EDDA - Assignment 1

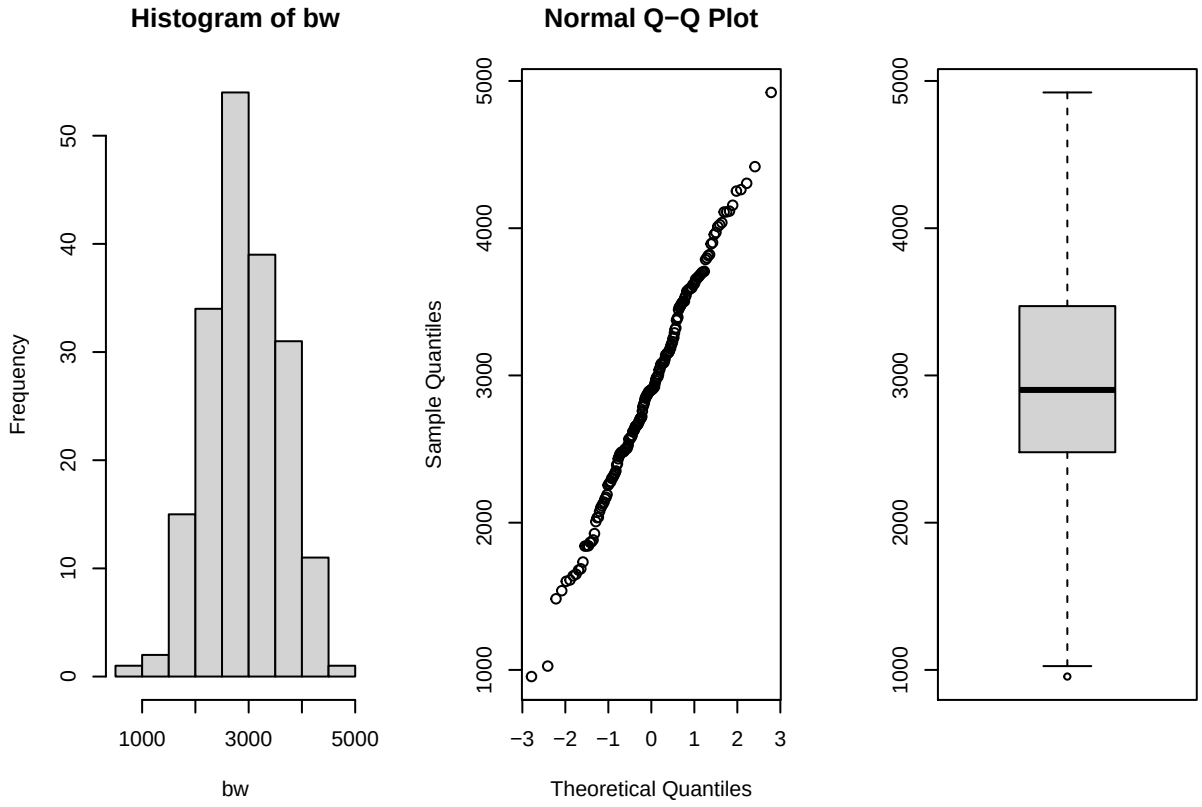
Exercise 1

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

a) Check normality of the data. Compute a point estimate for mu. Derive, assuming normality (irrespective of your conclusion about normality od 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)
```



```
shapiro.test(bw)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bw
## W = 1, p-value = 0.9
```

The graphical methods show that the data is normal. The Shapiro-Wilk test reinforces this assumption as it shows a p-value of 0.8995, meaning that the H0 is not rejected and therefore the data is normal. Furthermore, a point estimate for μ is conducted along side a 90% confidence interval.

```
m = mean(bw)
sd = sd(bw)
n = length(bw)
error = qnorm(0.95)*sd/sqrt(n)
ci = c(m-error, m+error)
m
```

```
## [1] 2913
```

```
ci
```

```
## [1] 2830 2997
```

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

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 α of 0.1. The claim is significant for all α's above 0.014 and insignificant for α's below 0.014.

