Assignment 2

Tommy Maaiveld, Krishnakanth Sasi, Halil Kaan Kara, Group 6

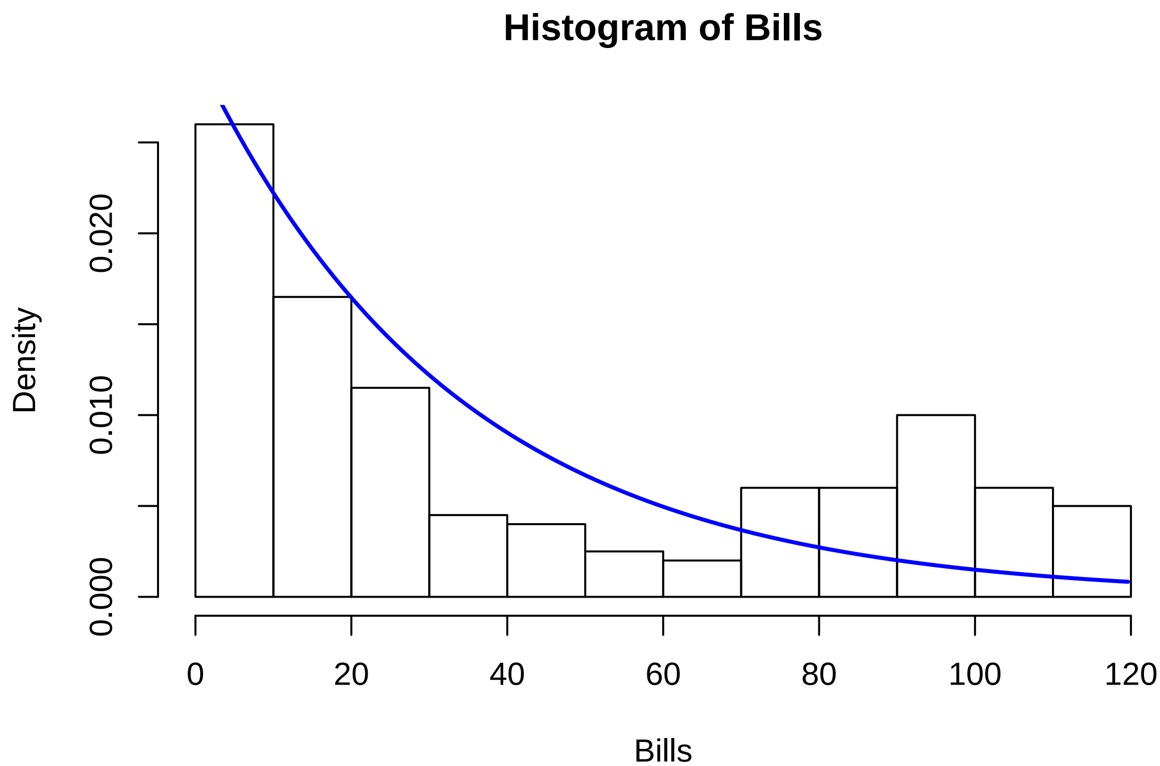
Introduction

This document describes the solutions found and implemented for the exercises of assignment 2. Exercises can be found in their corresponding sections. This document is created by Rmd, and figure captions are omitted since it changes the structure of the document in a bad way that makes it hard to follow.

Exercise 1

In this exercise, bootstrap test is used to determine whether the distribution of the data set is exponential and its λ is between [0.01, 0.1].

```r
data = read.table("./data/telephone.txt", header = TRUE)

B = 2000 # Iterations times

# Bounds for rate of exponential distribution
rateLower = 0.01
rateUpper = 0.1
rateIncrease = 0.01

n = length(data$Bills)

# Setup for loop
tStar = numeric(B)
t = median(data$Bills)


hist(data$Bills, freq = FALSE, main="Histogram of Bills", xlab = "Bills")
x=seq(0, max(data$Bills), length=1000)
lines(x, dexp(x, 0.03), type = "l", col="blue", lwd=2)
```



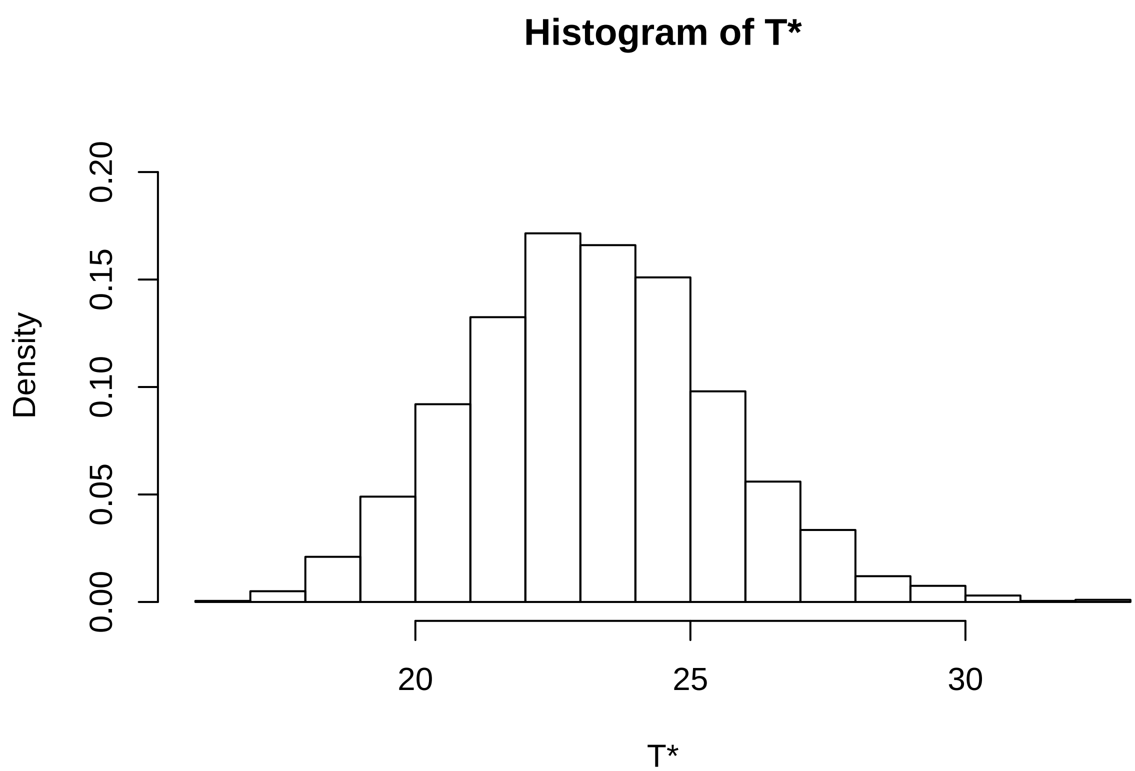
The loop that tries values for λ is given below. All values of $H_0 : \lambda = [0.01, 0.1]$ are rejected except for $H_0 : \lambda = 0.03$. The curve for $\lambda = 0.03$ can be seen in the first graph. However, this density curve does not look like the distribution of T* histogram which can be seen below. Since p-value of the $H_0 : \lambda = 0.03$ is greater than 5%, the test is inconclusive. The result of the p-value for $H_0 \lambda = 0.03$ can be seen after the code snippet.

```
# Try for all rates
for(rate in seq(from=rateLower, to=rateUpper, by=rateIncrease)) {
  for(iter in seq(from=0, to=B, by=1)) {
    # Get surrogate X*s from exponential distribution
    # with same size as the original data set
    sample = rexp(n, rate)

    # Store T* values for future comparison
    tStar[iter] = median(sample)
  }

  # Calculate p-value according to the slides of week-2
  pl = sum(tStar<t) / B
  pr = sum(tStar>t) / B
  p = 2*min(pl, pr)

  if (p > 0.05) {
    print(sprintf("H0: Rate: %.2f P-Value: %.2f is not rejected.", rate, p))
    break
```

```
  }
}
```

```
## [1] "H0: Rate: 0.03 P-Value: 0.13 is not rejected."
```

```
# Try to plot it with same graph style in week-2/30th slide
par(mfrow=c(1,1))
hist(tStar, probability=TRUE, ylim=c(0, 0.22), main="Histogram of T*", xlab = "T*")
```

