EDDA - Assignment 1 - Group 77

Dante de Lang, Ignas Krikštaponis and Kamiel Gülpen

Exercise 1

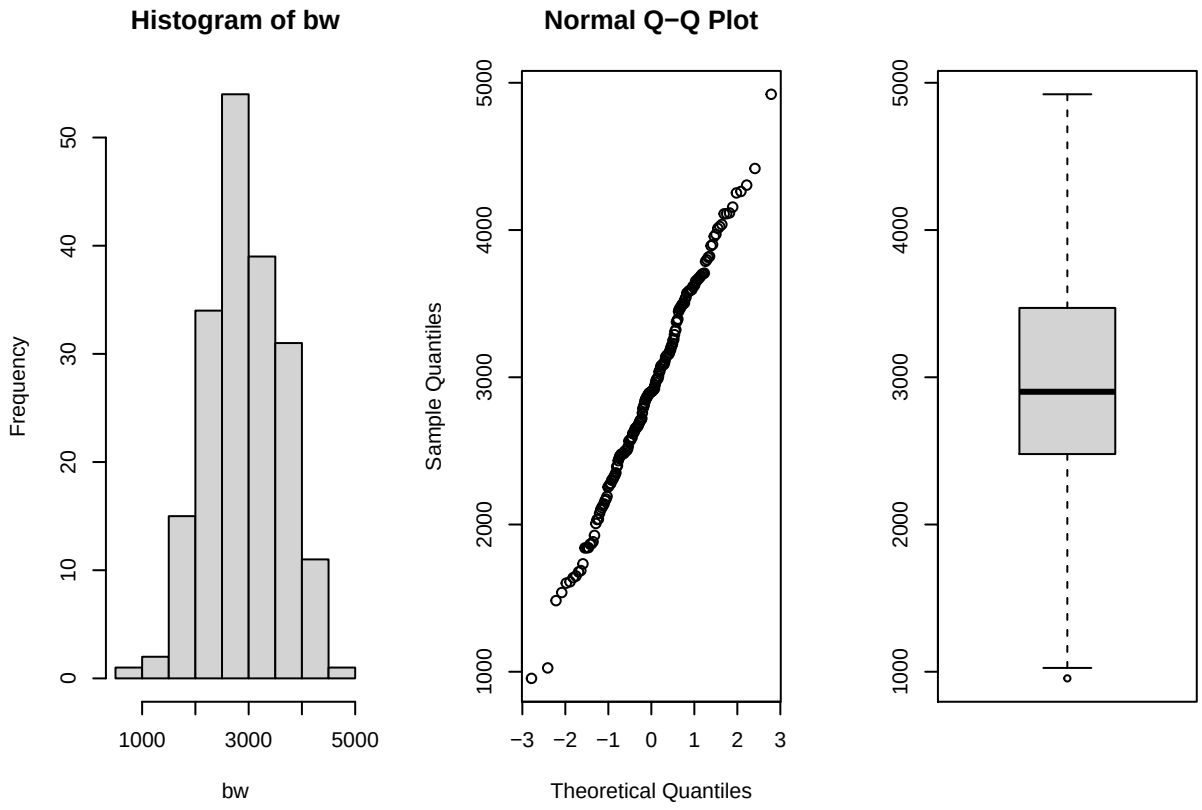
The data set birthweight.txt contains the birthweights of 188 newborn babies. We are interested in finding the underlying (population) mean mu of birthweights.

a) Check normality of the data. Compute a point estimate for mu. Derive, assuming normality (irrespective of your conclusion about normality of the data), a bounded 90% confidence interval for μ.

To check normality for the data we use a qqplot, historgram, box plot and Shapiro-Wilks test.

```
par(mfrow=c(1,3))
data <- read.table(file="data/birthweight.txt",header=TRUE)

bw <- data$birthweight
hist(bw)
qqnorm(bw)
boxplot(bw)
```



```r
# perform Shapiro-Wilk normality test
round(shapiro.test(bw)$p.value, 3)
```

```
## [1] 0.9
```

The graphical methods show that the data is normal, due to the bell shaped histogram and straight diagonal QQ-plot. The Shapiro-Wilk test reinforces this assumption as it shows a p-value of > 0.05, meaning that the H_0 is not rejected and therefore the data is normal. Furthermore, a point estimate for μ is conducted along side a 90% confidence interval.

```r
m = mean(bw)
sd = sd(bw)
n = length(bw)
error = qnorm(0.95)*sd/sqrt(n)
ci = c(m-error, m+error)
paste0("mean: ", round(m, 3))
```

```
## [1] "mean: 2913.293"
```

```r
paste0("90% confidence interval: (", round(ci[1], 3), ",", round(ci[2], 3), ")")
```

```
## [1] "90% confidence interval: (2829.618,2996.967)"
```

b) An expert claims that the mean birthweight is bigger than 2800, verify this claim by using a t-test. What is the outcome of the test if you take $\alpha = 0.1$? And other values of α?

```r
round(t.test(bw, mu=2800, alternative = "greater", conf.level = 0.95)$p.value, 3)
```

```
## [1] 0.014
```