```
t.test(bw, mu=2800, alternative = "greater", conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  bw
## t = 2.2, df = 187, p-value = 0.01
## alternative hypothesis: true mean is greater than 2800
## 95 percent confidence interval:
##  2829  Inf
## sample estimates:
## mean of x
##      2913
```

c) In the R-output of the test from b), also a confidence interval is given, but why is it different from theconfidence 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, μ=180, ν=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: μ=ν. You can evaluate the power (at point ν=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 ν in the grid seq(175,185,by=0.25). Plot the power as a function of ν.

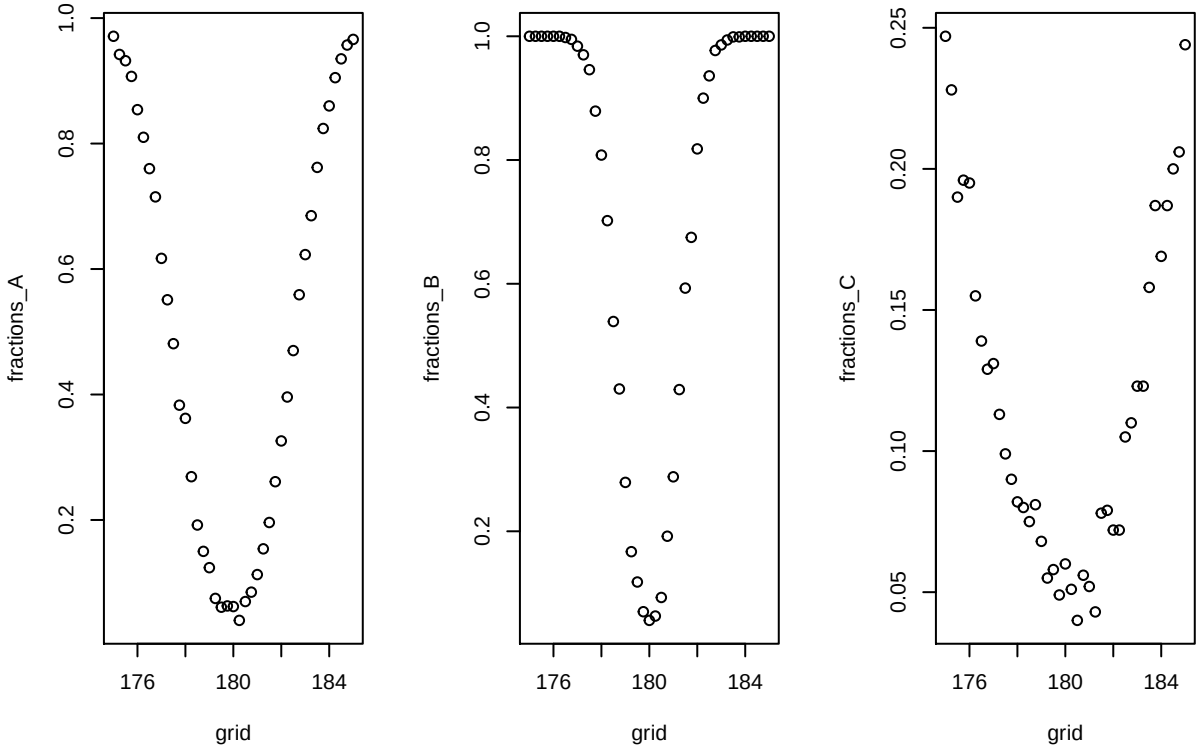
```
n <- m <- 30
mu <- 180
nu <- 175
sd <- 5
grid <- seq(175,185, by=0.25)

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)
```

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

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

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

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

```
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.

