assignment1

nihat uzunalioglu - 2660298, emiel kempen - 2640580, saurabh jain - 2666959

2/13/2020

Exercise 1

a) Set n=m=30, mu=180 and sd=5. Calculate now the power of the t-test for every value of nu in the grid seq(175,185,by=0.25).

```
n = m = 30; mu = 180; sd = 5
# Assign nu
nu = seq(175, 185, by=0.25)
# Calculation of the power
secA <- powersFromDifferentNuVals(n, m, mu, nu, sd)
```

b) Set n=m=100, mu=180 and sd=5. Repeat the preceding exercise. Add the plot to the preceding plot.

```
n=m=100; mu=180; sd=5
# Repeat the preceding exercise.
# Calculation of the power
secB <- powersFromDifferentNuVals(n, m, mu, nu, sd)
```

c) Set n=m=30, mu=180 and sd=15. Repeat the preceding exercise.

```
n=m=30; mu=180; sd=15
# Repeat the preceding exercise.
# Calculation of the power
secC <- powersFromDifferentNuVals(n, m, mu, nu, sd)
```

d) Explain your findings.

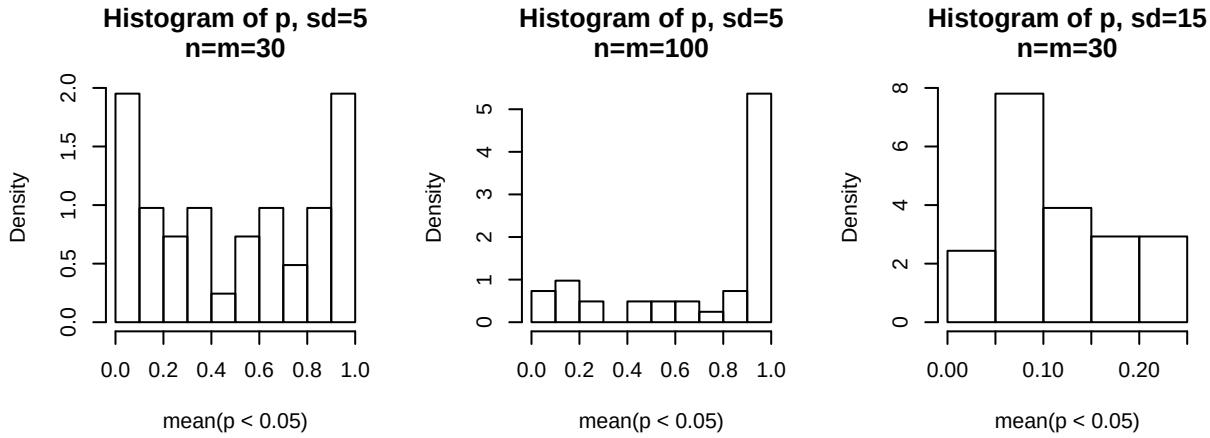
```
# Set aligning for 3 different histograms
par(mfrow=c(1,3))
# X axis label
xlab <- "mean(p < 0.05)"

# Plot histograms
hist(secA, freq = FALSE,
     main = "Histogram of p, sd=5\nn=m=30",
     xlab = xlab)

hist(secB, freq = FALSE,
     main = "Histogram of p, sd=5\nn=m=100",
```

```
     xlab = xlab)

hist(secC, freq = FALSE,
     main = "Histogram of p, sd=15\nn=m=30",
     xlab = xlab)
```



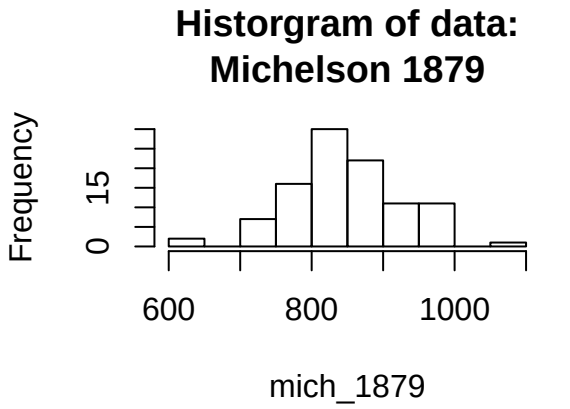
- Number of observations play a significant role in terms of the power of the test. While number of observations in x and y samples is 30, we came to the conclusion of H0, the null hypothesis, can be accepted even though we see an accumulation between 5% and 10% while standard deviation is equal to 15.