A t-test is performed to verify the claim that the mean birthweight is bigger than 2800. The t-test shows a p-value of 0.014. This means that this claim is significant for an alpha of 0.1. The claim is significant for all alpha's above 0.014 and insignificant for alpha's below 0.014.

c) In the R-output of the test from b), also a confidence interval is given, but why is it different from the confidence interval found in a) and why is it one-sided?

The confidence interval interval is different because the one-sample t-test returns a 95% confidence interval while a 90% confidence interval is conducted in 1b). The confidence interval is one sided because the critical area of the weight distribution is compared to a mean where it is greater than 2800, but not both greater and less than 2800.

Exercise 2

We study the power function of the two-sample t-test (see Section 1.9 of Assignment 0). For n=m=30, mu=180, nu=175 and sd=5, generate 1000 samples x=rnorm(n,mu,sd) and y=rnorm(m,nu,sd), and record the 1000 p-values for testing H0: mu=nu. You can evaluate the power (at point nu=175) of this t-test as fraction of p-values that are smaller than 0.05.

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). Plot the power as a function of nu.

```r
n <- m <- 30; mu <- 180; nu <- 175; sd <- 5
grid <- seq(175,185, by=0.25)

power_function<-function(grid,n,m,mu,sd) {
  B <- 1000
  p <- numeric(B)
  G <- length(grid)
  fractions <- numeric(G)
  for (grid_nu in 1:G){
    p <- numeric(B)
    for (b in 1:B){
      x <- rnorm(n,mu,sd)
      y <- rnorm(m,grid[grid_nu],sd)
      p[b] <- t.test(x,y, var.equal = TRUE)[[3]]
    }
    fractions[grid_nu] <- mean(p<0.05)
  }
  return(fractions)
}

fractions_A <- power_function(grid,n,m,mu,sd)
```

Plots are shown further below.

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

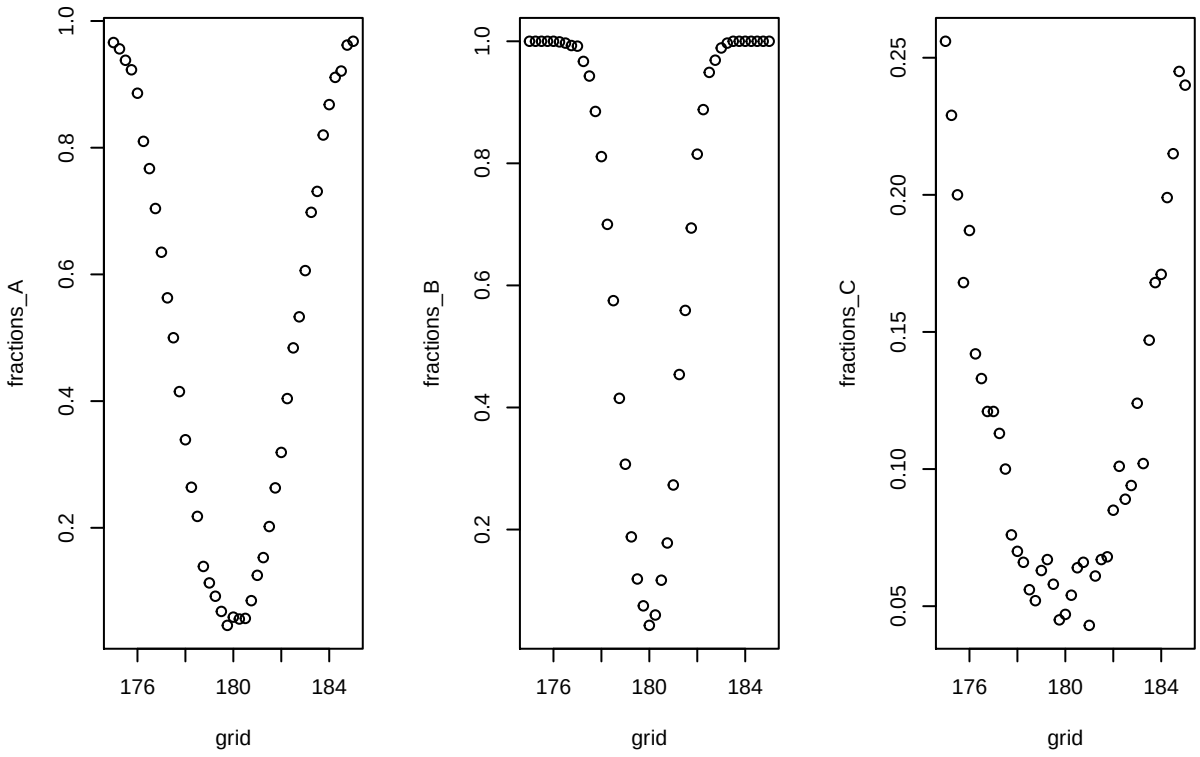
```r
n <- m <- 100; mu <- 180; sd <- 5

fractions_B <- power_function(grid,n,m,mu,sd)
```