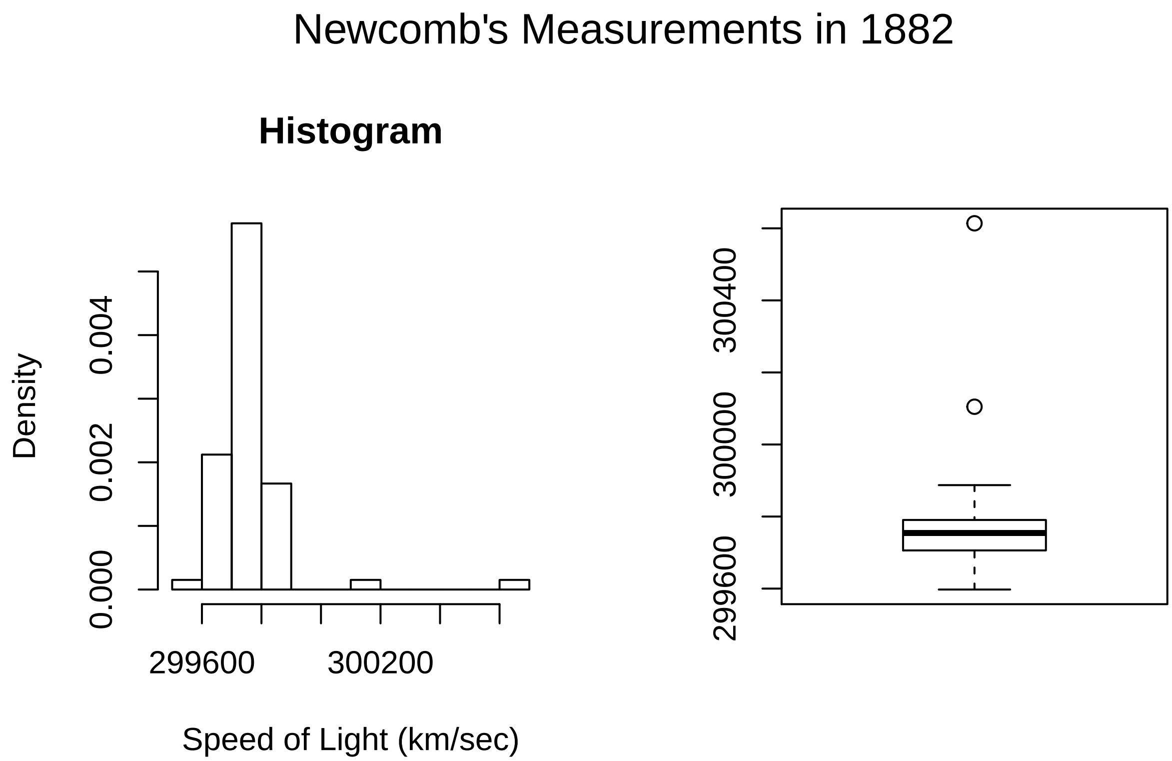


Exercise 2

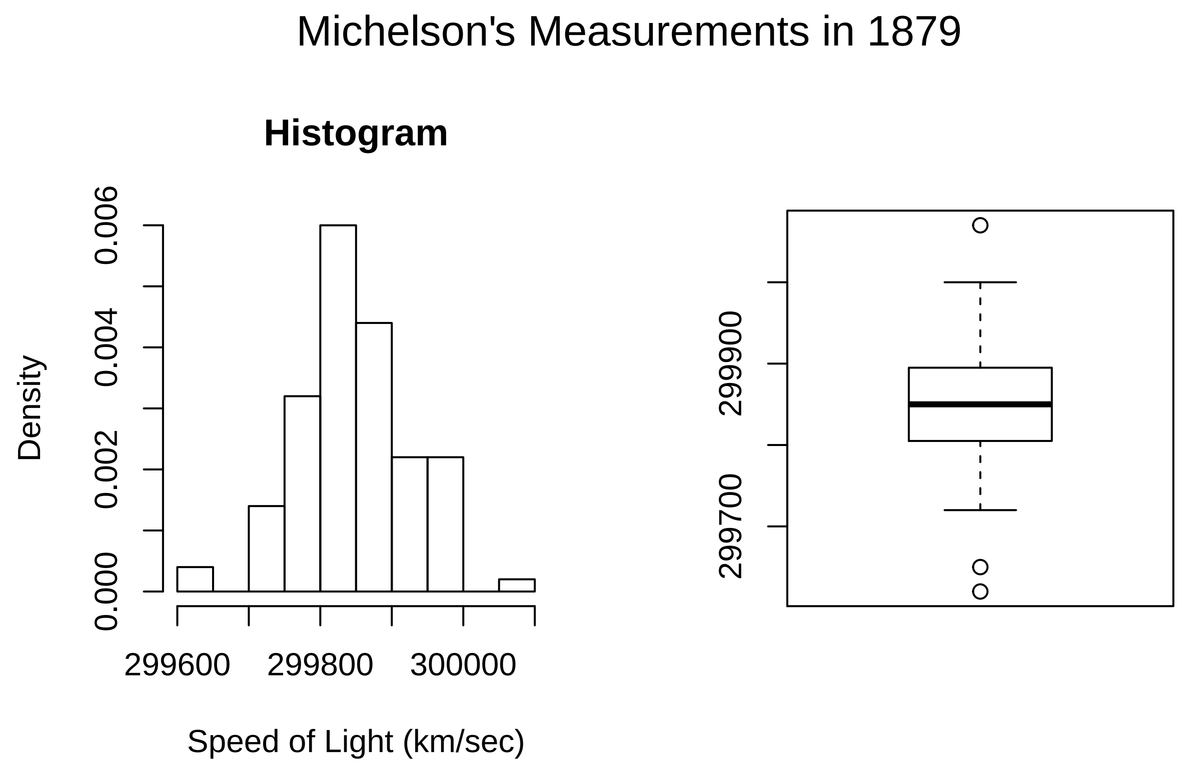
This exercise inspects the measurements of the speed of light done by two scientists in three different times. Histograms and box plots of these measurements can be seen below. From the box plot, it can be seen that the mean of all measurements are around the speed of light, however, the measurements also seem to have some outliers. Histograms of the measurements suggest that these measurements are probably from the same distribution. To figure out confidence intervals, we used median since it is more reliable against the outliers that are present in all of the measurements.

```
 # Newcomb's measurements made in 1882 on three days
lightMicro = (light / 1000) + 24.8 # Microseconds to travel 7442 kilometers
light = 7442 / (lightMicro * 10^(-3))

par(mfrow=c(1,2), oma = c(0, 0, 3, 0)) # Two graphs side by side
```

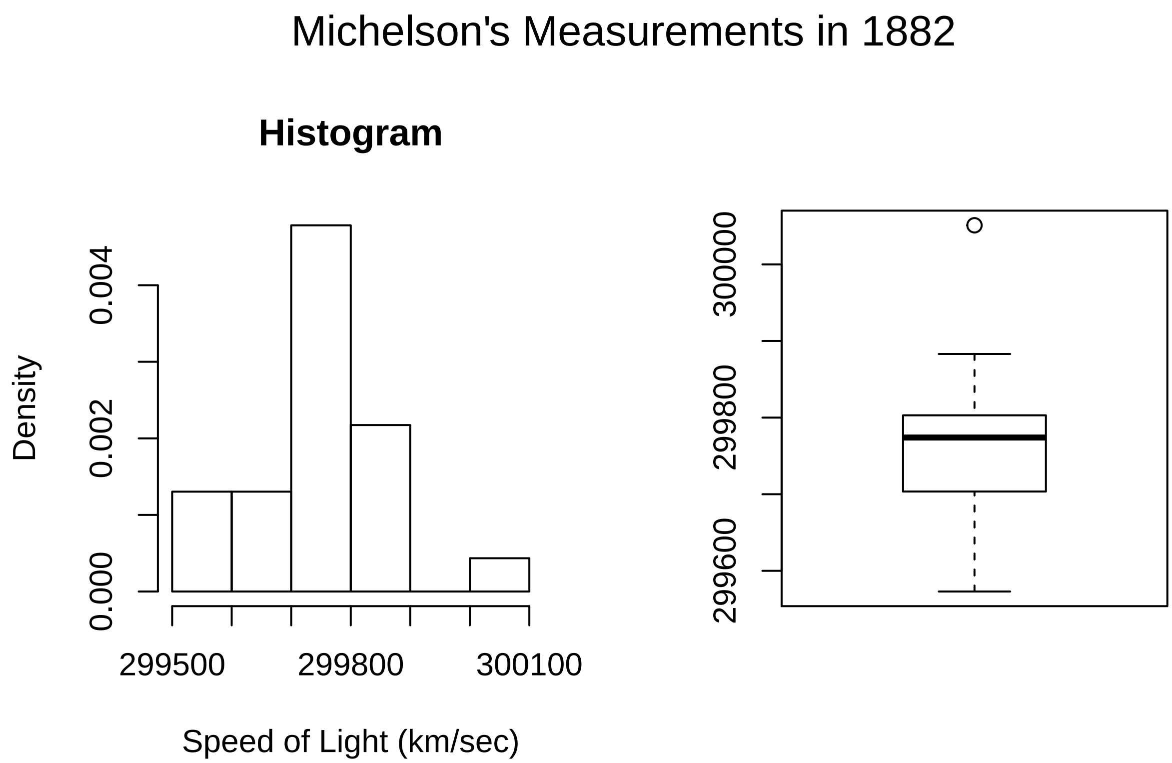
```
hist(light$V1, freq=FALSE, main = "Histogram", xlab="Speed of Light (km/sec)")
boxplot(light$V1)

mtext("Newcomb's Measurements in 1882", outer = TRUE, cex = 1.3)
```




```
par(mfrow=c(1,2), oma = c(0, 0, 3, 0)) # Two graphs side by side
hist(light1879Stacked$values, freq=FALSE,
     main = "Histogram", xlab="Speed of Light (km/sec)")
boxplot(light1879Stacked$values)
mtext("Michelson's Measurements in 1879", outer = TRUE, cex = 1.3)
```




```r
par(mfrow=c(1,2), oma = c(0, 0, 3, 0)) # Two graphs side by side
hist(light1882Stacked$values, freq=FALSE
     , main = "Histogram", xlab="Speed of Light (km/sec)")
boxplot(light1882Stacked$values)
mtext("Michelson's Measurements in 1882", outer = TRUE, cex = 1.3)
```