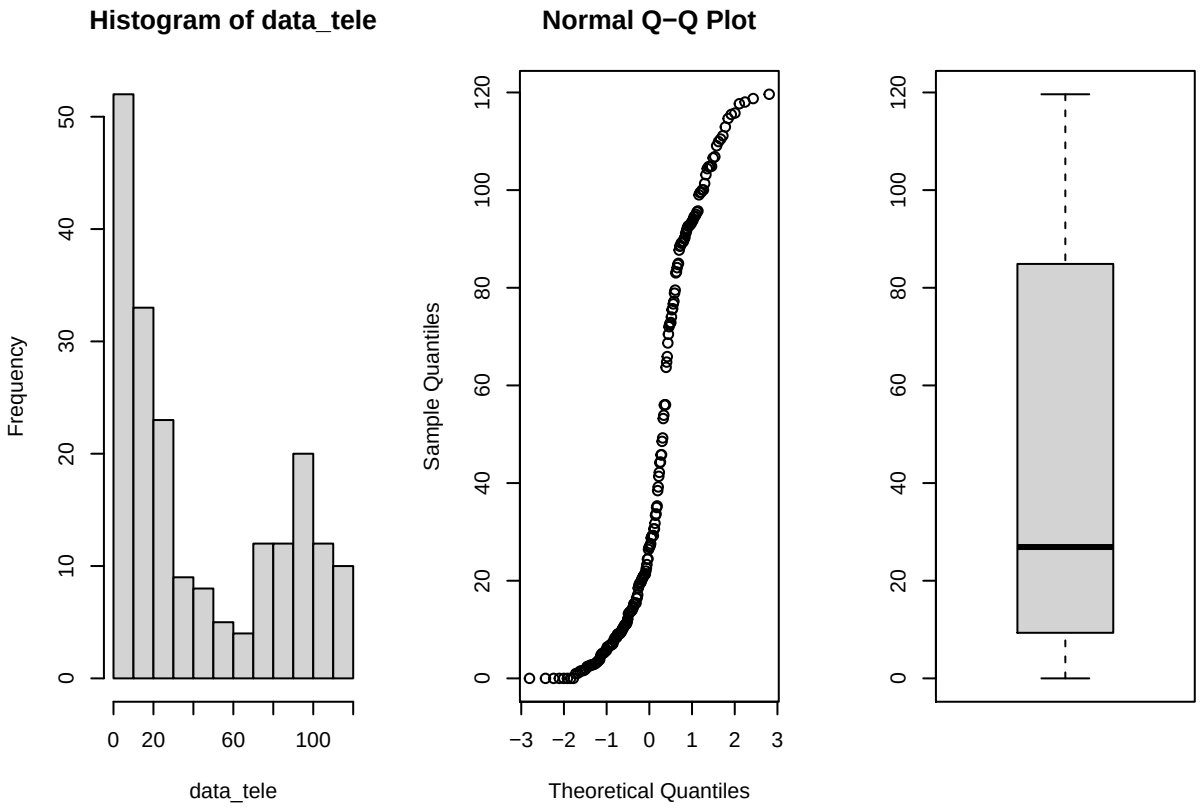
The more datapoints seems to have an influence on the narrowness of the plot. Furthermore, a bigger std gives a more wider distribution of fractions as presented in the plot of C.

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.

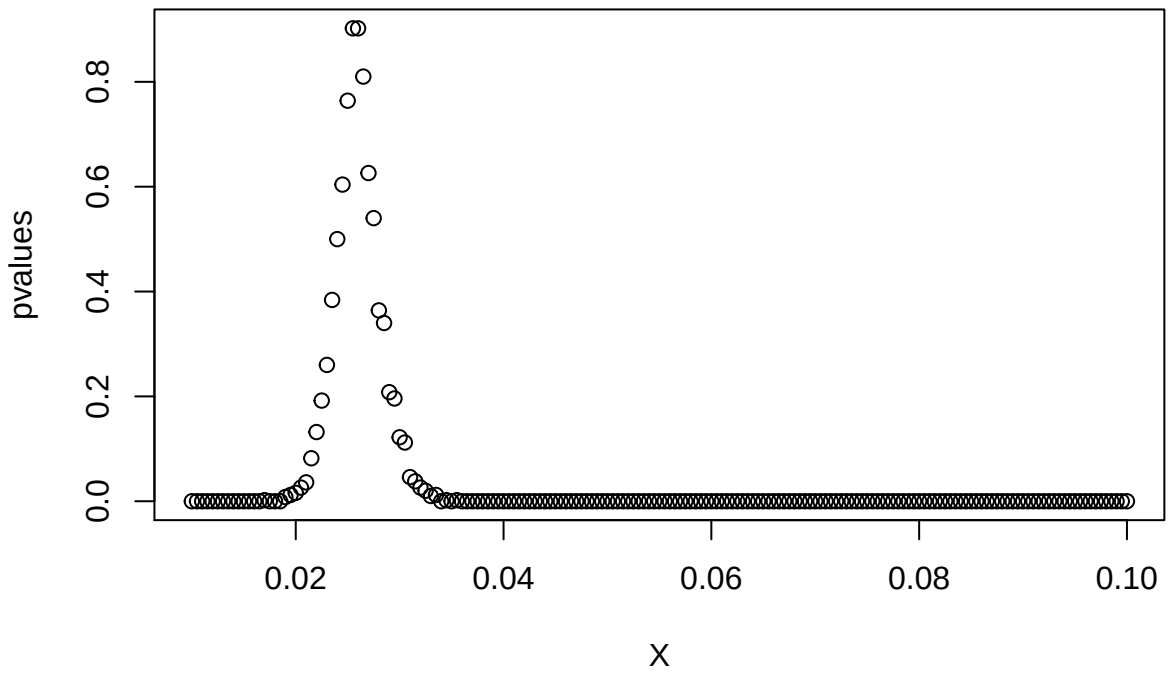
```
data<-read.table(file="data/telephone.txt",header=TRUE)
data_tele <- data$Bills
par(mfrow=c(1,3))
hist(data_tele)
qqnorm(data_tele)
boxplot(data_tele)
```