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

```r
n <- m <- 30
mu <- 180
sd <- 15

fractions_C <- power_function(grid,n,m,mu,sd)
par(mfrow=c(1,3))
plot(grid,fractions_A)
plot(grid,fractions_B)
plot(grid,fractions_C)
```



d) Explain your findings.

More data points seems to have an influence on the narrowness of the plot and therefore seems to give a more precise outcome. Furthermore, a bigger standard deviations gives a more wider distribution of fractions as presented in the plot of C with lower values of the fractions. This can be explained by the fact that a higher standard deviations gives a higher uncertainty which results in a lower amount of fractions with a p-value below 0.05.

Exercise 3

A telecommunication company has entered the market for mobile phones in a new country. The company's marketing manager conducts a survey of 200 new subscribers for mobile phones. The results of the survey are in the data set telephone.txt, which contains the first month bills X_1,. . . ,X_200, in euros.

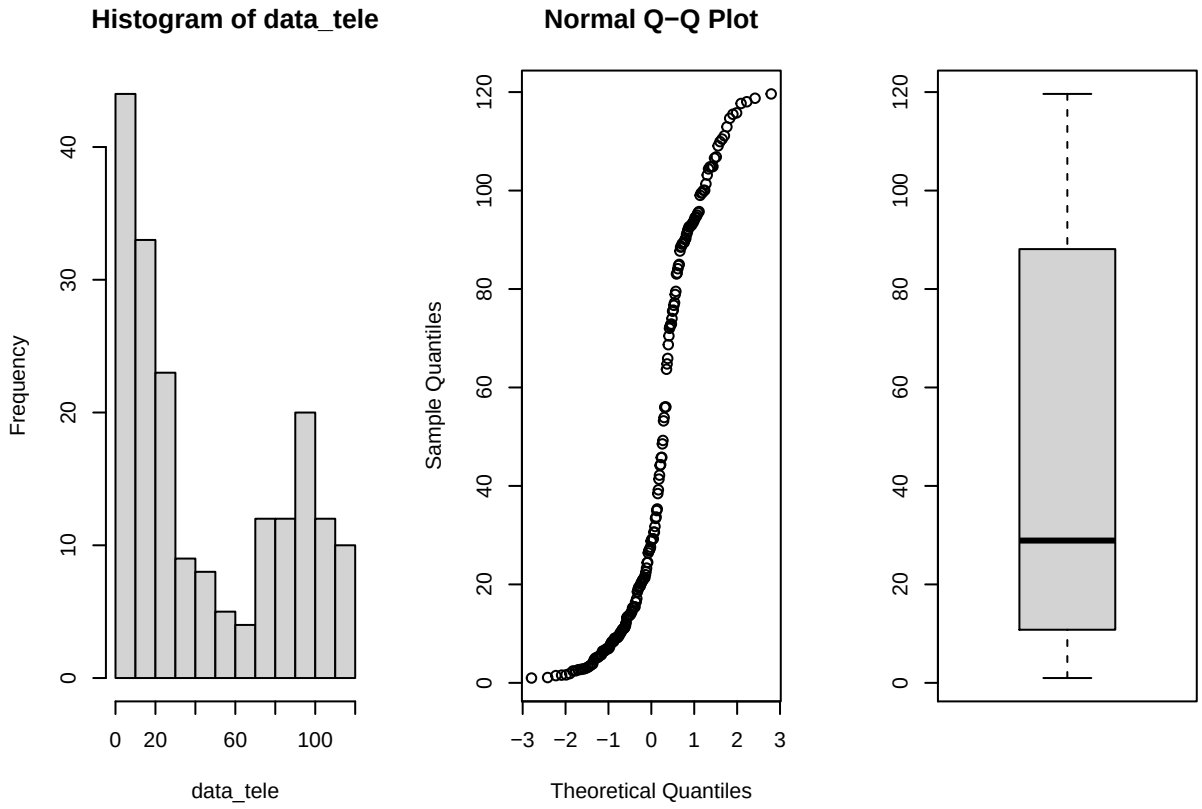
a) Make an appropriate plot of this data set. What marketing advice(s) would you give to the marketing manager? Are there any inconsistencies in the data? If so, try to fix these.

```
data<-read.table(file="data/telephone.txt",header=TRUE)

# remove zeros
data <- data %>% filter(Bills > 0)

data_tele <- data$Bills
par(mfrow=c(1,3))
hist(data_tele)
```

```
qqnorm(data_tele)
boxplot(data_tele)
```