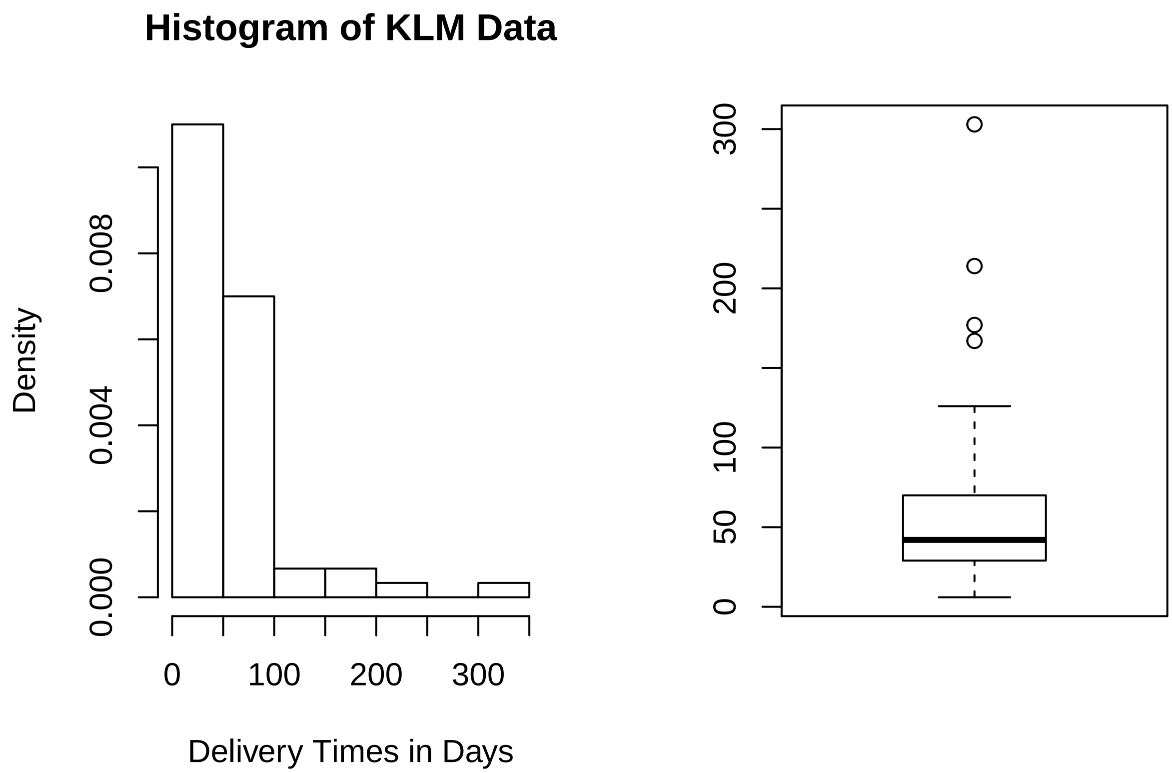
The exact value of the speed of light in vacuum denoted by c is 299,792,458 metres/second. Confidence intervals for the given three different data sets in kilometre/second can be seen below. Given the confidence intervals and the exact value of the speed of light, it can be said that it is consistent with the measurements of Michelson's measurements done in 1882.

```
## [1] "Method              97.5%    2.5%"

## [1] "Newcomb's   1882    299742.226    299772.409"

## [1] "Michelson's 1879    299830.000    299860.000"

## [1] "Michelson's 1882    299752.000    299825.000"
```

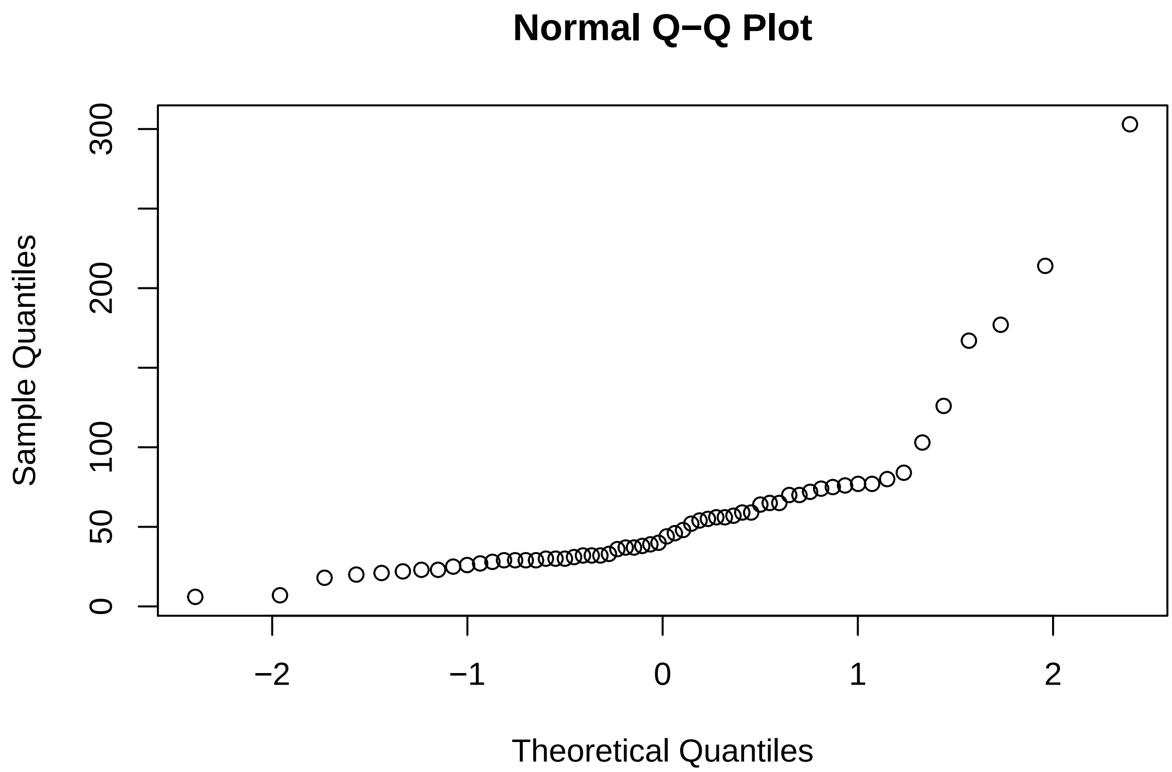
Exercise 3

For testing median values we used sign-test since from the graphs, the sample does not seem to have a normal distribution or from a symmetrical population. Histogram and QQ-Plot of the data set can be seen below. For the test, hypothesis $H_0 : \mu \leq 31$ is tested against the alternative hypothesis, $H_1 : \mu > 31$.

```r
klmData = scan("./data/klm.txt")

par(mfrow=c(1,2))
# This doen't look like it is from normal distribution?
hist(klmData, freq=FALSE, main="Histogram of KLM Data", xlab = "Delivery Times in Days")
boxplot(klmData)
```




```r
par(mfrow=c(1,1))
qqnorm(klmData)
```