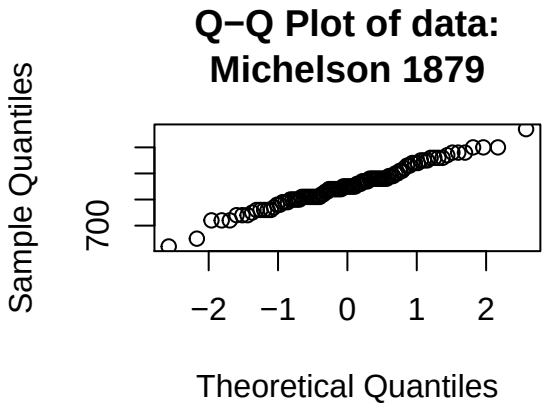
Exercise 2

```
# Read files, need to use filling for light1882 as the last row
# contains less values than the others
mich_1879 = data.matrix(read.table("light1879.txt"))
mich_1882 = data.matrix(read.table("light1882.txt", fill=TRUE))
newcomb = data.matrix(read.table("light.txt"))
```

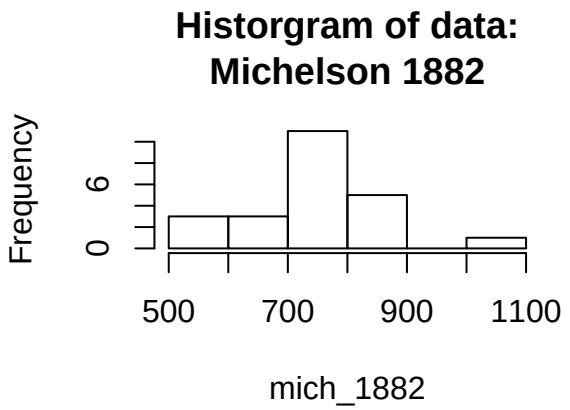
a) Investigate the normality for all three data sets.

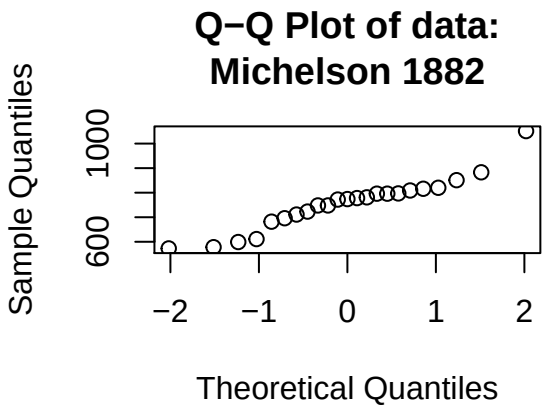
```
# We determine normality by plotting a histogram and QQ Plot for every dataset.
par(mfrow=c(1,2))
hist(mich_1879, main="Historgram of data:\nMichelson 1879")
qqnorm(mich_1879, main="Q-Q Plot of data:\nMichelson 1879")
```



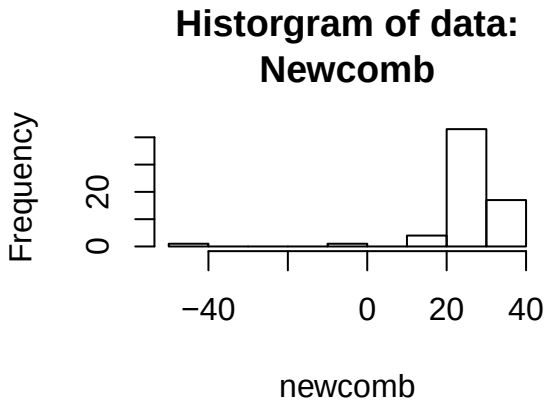


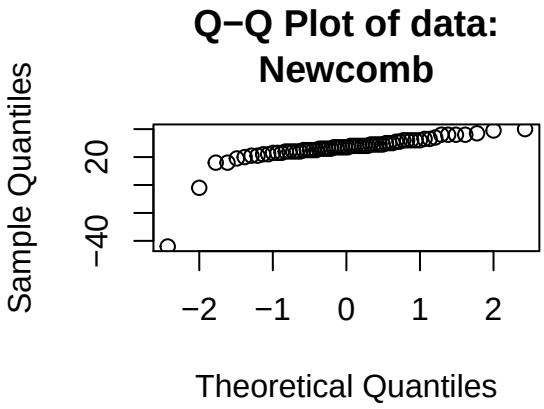

```
par(mfrow=c(1,2))
hist(mich_1882, main="Historgram of data:\nMichelson 1882")
qqnorm(mich_1882, main="Q-Q Plot of data:\nMichelson 1882")
```






```
par(mfrow=c(1,2))
hist(newcomb, main="Historgram of data:\nNewcomb")
qqnorm(newcomb, main="Q-Q Plot of data:\nNewcomb")
```





- The plots show a normal distribution for both of Michelson's measurements, as the histogram takes a bell-shaped curve and the QQ-plot follows a straight line. In Newcomb's measurements, these characteristics of a normal distribution do not show.

b) Determine confidence intervals for the speed of light in km/sec for all three data sets (use population means). Comment on the intervals found.