From the survey it seems that there are two distinct peaks,therefore it would be good to run two separate marketing campaigns: a "premium" service campaign for customers who are willing to spend more and a campaign aimed at savers, usually people who use pre-paid services, - establishing a separate "cheaper" brand would be a good strategy here.

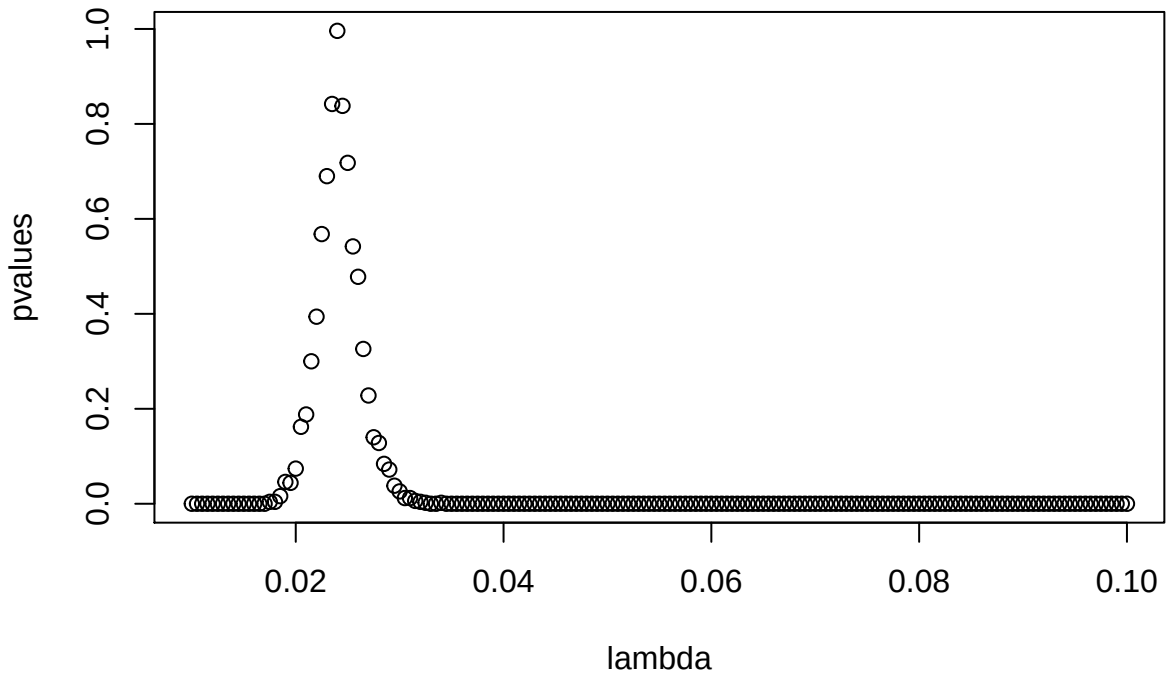
The survey data also encompassed people who did not have any spendings on the phone bills, therefore they were removed from the analysis.

b) By using a bootstrap test with the test statistic T = median($X_1, ..., X_200$), test whether the data telephone.txt stems from the exponential distribution Exp(λ) with some lambda from [0.01, 0.1].

```
lambda <- seq(0.01, 0.1, 0.0005)
pvalues <- c()
t <- median(data_tele)
for (x in lambda){
  B <- 1000
  tstar <- numeric(B)
  n <- length(data_tele)

  for (i in 1:B){
    xstar <- rexp(n,x)
    tstar[i] <- median(xstar)
  }
```

```
  pl<-sum(tstar<t)/B
  pr<-sum(tstar>t)/B
  p<-2*min(pl,pr)
  pl;pr;p
  pvalues <- c(pvalues,p)
}
plot(lambda, pvalues)
```




```
best_p <- which(pvalues > 0.05)
lambda[best_p]
```

```
##  [1] 0.0200 0.0205 0.0210 0.0215 0.0220 0.0225 0.0230 0.0235 0.0240 0.0245
## [11] 0.0250 0.0255 0.0260 0.0265 0.0270 0.0275 0.0280 0.0285 0.0290
```

The figure above shows the p-values for different lambda values. There can be concluded that the data stems from a exponential distribution for the lambda values 0.02 to 0.029.