We expect median to divide the data set into two equal parts so that when a random sample is chosen, the probability of it being smaller or greater than the median should be equal to tossing a coin.

```
# H_0 median duration is <= 31 days
testMedian = 31

klmMedian = median(klmData)

sumOut = sum(klmData <= testMedian) # Get values smaller than the test value
klmbinom = binom.test(sumOut, length(klmData), p=0.5, alternative = "less")
```

```
## [1] "Testing H0 lambda <= median"

## [1] "p-value is 0.006745"

## [1] "confidence interval is 0.000 - 0.446"

## [1] "probability of success 0.333"
```

It can be seen that the $H_0 : \mu \leq 31$ is rejected with the p-value of 0.0067446 since $H_0 : \mu \leq 31$ is not greater than the 50% of the sample. It is located in the first 33% of the data, therefore $H_1 : \mu > 31$ is accepted.

For the seconds part of this exercise, we filtered the delivery dates which are overdue and used binomial test with the probability of 10% since we are looking whether the deliveries are mostly made on time by Boeing without violating the criteria that is demanded by KLM.

```r
lateDays = sum(klmData > 72) # Days greater than max delivery days of 72
lateBinom = binom.test(lateDays, length(klmData), p=0.1, alternative = "greater")
```

```
## [1] "Testing H0 deliveryTime < 72 days"

## [1] "p-value is 0.006"

## [1] "confidence interval is 0.133 - 1.000"

## [1] "probability of success 0.217"
```

From the output of the test, it can be seen that $H_0 : d \leq 10\%$ is rejected with the p-value $= 0.005681$ and the alternative hypothesis $H_1 : d > 10\%$ is accepted. This yiels that Boeing is failing to meet the criteria by delivering more than 10% of the parts late.

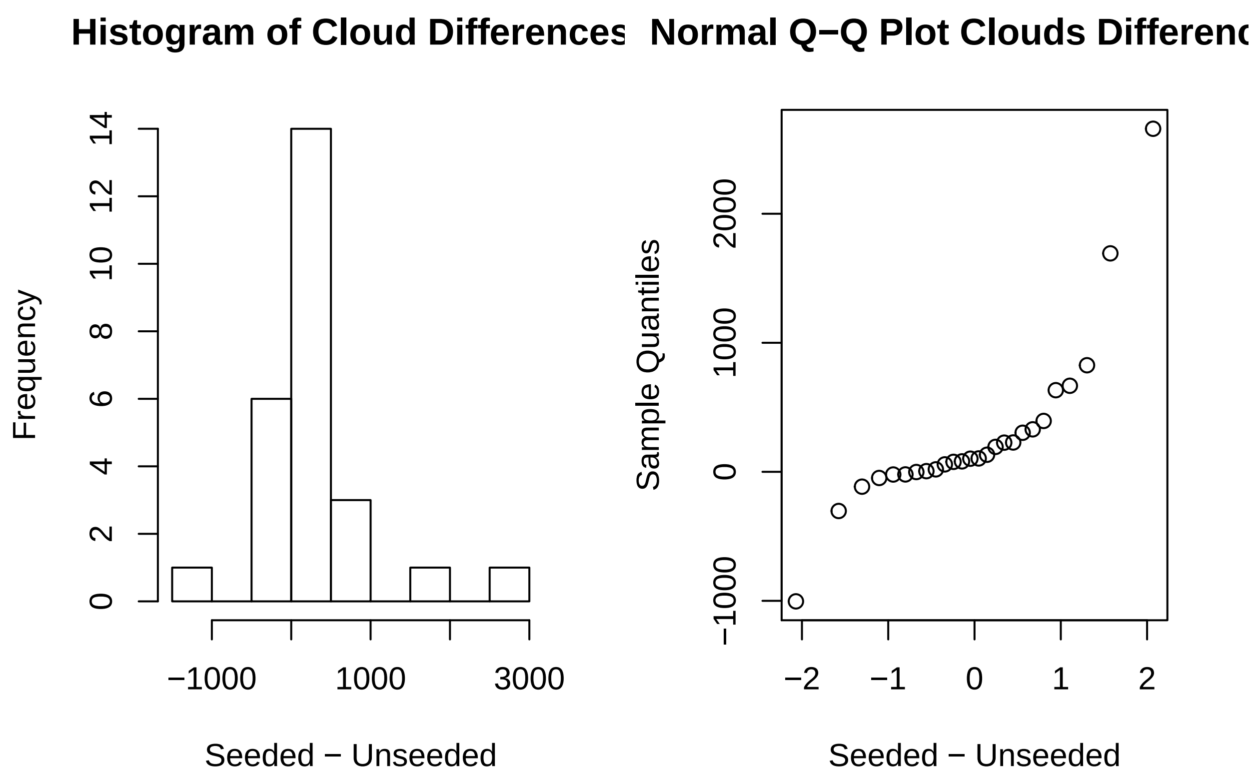
Exercise 4

In this exercise, effects of silver nitrate to the clouds on rainfall is investigated. There are two data sets with 26 observations each. In the first section, t-test, Mann-Whitnet test, and Kolmogorov-Smirnov test are used. In section two, same tests applied to the square root of the data. In the last section, same tests applied to the square root of square rooted data.

Section 1

Test reults of the original data can be seen below. From this output, we fail to reject H_0 by using t-test becuase of its p-value being greater than 5%. Unfortunately, from the histogram and the QQ-Plot, we suspect that the distributions are not normal, therefore using t-test is meaningless. In the Mann-Whitney test and Kolmogorov - Smirnov tests, we rejected H_0 with the p-value$= 0.013$ concluding that the distributions of the samples are different.

```r
par(mfrow=c(1,2))
hist(clouds$seeded - clouds$unseeded, main = "Histogram of Cloud Differences"
     , xlab = "Seeded - Unseeded")
qqnorm(clouds$seeded - clouds$unseeded, main = "Normal Q-Q Plot Clouds Differences"
       , xlab = "Seeded - Unseeded")
```




```r
# T - Test

# Differences of the data samples does not seem to be normal
# However, histogram of these differences is seem to be?
# This is probably not paired since QQ Normal Plot suggests that
# the distribution is not normal
t1Test = t.test(clouds$seeded, clouds$unseeded)

# Mann - Whitney Test

t1Wilcox = wilcox.test(clouds$seeded, clouds$unseeded)
```

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

```r
# Kolmogorov - Smirnov Test

t1Ks = ks.test(clouds$seeded, clouds$unseeded)
```

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

```
## [1] "T-Test - Testing H0 Both samples come from a normal population"
```