- First, using data from the exercise, the measurements have to be transformed to km/sec following the calculation below:

```
mich_1879_km_s = mich_1879 + 299000
mich_1882_km_s = mich_1882 + 299000
newcomb_km_s = 7.442/(((newcomb/1000)+24.8)/1000000)
```

- Then we calculate the confidence intervals for the speed of light in km/sec for all three data sets using the One Sample t-test:

```
t.test(mich_1879_km_s)[[4]]
```

```
## [1] 299836.7 299868.1
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(mich_1882_km_s)[[4]]
```

```
## [1] 299709.9 299802.5
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(newcomb_km_s)[[4]]
```

```
## [1] 299731.9 299795.8
## attr(,"conf.level")
## [1] 0.95
```

- From the confidence intervals, we learn that the 1879 measurements from Michelsen are the most consistent, as the upper and lower bounds value differ least from each other when compared to the confidence intervals of the other measurements. Newcomb's measurements come in second, and Michelsen's 1882 measurements deviate the most.

c) Find on the internet the currently most accurate value for the speed of light. Is it consistent with the measurements of Michelson and Newcomb?

- From Wikipedia, we learn that the currently most accurate value for the speed of light equals $299792.458 km/s$.

```
speed_of_light = 299792.458
```

- We then check whether this value falls within the upper and lower confidence levels of the three seperate datasets, to see whether it is consistent with the measurements of Michelson and Newcomb:

```
# Extracting upper and lower confidence bounds
# by indexing the One Sample t-test results
t.test(mich_1879_km_s)[[4]][1]<=speed_of_light & speed_of_light<=t.test(mich_1879_km_s)[[4]][2]
```

```
## [1] FALSE
```

```
t.test(mich_1882_km_s)[[4]][1]<=speed_of_light & speed_of_light<=t.test(mich_1882_km_s)[[4]][2]
```

```
## [1] TRUE
```

```
t.test(newcomb_km_s)[[4]][1]<=speed_of_light &     speed_of_light<=t.test(newcomb_km_s)[[4]][2]
```

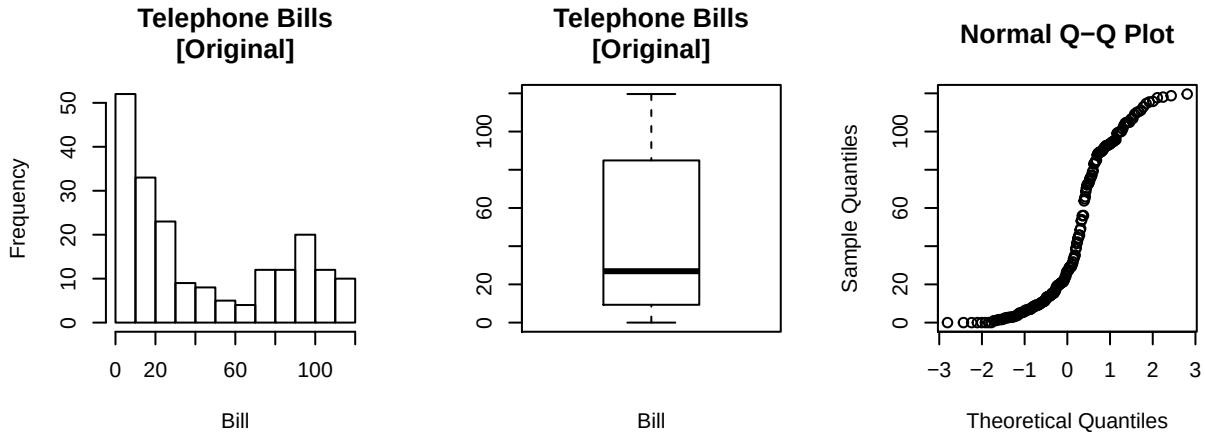
```
## [1] TRUE
```

- The currently most accurate value for the speed of light is thus consistent with the measurements of Newcomb and the 1882 measurements of Michelsen, but not his 1879 measurements.

Exercise 3

a) Make an appropriate plot of this data set.