The data seems oddly distributed with two subpeaks, would have expected a more normally distributed set. Therefore perhaps a good idea if the manager arranges the prices better.

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(lambda) with some lambda from [0.01, 0.1].

```
X <- seq(0.01, 0.1, 0.0005)
pvalues <- c()
t <- median(data_tele)
for (x in X){
  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(X, pvalues)
```



There exist an Exp function that fits the hypothesis

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

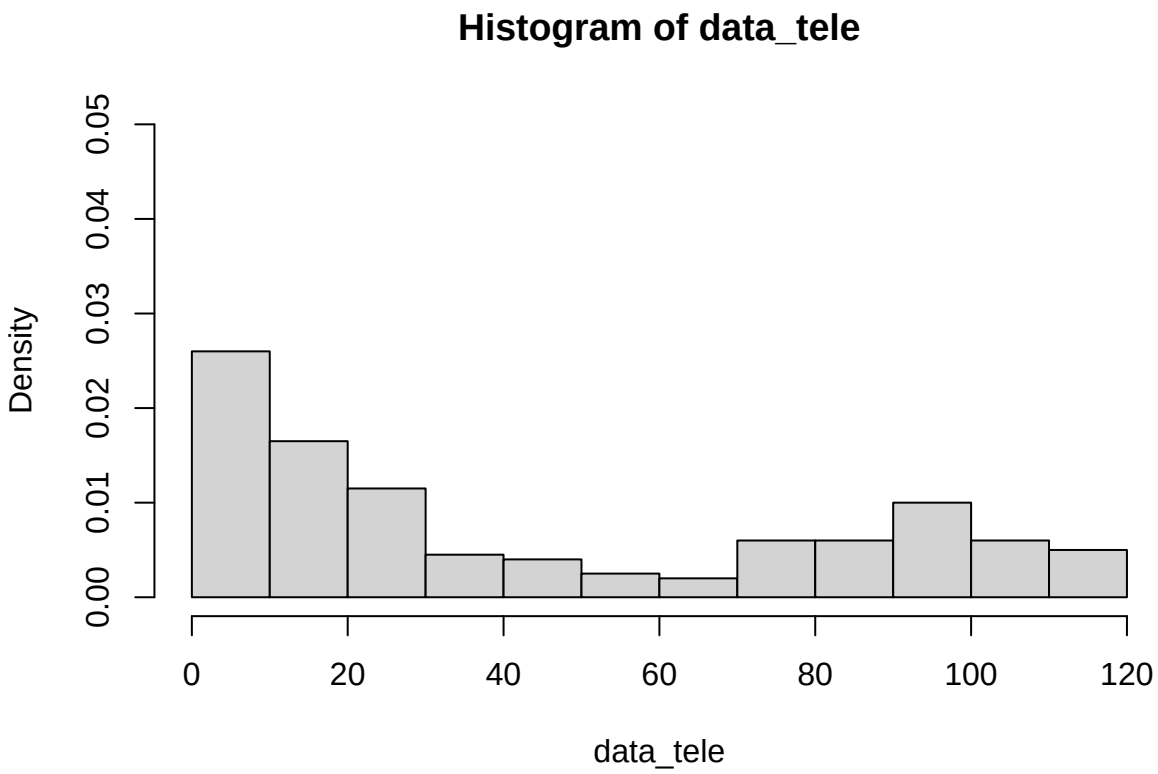
```
B <- 1000
T1 <- median(data_tele)
Tstar <- numeric(B)
for (i in 1:B){
  Xstar <- sample(data_tele,replace=TRUE)
  Tstar[i] <- median(Xstar)
}
Tstar25 <- quantile(Tstar,0.025)
Tstar975 <- quantile(Tstar, 0.975)

T1
```

```
## [1] 26.91
```

```
c(2*T1-Tstar975, 2*T1-Tstar25)
```

```
## 97.5%  2.5%
## 18.81 33.55
```

```
hist(data_tele, prob=T, ylim=c(0,0.05))
x<-seq(0,max(data_tele),length=1000)
lines(x,exp(x),type="l",col="blue",lwd=2)
```