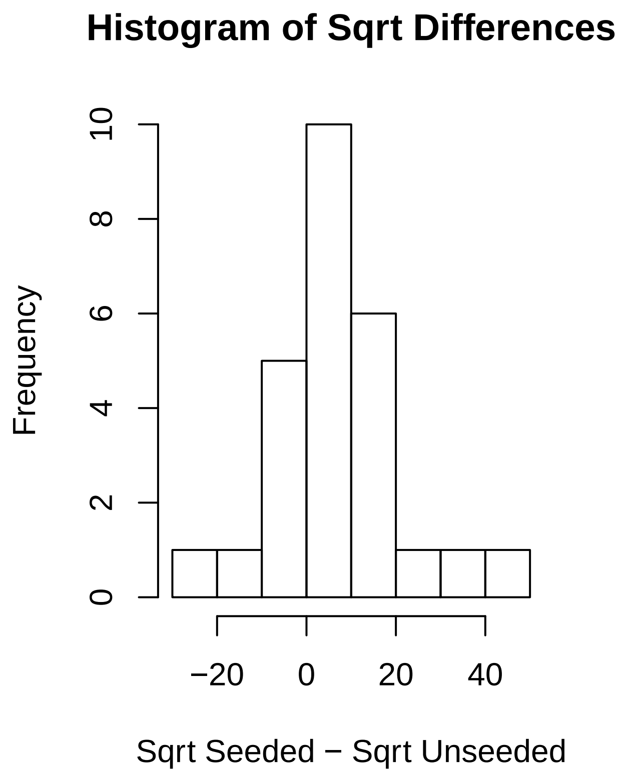
```
## [1] "p-value is 0.054"
```

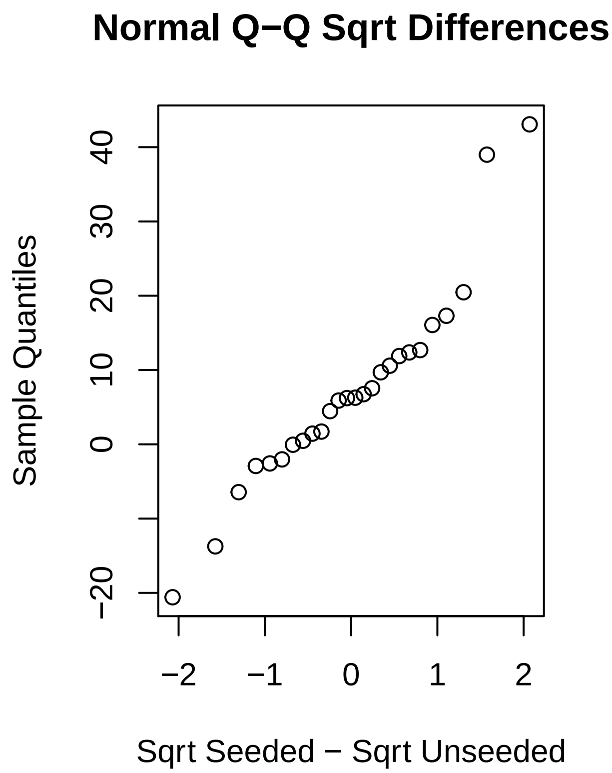
```
## [1] "confidence interval is -4.740 - 559.586"

## [1] "probability of success 441.985" "probability of success 164.562"

## [1] "Mann - Whitney Test - Testing H0 Both samples are from the same population"

## [1] "p-value is 0.014"

## [1] "W value is 473.000"

## [1] "Kolmogorov - Smirnov Test - Testing H0 Both samples are from the same population"

## [1] "p-value is 0.019"

## [1] "D value is 0.423"
```

Section 2

This time we observe from the histogram and the QQ-Plot that the samples are from a normal distribution. This means that we can use *t*-test this time. From the test results, as before, it can be concluded that the distributions differ significantly for the seeded and unseeded samples.

```
par(mfrow=c(1,2))
hist(sqrtClouds$seeded - sqrtClouds$unseeded, main = "Histogram of Sqrt Differences"
     , xlab = "Sqrt Seeded - Sqrt Unseeded")
qqnorm(sqrtClouds$seeded - sqrtClouds$unseeded, main = "Normal Q-Q Sqrt Differences"
       , xlab = "Sqrt Seeded - Sqrt Unseeded")
```






```r
# T - Test

t2test = t.test(sqrtClouds$seeded, sqrtClouds$unseeded)

# Mann - Whitney Test

t2wilcox = wilcox.test(sqrtClouds$seeded, sqrtClouds$unseeded)
```

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

```r
# Kolmogorov - Smirnov Test

t2ls = ks.test(sqrtClouds$seeded, sqrtClouds$unseeded)
```

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

```
## [1] "T-Test - Testing H0 Both samples come from a normal population"
```

```
## [1] "p-value is 0.020"
```

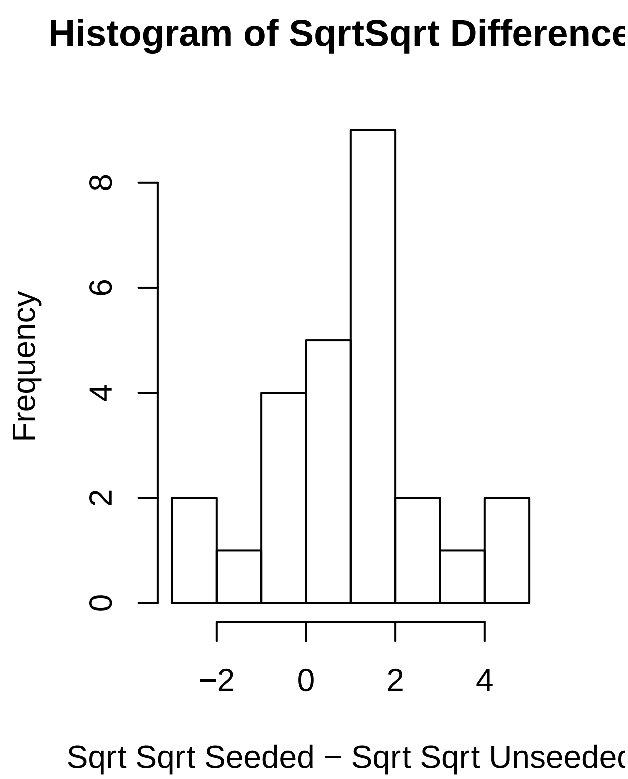
```
## [1] "confidence interval is 1.202 - 13.071"
```

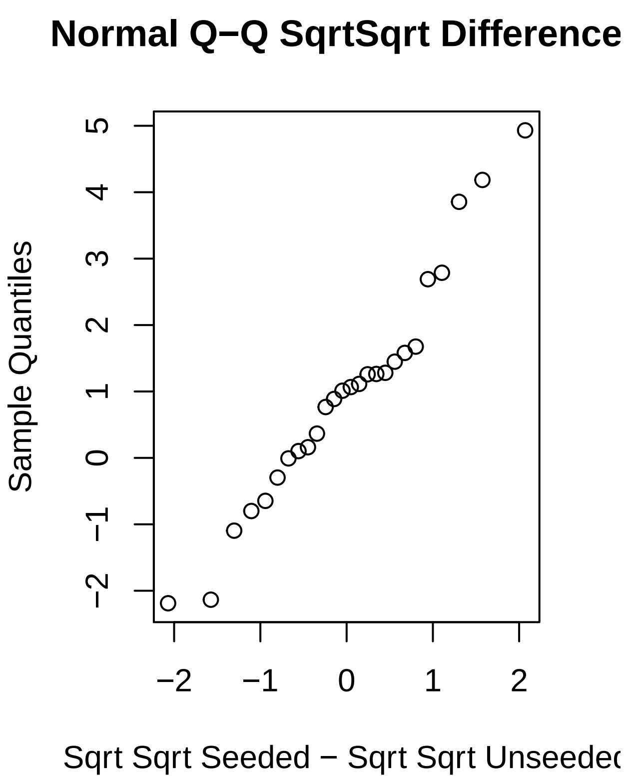
```
## [1] "probability of success 17.068" "probability of success 9.931"

## [1] "Mann - Whitney Test - Testing H0 Both samples are from the same population"

## [1] "p-value is 0.014"

## [1] "W value is 473.000"

## [1] "Kolmogorov - Smirnov Test - Testing H0 Both samples are from the same population"

## [1] "p-value is 0.019"

## [1] "D value is 0.423"
```

Section 3

As in previous section, the histogram and the QQ-Plot of the square root of square rooted sample can be assumed that it is from a normal distribution. All of the tests conclude that the distributions of both samples differ significantly again. The results of all three tests can be seen below.

```r
par(mfrow=c(1,2))
hist(sqrtSqrtClouds$seeded - sqrtSqrtClouds$unseeded, main = "Histogram of SqrtSqrt Differences"
     , xlab = "Sqrt Sqrt Seeded - Sqrt Sqrt Unseeded")
qqnorm(sqrtSqrtClouds$seeded - sqrtSqrtClouds$unseeded, main = "Normal Q-Q SqrtSqrt Differences"
       , xlab = "Sqrt Sqrt Seeded - Sqrt Sqrt Unseeded")
```



