = 5.6442, df = 2, p-value = 0.05948

**Exercise 7**

1. According to the QQ-plot of these three drugs, we just can assume samples in drug 3(cyclopropane) are taken from normal population. Since the sample size isn’t large enough, the normality can be doubtful. Drug 1 and 2 are certainly not normal.
2. The estimated concentration for drug 1 is equal 0.4340. drug 2 and drug 3 are 0.399 and 0.015 respectively. Since the p-value is 0.11, we reject null hypothesis which means for some (i, j).
3. In this part, the p-value is 0.059. Therefore, the null hypothesis in not rejected (but not with a great difference). . As we concluded in first part, the samples are not taken from normal population. However, ANOVA assumes normality to test the data. Therefore, the conclusion of ANOVA is against Kruskal-Wallis result.