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.

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

opt_Lambda <- X[max_index]
```

The variable opt_Lambda is the optimal lambda value with the highest P-value.

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_bigeq40 <- sum(data_tele>=40)
bill_smal40 <- sum(data_tele<40)

binom.test(bill_bigeq40, length(data_tele),p=0.5)
```

```
##
##  Exact binomial test
##
## data:  bill_bigeq40 and length(data_tele)
## number of successes = 83, number of trials = 200, p-value = 0.02
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
```

```
##  0.3459 0.4866
## sample estimates:
## probability of success
##                  0.415
```

```
binom.test(bill_smal40, length(data_tele),p=0.5)
```

```
##
##  Exact binomial test
##
## data:  bill_smal40 and length(data_tele)
## number of successes = 117, number of trials = 200, p-value = 0.02
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5134 0.6541
## sample estimates:
## probability of success
##                  0.585
```

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

```
## [1] 0.26
```

Exercise 4

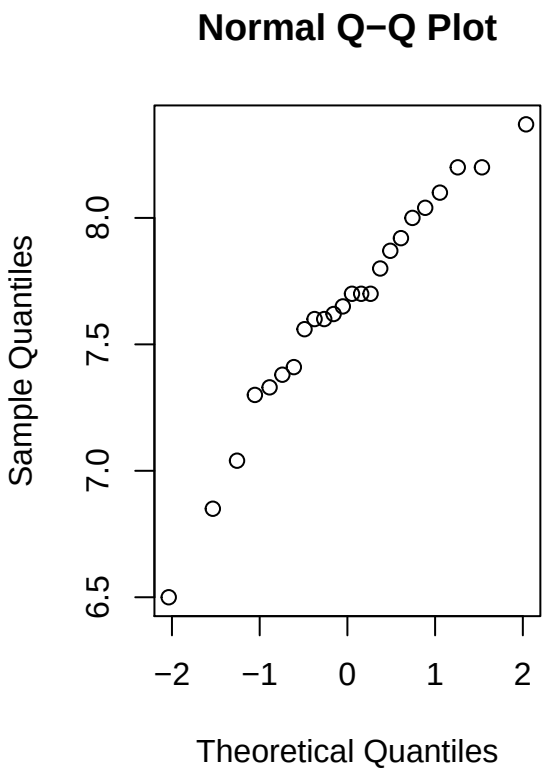
To study the effect of energy drink a sample of 24 high school pupils were randomized to drinking either a softdrink or an energy drink after running for 60 meters. After half an hour they were asked to run again. For both sprints they were asked to sprint as fast they could, and the sprinting time was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

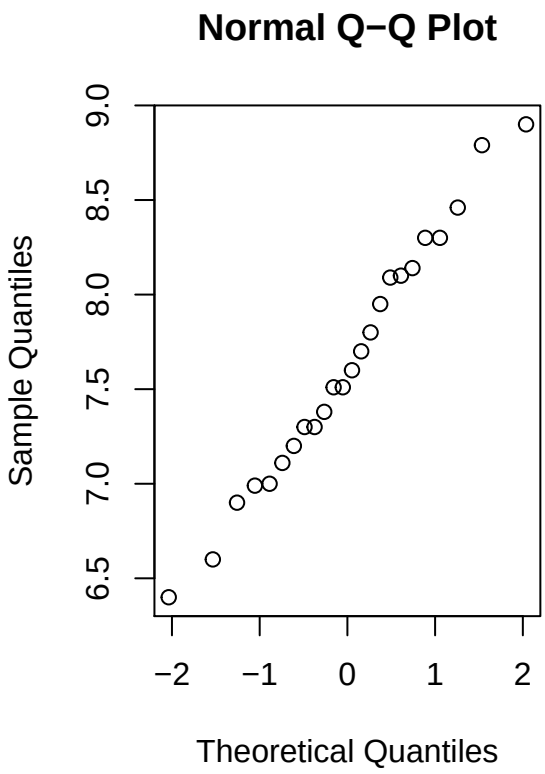
a) Disregarding the type of drink, test whether the run times before drink and after are correlated.

```
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 Person 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

lemo <- data %>%
  filter(drink == "lemo")

paste0("p-value for soft drink: ", round(t.test(lemo$before, lemo$after, paired = TRUE)$p.value, 4))
```

```
## [1] "p-value for soft drink: 0.4373"
```

```
# filter for energy

energy <- data %>%
  filter(drink == "energy")

paste0("p-value for energy drink: ", round(t.test(energy$before, energy$after, paired = TRUE)$p.value,
```

```
## [1] "p-value for energy drink: 0.1264"
```