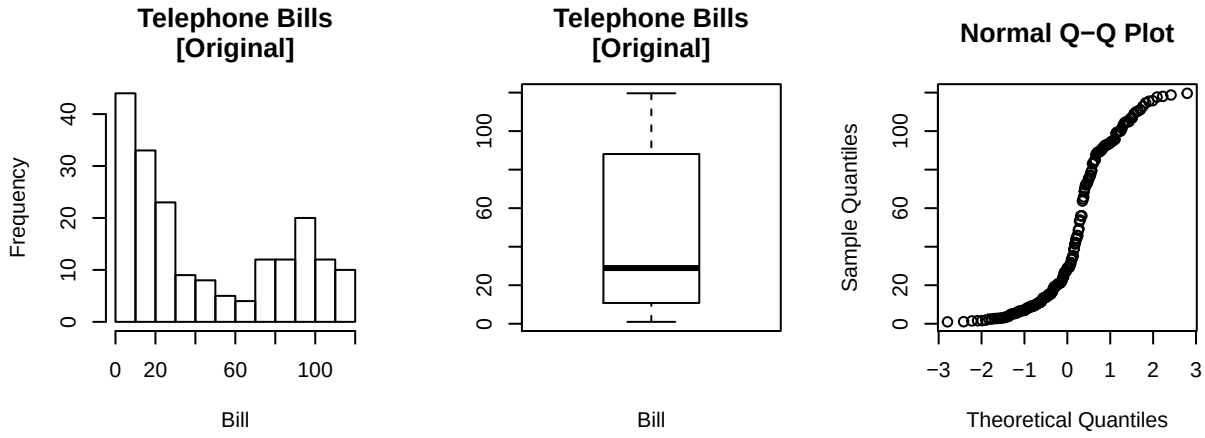
```
# Read telephone.txt
tel <- read.table(file = "telephone.txt", header = TRUE)
# Divide row 2 for histogram and boxplot
par(mfrow=c(1,3))
# Plotting telephone.txt
hist(x = tel$Bills,
     main = "Telephone Bills\n[Original]",
     xlab = "Bill")
boxplot(x = tel$Bills,
        main = "Telephone Bills\n[Original]",
        xlab = "Bill")
qqnorm(tel$Bills)
```



What marketing advice(s) would you give to the marketing manager? - There are two types of users occuring according to the tables, low paying and high paying. A campaign could be offered to the users for the middle class and as well as a motivation for the people who have low bills. This way, it will be well distributed along in each class and prefered for long time usage.

Are there any inconsistencies in the data? If so, try to fix these. - When we look at the data, we can see that there are zero paying bills in the data which does not show any insight for us and can be misleading for the following campaign strategies so they are extracted from the data and differences can be seen in the following charts.

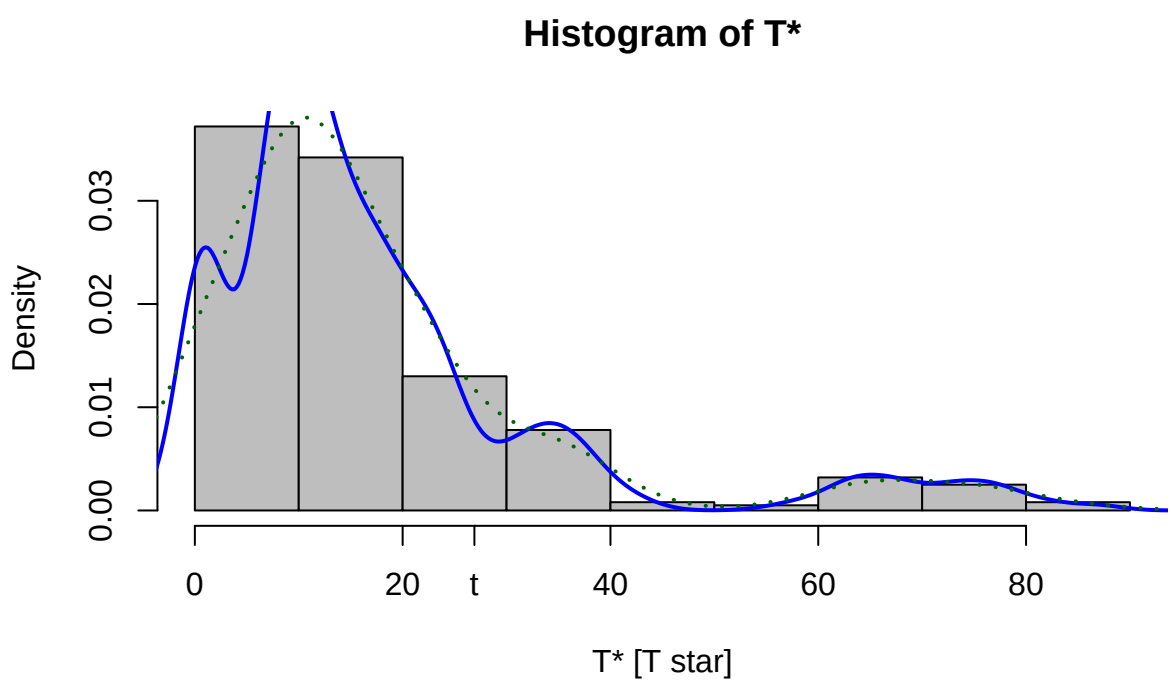
```
# Extract zero bills from the data
nonZeroBills <- tel[tel$Bills != 0, ]
# Divide row 2 for histogram and boxplot
par(mfrow=c(1,3))
# Plotting telephone.txt
hist(x = nonZeroBills,
     main = "Telephone Bills\n[Original]",
     xlab = "Bill")
boxplot(x = nonZeroBills,
        main = "Telephone Bills\n[Original]",
        xlab = "Bill")
qqnorm(nonZeroBills)
```