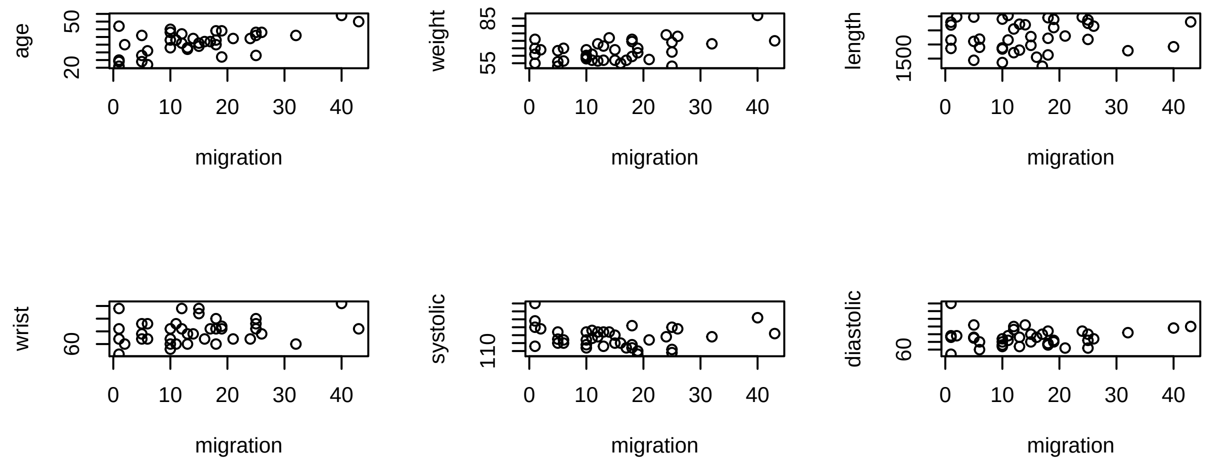


```
## Warning in wilcox.test.default(sqrtSqrtClouds$seeded,
## sqrtSqrtClouds$unseeded): cannot compute exact p-value with ties

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

## [1] "T-Test - Testing H0 Both samples come from a normal population"

## [1] "p-value is 0.012"

## [1] "confidence interval is 0.220 - 1.724"

## [1] "probability of success 3.879" "probability of success 2.907"

## [1] "Mann - Whitney Test - Testing H0 Both samples are from the same population"

## [1] "p-value is 0.014"

## [1] "W value is 473.000"

## [1] "Kolmogorov - Smirnov Test - Testing H0 Both samples are from the same population"

## [1] "p-value is 0.019"

## [1] "D value is 0.423"
```

Exercise 5

Section 1

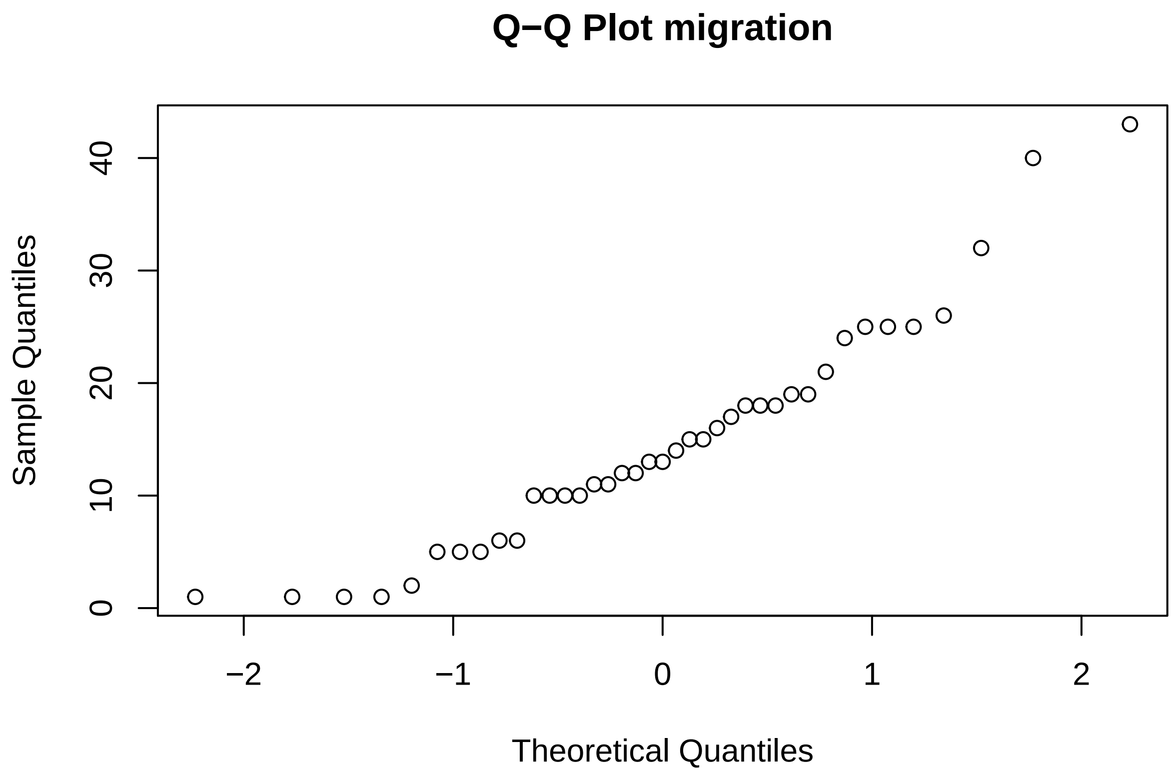
```r
par(mfrow=c(3,3))
plot(age~migration, peruvians)
plot(weight~migration, peruvians)
plot(length~migration, peruvians)
plot(wrist~migration, peruvians)
plot(systolic~migration, peruvians)
plot(diastolic~migration, peruvians)
```



From the plots, there seems to be a dependence between age, and weight to migration years.Apart from this none of the other variables seems to display a significant correlation to migration.

Section 2

```
par(mfrow=c(1,1))
# Checking normality for migration sample
qqnorm(peruvians$migration,main="Q-Q Plot migration")
```




```
#Normality is not evident for migration sample, hence we use Spearman's correlation test to check for d

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

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

```
##
##  Spearman's rank correlation rho
##
## data:  peruvians$age and peruvians$migration
## S = 5176.6, p-value = 0.002189
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4760575
```

```
# Moderate correlation observed

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

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

```
##
##  Spearman's rank correlation rho
##
## data:  peruvians$weight and peruvians$migration
## S = 6415.1, p-value = 0.02861
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3506956
```

```r
# Moderate correlation observed

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

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

```
##
##  Spearman's rank correlation rho
##
## data:  peruvians$length and peruvians$migration
## S = 9044.3, p-value = 0.6087
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## 0.08458432
```

```r
# Insignificant correlation

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

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

```
##
##  Spearman's rank correlation rho
##
## data:  peruvians$wrist and peruvians$migration
## S = 7712.8, p-value = 0.1797
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2193498
```

```r
# Weak correlation observed

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

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

```
##
##  Spearman's rank correlation rho
##
## data:  peruvians$systolic and peruvians$migration
## S = 11544, p-value = 0.3054
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.1684286
```

```r
# Weak but inverse correlation observed

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

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

```
##
##  Spearman's rank correlation rho
##
## data:  peruvians$diastolic and peruvians$migration
## S = 9137.6, p-value = 0.6494
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## 0.07514098
```

```r
# Insignificant correlation observed
```

Both age and weight seems to show moderate correlation to migration. Other variables, display either insignificant or weak correlation.