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.

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] 0.1586
```

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

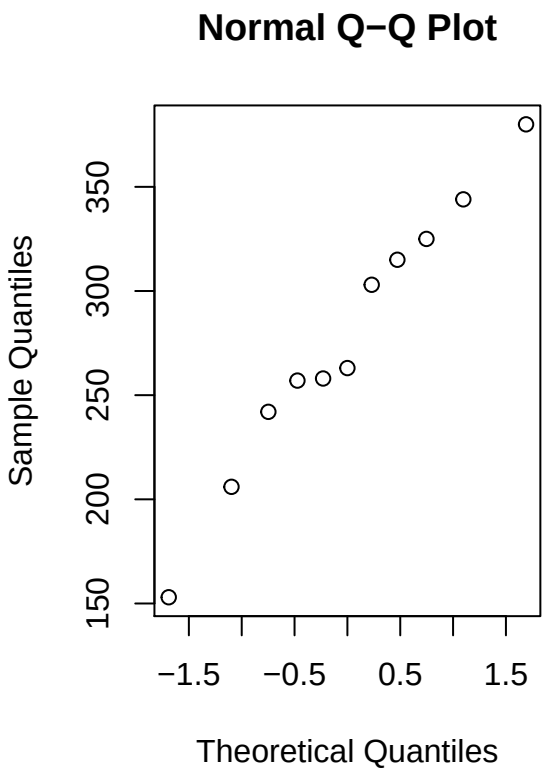
d) Can you think of a plausible objection to the design of the experiment in b) if the main aim was to test whether drinking the energy drink speeds up the running? Is there a similar objection to the design of the experiment in c)? Comment on all your findings in this exercise.

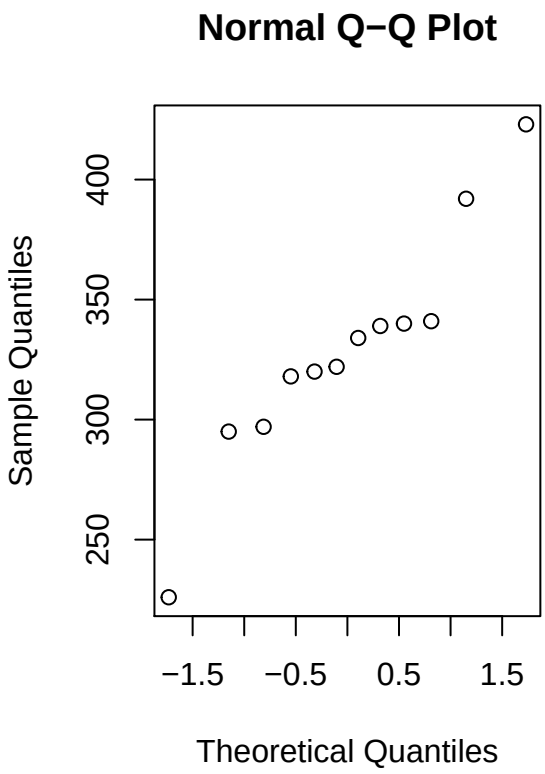
In both experiments the participants were asked to run on the same day. This could strongly influence the outcomes in data. Therefore, the setup was certainly not ideal to check the influence of both drinks.

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, 4))
```

```
## [1] "t-test p-value: 0.0444"
```

```
# Mann-Whitney test

paste0("Mann-Whitney test p-value: ", round(wilcox.test(meatmeal$weight, sunflower$weight)$p.value, 4))
```

```
## [1] "Mann-Whitney test p-value: 0.0688"
```

```
# Kolmogorov-Smirnov test

paste0("Kolmogorov-Smirnov test: ", round(ks.test(meatmeal$weight, sunflower$weight)$p.value, 4))
```

```
## [1] "Kolmogorov-Smirnov test: 0.1085"
```