b) By using a bootstrap test with the test statistic T = median(X1,...,X200), test whether the data telephone.txt stems from the exponential distribution Exp(λ) with some λ from [0.01, 0.1].

```
## [1] "The median equals 26.905"
```

```
# Simulate B times
for (i in 1:B) {
  # Get sample from Exp(lambda)
  xstar = rexp(n, rate = sample(lamb, 1))
  # Receives tstar by the test statistics the median
  tstar[i] = median(xstar)
}
# Draw histogram
hist(tstar, prob=T, col = "grey",
     main = "Histogram of T*",
     xlab = "T* [T star]")
# Draw density curve of T
lines(density(tstar), col="blue", lwd=2) # add a density estimate with defaults
lines(density(tstar, adjust=2), lty="dotted", col="darkgreen", lwd=2)
axis(1, t, expression(paste("t") ) )
```




```
# P left
pl = sum(tstar < t) / B
# P right
pr = sum(tstar > t) / B
p = 2 * min(pl, pr)
print(paste("The p value is", p, "and H0 is not rejected"))
```

```
## [1] "The p value is 0.332 and H0 is not rejected"
```

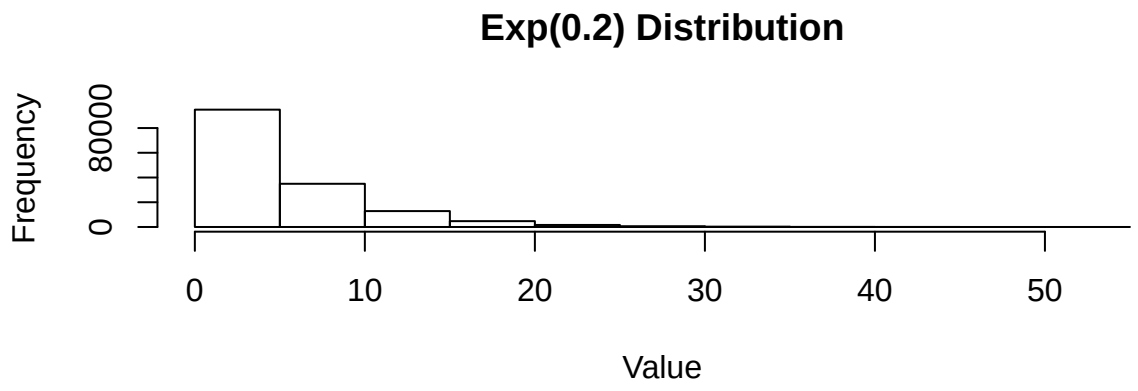
c) Construct a 95% bootstrap confidence interval for the population median of the sample.

```
confLvl <- t.test(tel$Bills)[[4]]
print(paste("%95 confidence interval for the sample: [",
            confLvl[1], ",", confLvl[2], "]"))
```

```
## [1] "%95 confidence interval for the sample: [ 38.1537266714263 , 49.0214733285737 ]"
```

d) Assuming $X_1, ..., X_N \sim Exp(\lambda)$ and using the central limit theorem for the sample mean, estimate λ and construct again a 95% confidence interval for the population median. Comment on your findings.