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

```
B <- 1000
T1 <- median(data_tele)
Tstar <- numeric(B)
for (i in 1:B){
  Xstar <- sample(data_tele,replace=TRUE)
  Tstar[i] <- median(Xstar)
```

```
}
Tstar25 <- quantile(Tstar,0.025)
Tstar975 <- quantile(Tstar, 0.975)

paste0("Data median: ", round(T1, 3))
```

```
## [1] "Data median: 28.905"
```

```
paste0("95% confidence interval: (", round(2*T1-Tstar975, 3),
       ",", round(2*T1-Tstar25, 3), ")")
```

```
## [1] "95% confidence interval: (15.62,36.721)"
```

The 95% bootstrap confidence interval of the median can be observed above. The confidence interval of the bootstrap method changes each time it is executed because of the random nature of the method.

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

The variable opt_Lambda is the lambda value for which the p-value was the highest. The CLT allows us to compute a normal confidence intervals to data that are not themselves normally distributed and therefore can be used to the exponentially distributed data.

```
max_index <- which.max(pvalues)

opt_Lambda <- lambda[max_index]

paste0("Optimal lambda value found: ", round(opt_Lambda, 3))
```

```
## [1] "Optimal lambda value found: 0.024"
```

```
# simulate exponential distribution and calculate the ci
medians <- c()
B <- 1000
n <- length(data_tele)

for (i in 1:B){
  medians <- c(medians, median(rexp(n,opt_Lambda)))
}

sd = sd(medians)
m = mean(medians)
error = qnorm(0.975)*sd/sqrt(B)
ci = c(m-error, m+error)
paste0("Median estimate: ", round(m, 3))
```

```
## [1] "Median estimate: 28.863"
```

```
paste0("95% confidence interval: (", round(ci[1], 3), ",", round(ci[2], 3), ")")
```

```
## [1] "95% confidence interval: (28.673,29.053)"
```

Optimal λ was selected from the experiment in b) that resulted in the highest p-value. The results show a 95% confidence interval that is more narrow than the one found via bootstrapping. This is attributed to the stochastic nature of the bootstrap and the fact that bootstrap used real data clearly deviated from perfect exponential distribution.

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

```
bill_smal40 <- sum(data_tele<40)

binom.test(bill_smal40, length(data_tele))
```

```
##
##  Exact binomial test
##
## data:  bill_smal40 and length(data_tele)
## number of successes = 109, number of trials = 192, p-value = 0.07
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.494 0.639
## sample estimates:
## probability of success
##                  0.568
```

```
bill_less10 <- sum(data_tele < 10)
bill_less10/length(data_tele)
```

```
## [1] 0.229
```

In the above cell one can observe that a binomial test is used with as x-value the data of the telephone company where the bills where lower than 40 euro. The null hypothesis of this test states that that the median bill is bigger or equal to 40 euro while the alternative hypothesis states that the the median bill is smaller than 40 euro. After conducted this test a p-value of 0.07 is found, which means that the Null hypothesis can not be rejected as it is higher than the standard alpha of 0.05. This also means that the alternative hypothesis can not be accepted. Furthermore a test is conducted to see whether the fraction of the bills less than 10 euro is less than 25%. In the above cell can be seen that the fraction is equal to 0.229 which is smaller than 0.25 and therefore we can state that the fraction is less than 25%.