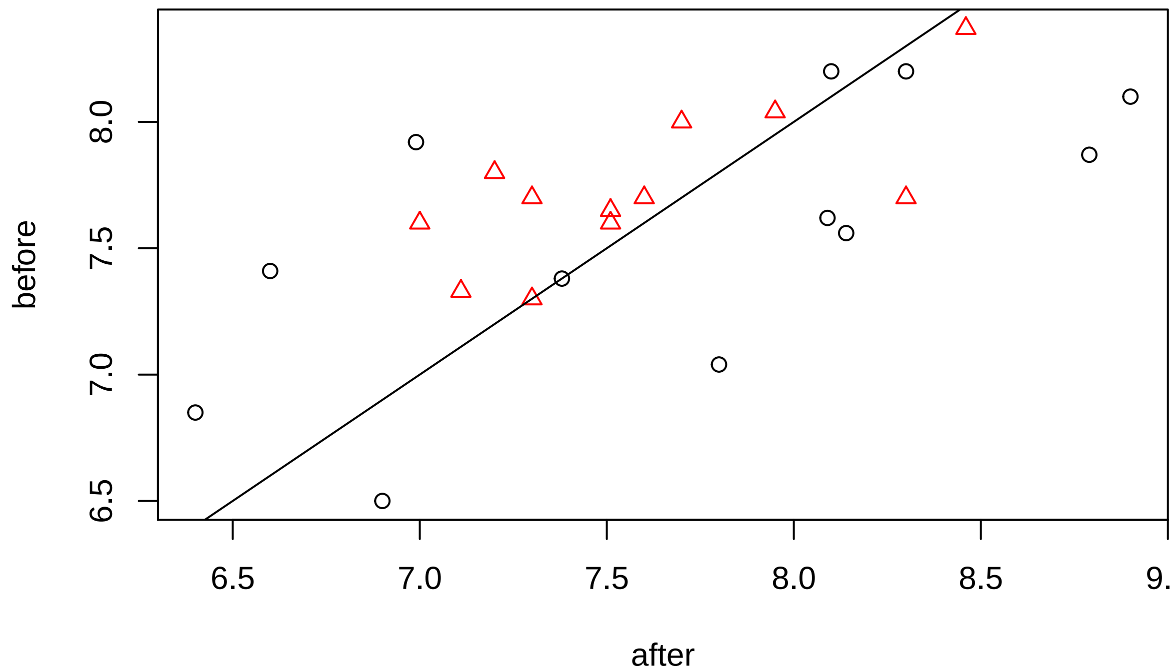
Exercise 6

Section 1

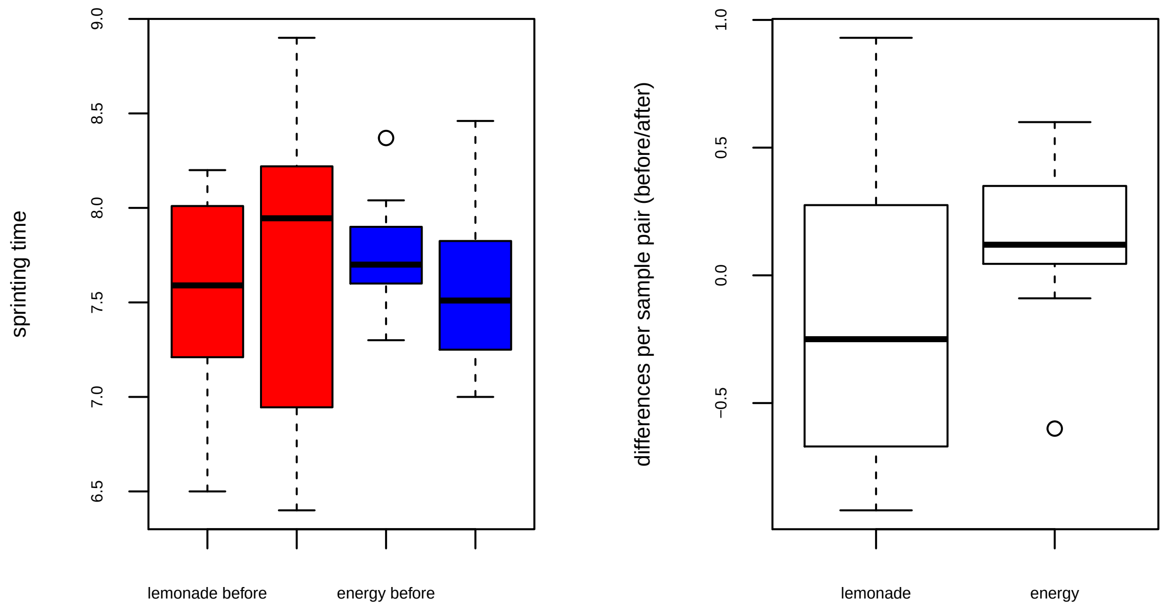
```r
df <- read.table("./data/run.txt")
df$cat <- c(rep(1,12), rep(2,12))

print(summary(df))
```

```
##      before          after           drink         cat     
##  Min.   :6.500   Min.   :6.400   energy:12   Min.   :1.0  
##  1st Qu.:7.402   1st Qu.:7.178   lemo  :12   1st Qu.:1.0  
##  Median :7.675   Median :7.555               Median :1.5  
##  Mean   :7.643   Mean   :7.639               Mean   :1.5  
##  3rd Qu.:7.940   3rd Qu.:8.110               3rd Qu.:2.0  
##  Max.   :8.370   Max.   :8.900               Max.   :2.0  
```

```r
par(mfrow=c(1,1)); plot(before~after, pch=cat, col=cat, data=df); abline(0,1)
```




```r
par(mfrow=c(1,2))
boxplot(df[1:12,1],df[1:12,2], df[13:24,1], df[13:24,2],
        col=c('red','red','blue','blue'),
        names=c("lemonade before","lemonade after","energy before","energy after"), cex.axis = 0.5,
        ylab = "sprinting time", cex.lab = 0.7)
boxplot(df[1:12,1]-df[1:12,2],df[13:24,1]-df[13:24,2],
        names = c("lemonade", "energy"), cex.axis = 0.5,
        ylab = "differences per sample pair (before/after)", cex.lab = 0.7)
```



Section 2

```
t.test(df[1:12,1],df[1:12,2],paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  df[1:12, 1] and df[1:12, 2]
## 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
```

```
t.test(df[13:24,1], df[13:24,2], paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  df[13:24, 1] and df[13:24, 2]
## 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
```

For the lemonade group, there is no cause to reject the null hypothesis and assume that the means are different. The same holds for the energy group, although the p-value is lower for this group (p-value = 0.126).

Section 3

```r
df$differences <- df$before - df$after
t.test(df[1:12,5],df[13:24,5])
```

```
##
##  Welch Two Sample t-test
##
## data:  df[1:12, 5] and df[13:24, 5]
## 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
```

Since the p-value is 0.159, there is no cause to reject the null hypothesis and assume that the means are different.

Section 4

Since the participants were asked to run two stretches within a relatively small timespan, the first measurement may be affecting the second (learning effect). there could be additional factors affecting performance on the second run such as fatigue or muscle activation (i.e. 'getting warmed up').

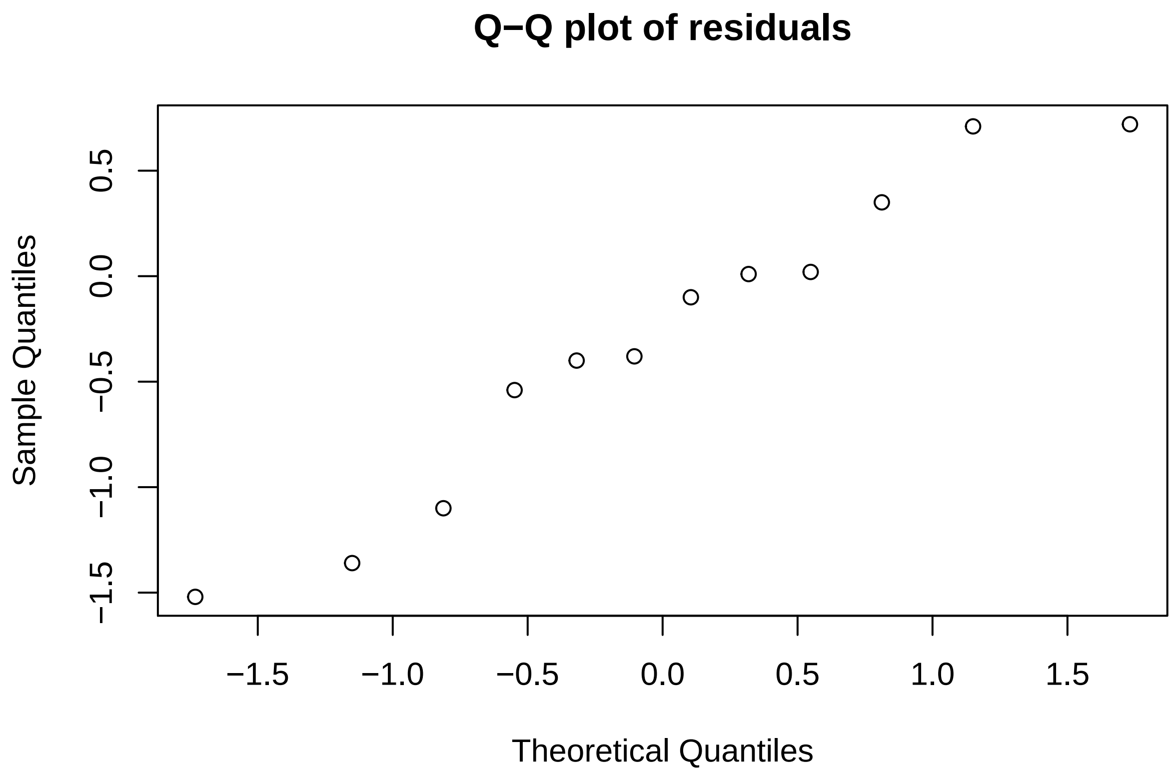
Section 5

Here, the samples are drawn from independent populations (different students) and one measurement does not affect the other, so there is no learning effect present.