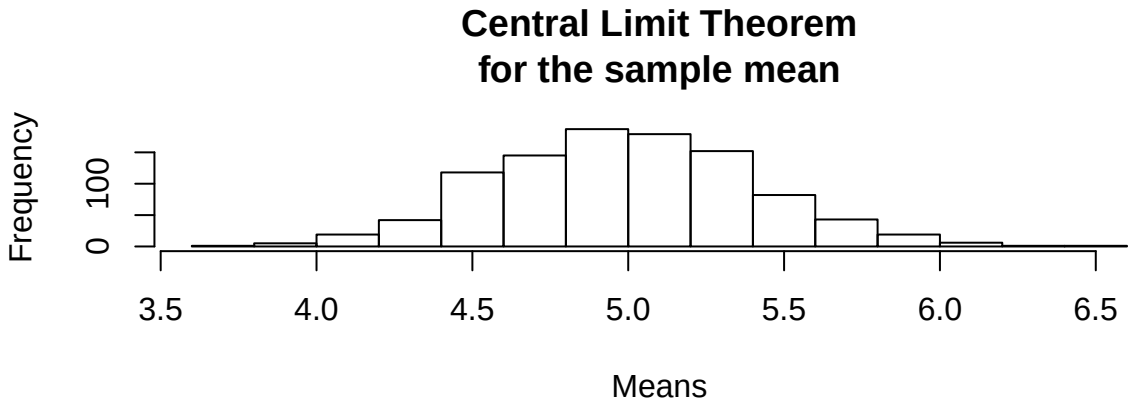
```
# Simulate
sim <- rexp(n*rows, lambda)
# Plot histogram
hist(sim, main = "Exp(0.2) Distribution",
     xlab = "Value")
```




```
# Create matrix
m <- matrix(sim, rows)

# Calculate sample means
sample.means <- rowMeans(m)

# Calculate mean and standard deviation
# of the simulation for CLT(Central Limit Theorem)
sm.avg <- mean(sample.means); sm.sd <- sd(sample.means)
# Plot average
hist(sample.means,
     main = "Central Limit Theorem\nfor the sample mean",
     xlab = "Means")
```




```
# Theoretical(Expected) standard deviation
sm.sd.clt <- sqrt(lambda/n)
# Average of the simulation, standard deviation
# and expected standard deviation
sm.avg; sm.sd; sm.sd.clt
```

```
## [1] 4.98489
```

```
## [1] 0.4116917
```

```
## [1] 0.03651484
```

```
# Calculation of %95 confidence level for the mean
confLvl <- t.test(sample.means)[[4]]
print(paste("%95 confidence interval for the sample: [",
            confLvl[1], ",", confLvl[2], "]"))
```

```
## [1] "%95 confidence interval for the sample: [ 4.95934303238425 , 5.0104378351341 ]"
```

- The figure on the left hand side shows $Exp(\lambda = 2)$ distribution and the figure on the right hand side shows the normal distribution after the application of Central Limit Theorem(CLT) for the sample mean. After doing 1000 simulations with samples have 150 sample size, Central Limit Theorem again proves that as the repetation goes infinity, a perfect normal distribution occurs.

```
# Estimated lambda
es_lambda = 1 / sm.avg
```

- The λ can be estimated by $\frac{1}{mu_{estimated}}$. Thus, this leads us $\lambda = 0.2006062$.

e) Using an appropriate test, test the null hypothesis that the median bill is bigger or equal to 40 euro against the alternative that the median bill is smaller than 40 euro.

- In the section A, as we plotted histogram, boxplot and QQ-Plot to be able to observe whether the data has normality or not, it is clear that the data does not have a normality within samples. For these kinds of non-normality situations we choose to apply either sign or Wilcoxon test on the dataset.

```r
# Get sum where bills are >= 40
cond <- sum(tel$Bills >= 40)
# Sign test
binom.test(cond, length(tel$Bills), alternative = "less")[3]
```

```
## $p.value
## [1] 0.009698472
```

```r
# Wilcoxon test
wilcox.test(tel$Bills, mu=40)[3]
```

```
## $p.value
## [1] 0.1692763
```

Next, design and perform a test to check whether the fraction of the bills less than 10 euro is at most 25%.

```r
# Sign test
cond <- sum(tel$Bills <= 9.999)
binom.test(cond, length(tel$Bills), alternative = "greater")[3]
```

```
## $p.value
## [1] 1
```

Exercise 4

a) Disregarding the type of drink, test whether the run times before drink and after are correlated.

```r
run <- read.table(file = "run.txt", header = TRUE)
res <- cor.test(run$before,run$after)
res[3]
```

```
## $p.value
## [1] 0.0007799659
```

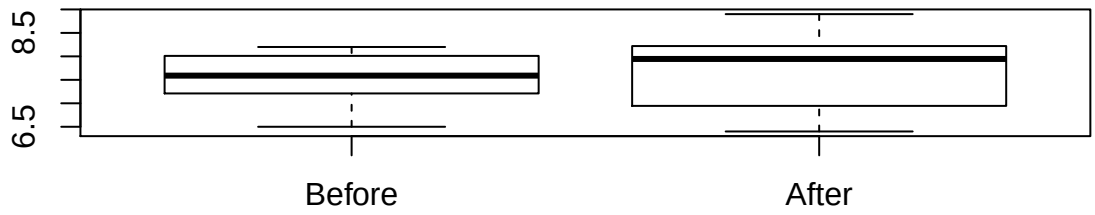
- p-value of the correlation test is 0.00077996588481048, which is less than the significance level $\alpha = 0.05$.