Exercise 4

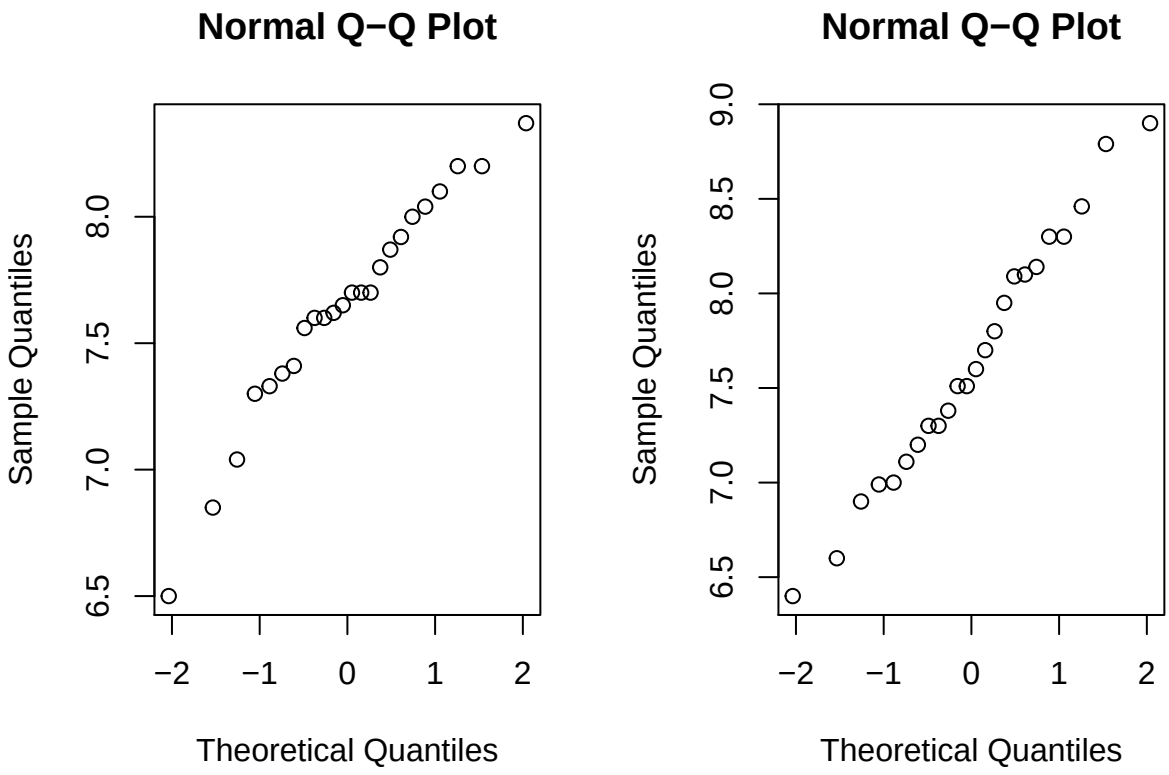
To study the effect of energy drink a sample of 24 high school pupils were randomized to drinking either a softdrink or an energy drink after running for 60 meters. After half an hour they were asked to run again. For both sprints they were asked to sprint as fast they could, and the sprinting time was measured. The data is given in the file run.txt. [Courtesy class 5E, Stedelijk Gymnasium Leiden, 2010.]

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

```
data <- read.table(file="data/run.txt",header=TRUE)
cor.test(data$before, data$after)$p.value
```

```
## [1] 0.00078
```

```
## disgnostics
par(mfrow=c(1,2)); qqnorm(data$before); qqnorm(data$after)
```



To test whether the data is correlated we run a Pearson correlation test. From the resulting p-value above we can conclude that there is significant correlation between the running times before drink and after drink. Assumption of normality needed for the performed test was confirmed by a qqnorm plot.

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

```
# calculate differences
data <- data %>%
  mutate(diff = before - after)
# filter for lemo and perdorm paired t-test
lemo <- data %>%
  filter(drink == "lemo")
paste0("p-value for soft drink: ",
       round(t.test(lemo$before, lemo$after, paired = TRUE)$p.value, 3))
```

```
## [1] "p-value for soft drink: 0.437"
```

```
# filter for energy and perform paired t-test
energy <- data %>%
  filter(drink == "energy")
paste0("p-value for energy drink: ",
       round(t.test(energy$before, energy$after, paired = TRUE)$p.value, 3))
```

```
## [1] "p-value for energy drink: 0.126"
```

Paired t-test was selected for this task as the experimental data was collected for the same individual after and before drink. For both energy and soft-drink groups there does not seem to be a significant difference in means of the running times as the p-values are < 0.05.

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.

```
# perform t-test
t.test(lemo$diff, energy$diff)$p.value
```

```
## [1] 0.159
```

The p-value is > 0.05, therefore the means of the two populations are not significantly different.