Question 6

The samples must come from a normal population. Whether this condition is satisfied can be examined by investigating the normality of the residuals.

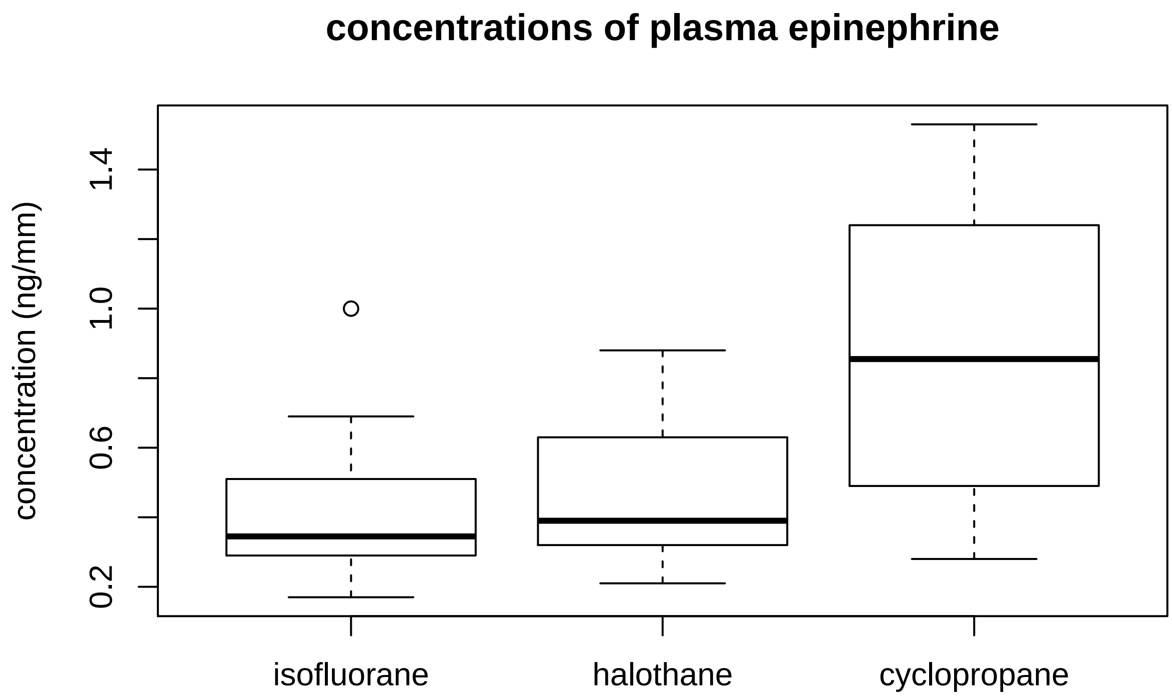
```r
residuals <- df[1:12,5] - df[13:24,5]
qqnorm(residuals, main='Q-Q plot of residuals')
```



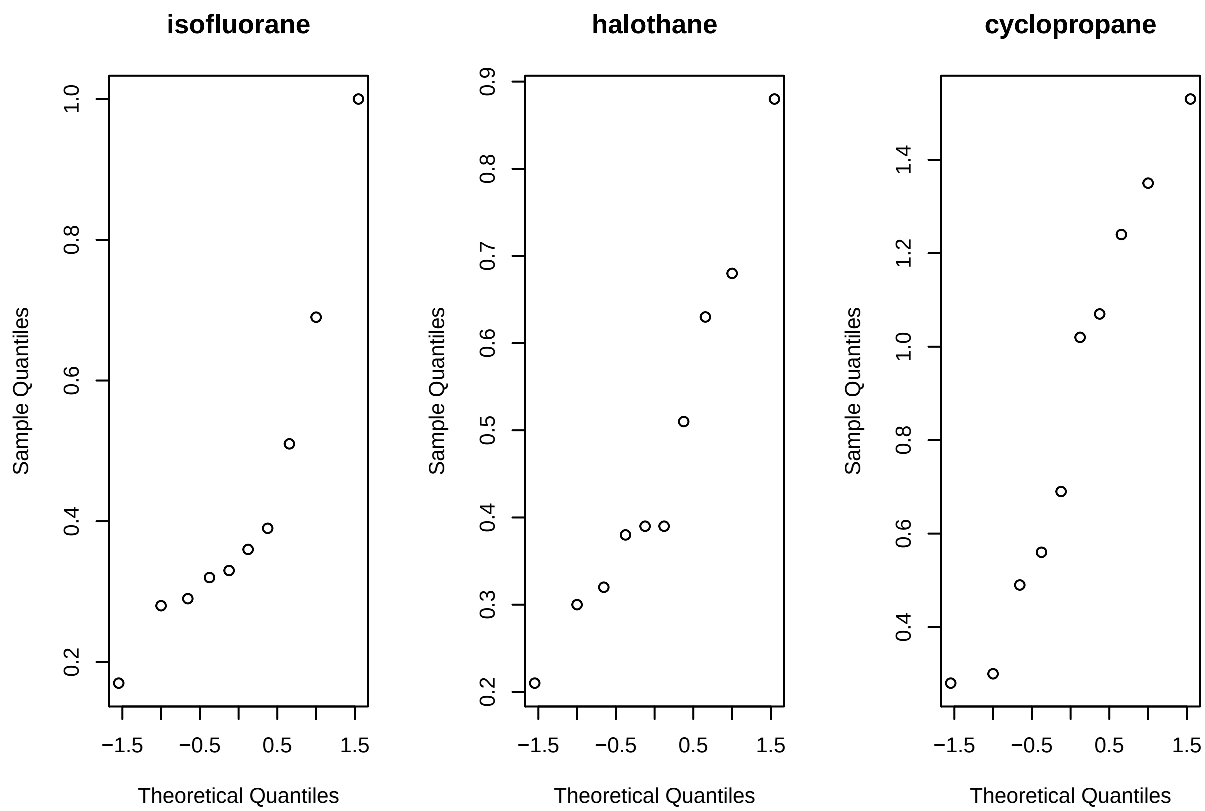
Exercise 7

Section 1

```
dogs <- read.table("./data/dogs.txt",stringsAsFactors = FALSE)
boxplot(as.numeric(dogs[2:11,1]), as.numeric(dogs[2:11,2]), as.numeric(dogs[2:11,3]),
        ylab = "concentration (ng/mm)",
        main = "concentrations of plasma epinephrine",
        names = c("isofluorane", "halothane", "cyclopropane"))
```




```
par(mfrow=c(1,3))
qqnorm(as.numeric(dogs[2:11,1]), main = dogs[1,1])
qqnorm(as.numeric(dogs[2:11,2]), main = dogs[1,2])
qqnorm(as.numeric(dogs[2:11,3]), main = dogs[1,3])
```



It is not reasonable to assume that the samples were taken from normal populations, since the plot for isofluorane appears skewed and could be nonnormal.

Section 2

```
dogsframe <- data.frame(concentration=as.numeric(as.matrix(dogs[2:11,])),
                        substance=factor(c(rep(dogs[1,1],10),rep(dogs[1,2],10),rep(dogs[1,3],10))))
dogsaov=lm(concentration~substance,data=dogsframe)
anova(dogsaov)
```

```
## Analysis of Variance Table
##
## Response: concentration
##           Df Sum Sq Mean Sq F value Pr(>F)
## substance  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 = concentration ~ substance, 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)            0.8530     0.1005   8.491 4.19e-09 ***
## substancehalothane    -0.3840     0.1421  -2.703   0.0117 *
## substanceisofluorane  -0.4190     0.1421  -2.949   0.0065 **
## ---
## 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
```

The p-value is low (0.011), so the null hypothesis would be rejected.

The estimated concentrations are as follows:
isofluorane: 0.469 halothane: 0.434 cyclopropane: 0.853

Section 3

```
kruskal.test(dogsframe$concentration,dogsframe$substance)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  dogsframe$concentration and dogsframe$substance
## Kruskal-Wallis chi-squared = 5.6442, df = 2, p-value = 0.05948
```

The p-value is 0.5948, which is larger than 0.05. The null hypothesis would not be rejected. The difference in results could indicate that the assumptions for a parametric one-way ANOVA test are not met. The populations tested here may be nonnormal, as seen in the Q-Q plots in Section 1, and the sample size is small (n=10).