b) Test separately, for both the softdrink and the energy drink conditions, whether there is a difference in speed in the two running tasks.

```
## Loading required package: ggplot2
```

```
## Loading required package: magrittr
```

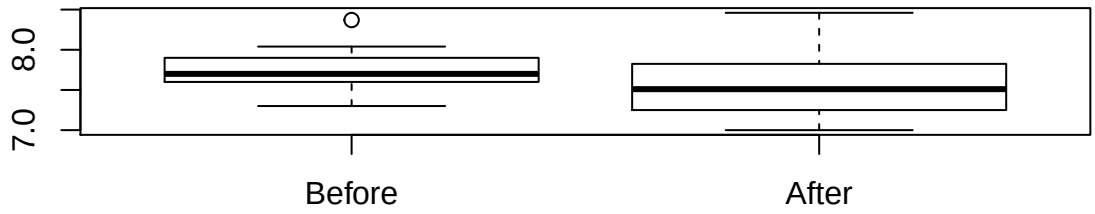
```r
softdrink <- subset(run, drink == "lemo")
boxplot(softdrink[,1],softdrink[,2],names=c("Before","After"))
```




```
cor.test(softdrink$before,softdrink$after)[3]
```

```
## $p.value
## [1] 0.01840684
```

```
energydrink <- subset(run, drink == "energy")
boxplot(energydrink[,1],energydrink[,2],names=c("Before","After"))
```




```
cor.test(energydrink$before,energydrink$after)[3]
```

```
## $p.value
## [1] 0.009257374
```

- In case of softdrinks, boxplots suggests increased distribution of speeds and increased median speed than before.
- In case of energy drink, boxplots suggests that running speed was decreased after taking the energy drink because the median of speed is lower than before and overall distribution also decreased.

c) For each pupil compute the time difference between the two running tasks. Test whether these time differences are effected by the type of drink.

```
time_diff <- numeric(length = length(run$before))
for (i in seq_along(time_diff)) {
  time_diff[i] <- run$before[i] - run$after[i]
}
time_diff
```

```
##  [1]  0.93 -0.58 -0.47  0.45 -0.92  0.81  0.00 -0.76 -0.80 -0.40  0.10 -0.10
## [13]  0.22 -0.60 -0.09  0.10  0.60  0.09  0.40  0.60  0.30  0.14  0.09  0.00
```

```
drink_type <- as.numeric(as.factor(run$drink))
cor.test(time_diff,drink_type)[3]
```

```
## $p.value
## [1] 0.1539964
```

- As per the test, there is a negative correlation but the value is significantly smaller which indicates that the time difference and type of drink is not strong correlated.

d) Can you think of a plausible objection to the design of the experiment in b) if the main aim was to test whether drinking the energy drink speeds up the running? Is there a similar objection to the design of the experiment in c)? Comment on all your findings in this exercise.

- If the main aim was to test whether drinking the energy drink speeds up the running then the other group could have been a placebo group because soft drinks could also hinder the performance of the participant. This could affect the results of the test.
- Yes, as mentioned earlier, if soft drinks had any impact on the running performance, then time difference would be significantly impacted.

Exercise 5

a) Test whether the distributions of the chicken weights for meatmeal and sunflower groups are different by performing three tests: the two samples t-test (argue whether the data are paired or not), the Mann-Whitney test and the Kolmogorov-Smirnov test.

```
# Two samples t-test
t.test(meatMeal$weight, sunFlower$weight)
```

```
##
##  Welch Two Sample t-test
##
## data:  meatMeal$weight and sunFlower$weight
## t = -2, df = 19, p-value = 0.04
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -102.57   -1.44
## sample estimates:
## mean of x mean of y
##       277       329
```

```r
# Mann-Whitney test
wilcox.test(meatMeal$weight, sunFlower$weight)
```

```
##
##  Wilcoxon rank sum test
##
## data:  meatMeal$weight and sunFlower$weight
## W = 36, p-value = 0.07
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Kolmogorov-Smirnov test
ks.test(meatMeal$weight, sunFlower$weight)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  meatMeal$weight and sunFlower$weight
## D = 0.5, p-value = 0.1
## alternative hypothesis: two-sided
```

```r
# Means of meatMeal and sunFlower
mM <- mean(meatMeal$weight)
mS <- mean(sunFlower$weight)
```

Comment on your findings. - H_0 is rejected for the two samples t-test. The true difference between the means are not zero thus we cannot assume that they are paired and observation number does not match as well to have a paired test. - H_0 of equal medians is rejected for the Mann-Whitney test. The underlying distribution of meatmeal is shifted to the left from that of sunflower. - H_0 of equal means should not be rejected since it is larger than 5%. But when we look at the means of both variables, 276.909 and 328.917, we can see that they differ from each other even though test results say that we should not to.