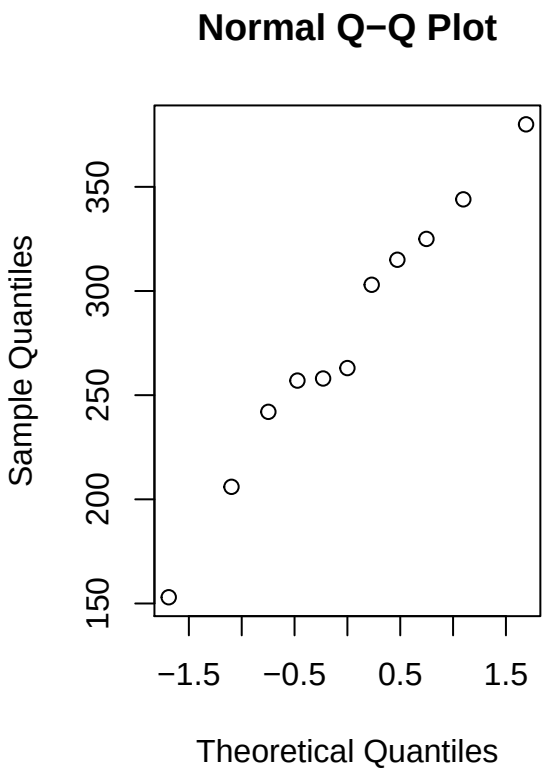
d) Can you think of a plausible objection to the design of the experiment in b) if the main aim was to test whether drinking the energy drink speeds up the running? Is there a similar objection to the design of the experiment in c)? Comment on all your findings in this exercise.

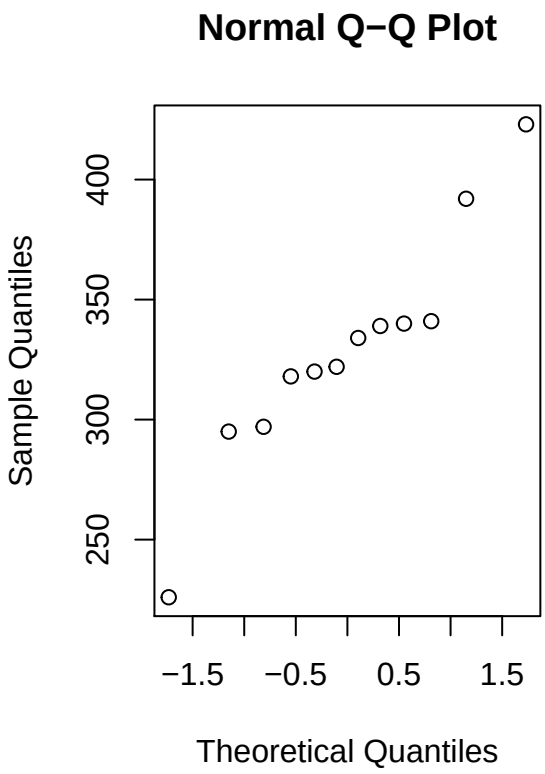
In both experiments the participants were asked to run on the same day. This could strongly influence the outcomes in data. Therefore, the setup was certainly not ideal to check the influence of both drinks. Same objection does not hold for the experiment in c) since both of the groups were independent and underwent same experimental conditions.

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. Comment on your findings.

```
# filter for meatmeal
meatmeal <- chickwts %>%
  filter(feed == "meatmeal") %>%
  select(weight)
# filter for sunflower
sunflower <- chickwts %>%
  filter(feed == "sunflower") %>%
  select(weight)
# check for data normality
par(mfrow=c(1,2))
qqnorm(meatmeal$weight)
qqnorm(sunflower$weight)
```






```
# perform t-test, the data is not paired
paste0("t-test p-value: ",
       round(t.test(meatmeal, sunflower)$p.value, 3))
```

```
## [1] "t-test p-value: 0.044"
```

```
# Mann-Whitney test
paste0("Mann-Whitney test p-value: ",
       round(wilcox.test(meatmeal$weight, sunflower$weight)$p.value, 3))
```

```
## [1] "Mann-Whitney test p-value: 0.069"
```

```
# Kolmogorov-Smirnov test
paste0("Kolmogorov-Smirnov test: ",
       round(ks.test(meatmeal$weight, sunflower$weight)$p.value, 3))
```

```
## [1] "Kolmogorov-Smirnov test: 0.108"
```

Data in chickwts is not paired as the "treatment" of different feed was applied to different newly-hatched chicks, therefore the data is independent. From t-test we can see that the p-values <0.05, this would conclude that the means between the two groups are significantly different. From Mann-Whitney test we can see that p-value is >0.05, therefore we cannot conclude that the medians of the two datasets are different. From Kolgomorov-Smirnov test we can see that p-value is >0.05, therefore we cannot conclude that the means are

different. From qqnorm plots we can observe that the "sunflower" feed data is not normal, therefore t-test here is inappropriate.

b)Conduct a one-way ANOVA to determine whether the type of feed supplement has an effect on the weight of the chicks. Give the estimated chick weights for each of the six feed supplements. What is the best feed supplement?

```
chickaov <- lm(weight~feed, data = chickwts)
# performing one-way ANOVA
anova(chickaov)
```

```
## 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
```

From the results of one-way ANOVA we can see that the p-values is < 0.05, therefore we can conclude that at least one of the means between different feed varieties are significantly different from the rest.