b) Conduct a one-way ANOVA to determine whether the type of feed supplement has an effect on the weight of the chicks.

```r
# One way ANOVA test
chicaov = lm(weight ~ feed, data = chickwts)
anova(chicaov)
```

```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value  Pr(>F)
## feed       5 231129   46226    15.4 5.9e-10 ***
## Residuals 65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(chicaov)$coefficients
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      323.58       15.8  20.436 5.33e-30
## feedhorsebean   -163.38       23.5  -6.957 2.07e-09
## feedlinseed     -104.83       22.4  -4.682 1.49e-05
## feedmeatmeal     -46.67       22.9  -2.039 4.56e-02
## feedsoybean      -77.15       21.6  -3.576 6.65e-04
## feedsunflower      5.33       22.4   0.238 8.12e-01
```

Give the estimated chick weights for each of the six feed supplements.

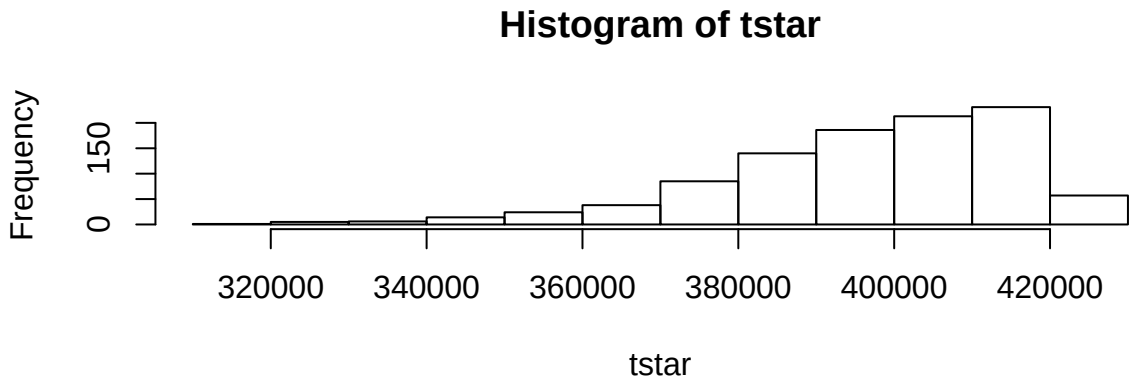
```
# Set the base mean mu1=323.583
mu_hat1 <- 323.583
mu_hat2 <- -163.383 + mu_hat1
mu_hat3 <- -104.833 + mu_hat1
mu_hat4 <- -46.674 + mu_hat1
mu_hat5 <- -77.155 + mu_hat1
mu_hat6 <- 5.333 + mu_hat1
```

Estimated chick weights for each of the six supplements are; with casein, it is 323.583, with horsebean, it is 160.2, with linseed, it is 218.75, with meatmeal, it is 276.909, with soybean, it is 246.428, with sunflower, it is 328.916.

What is the best feed supplement? - According to the ANOVA results, sunflower type of feed is the most relevant and has resemblance with the base level(Intercept).

c) Check the ANOVA model assumptions by using relevant diagnostic tools.

```
# permutation tests for independent samples
attach(chickwts, warn.conflicts = FALSE)
mystat = function(x) sum(residuals(x)^2)
B = 1000
tstar = numeric(B)
for (i in 1:B) {
  feedtstar = sample(feed)
  tstar[i] = mystat(lm(weight ~ feedtstar, data = chickwts))
}
myt = mystat(lm(weight ~ feed, data = chickwts))
hist(tstar)
```



```
myt
```

```
## [1] 195556
```

```
pl=sum(tstar<myt)/B
pr=sum(tstar>myt)/B
2 * pl
```

```
## [1] 0
```

- Permutation test for independent samples has the same setting as 1-way ANOVA. The p value that we had from the permutation test shows that feeding supplement plays a significant role for the weight of the chicken.

d) Does the Kruskal-Wallis test arrive at the same conclusion about the effect of feed supplement as the test in b)? Explain possible differences between this conclusion and the conclusion from b).

```
# the Kruskal-Wallis test
attach(chickwts, warn.conflicts = FALSE); kruskal.test(weight, feed)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  weight and feed
## Kruskal-Wallis chi-squared = 37, df = 5, p-value = 5e-07
```

The command kruskal.test performs the Kruskal-Wallis test and yields a p-value. The p-value for testing H_0 : F1 = F2 = F3 = F4 is 5.11283e-07, hence H_0 is rejected. We got similar p value from the ANOVA test as well which was 5.936e-10. The reason we say similar is because they are both very small values even though they differ each other by 10^3.