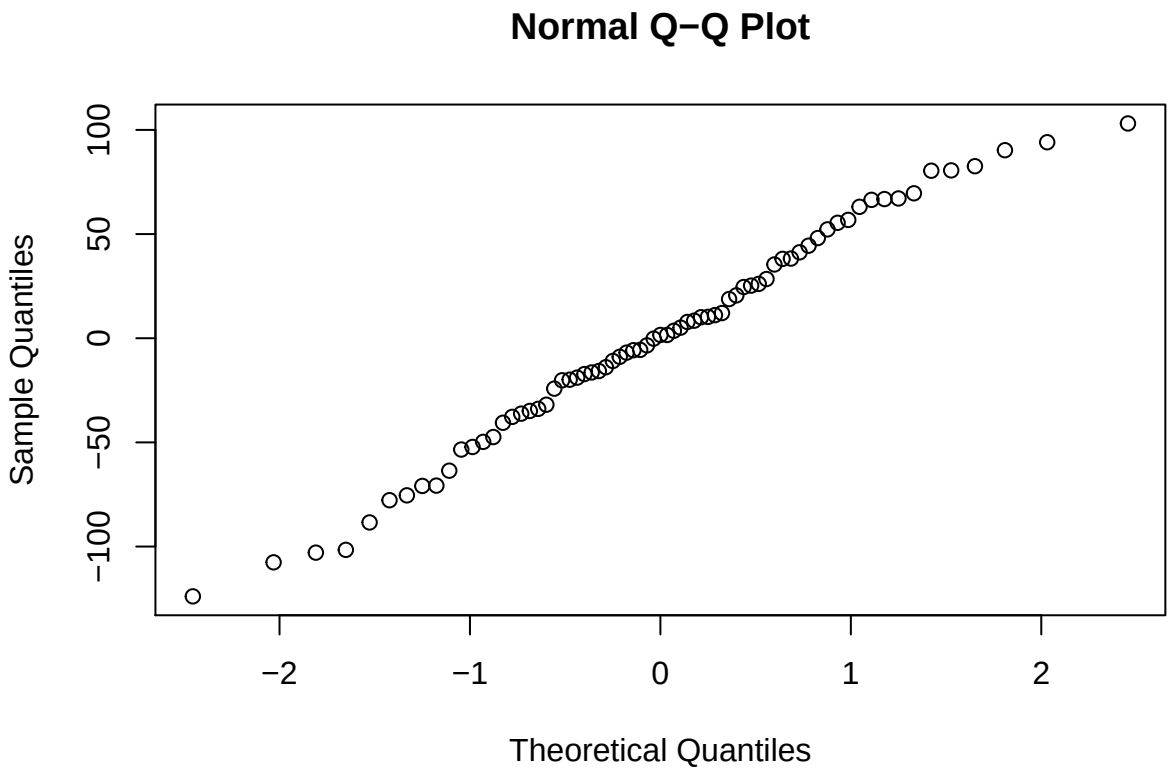
```
#extracting more information
table <- data.frame(summary(chickaov)$coefficients)
table %>%
  rename("p-value" = `Pr...t..`) %>%
  add_rownames("Type") %>%
  mutate(Type = case_when(
    Type == "(Intercept)" ~ "feedcasein",
    TRUE ~ Type
  ), Estimate = case_when(
    Type != "feedcasein" ~ as.numeric(table$Estimate[1]) + Estimate,
    TRUE ~ Estimate
  )) %>%  mutate_if(is.numeric, funs(as.character(signif(., 3)))) %>%
    knitr::kable(.)
```

Type	Estimate	Std..Error	t.value	p-value
feedcasein	324	15.8	20.4	5.33e-30
feedhorsebean	160	23.5	-6.96	2.07e-09
feedlinseed	219	22.4	-4.68	1.49e-05
feedmeatmeal	277	22.9	-2.04	0.0456
feedsoybean	246	21.6	-3.58	0.000665
feedsunflower	329	22.4	0.238	0.812

From summary statistics it seems that "sunflower" feed is the feed resulting in the highest weight, however by looking at the p-values we can see that it is not significantly different from "casein" feed variety, therefore we cannot conclude which one of these two is the best.

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

```
# check for normality
qqnorm(chickaov$residuals)
```



From qqplot the assumption of normality holds.

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 conclusions of the Kruskal-Wallis and ANOVA tests.

```
kruskal.test(weight~feed, data = chickwts)$p.value
```

```
## [1] 5.11e-07
```

With Kruskal-Wallis test we arrive to the same conclusion as with ANOVA. This is an expected outcome as ANOVA works with normal data (assumption verified in c) and Kruskal-Wallis test works with any type of data.