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
```

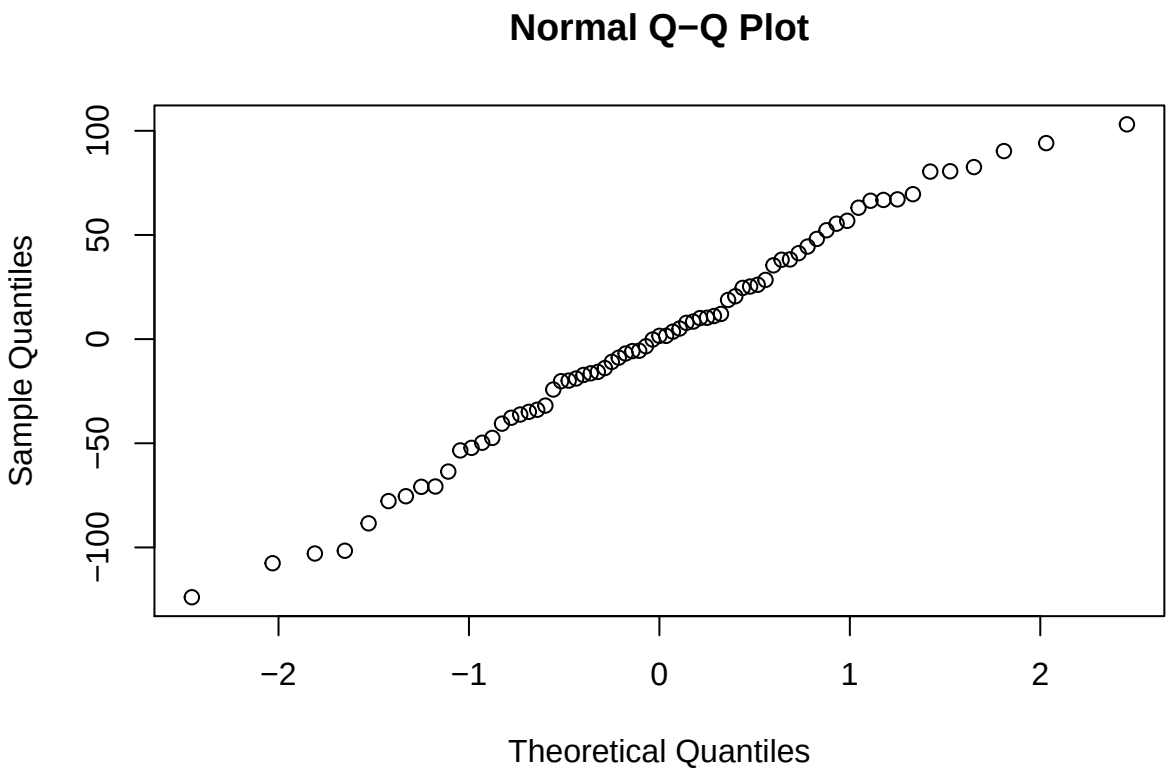
```
#extracting more information
summary_table <- as_tibble(summary(chickaov)$coefficients) %>%
  rename("p-value" = `Pr(>|t|)`)
knitr::kable(summary_table, "latex")
```

Estimate	Std. Error	t value	p-value
323.583	15.83	20.4361	0.0000
-163.383	23.49	-6.9568	0.0000
-104.833	22.39	-4.6816	0.0000
-46.674	22.90	-2.0386	0.0456
-77.155	21.58	-3.5756	0.0007
5.333	22.39	0.2382	0.8125

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

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

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




```
# check if the variances are equal
chickwts %>%
  group_by(feed) %>%
  summarise(variance = var(weight))
```

```
## # A tibble: 6 x 2
##   feed      variance
## * <fct>        <dbl>
## 1 casein       4152.
## 2 horsebean    1492.
## 3 linseed      2729.
## 4 meatmeal     4212.
## 5 soybean      2930.
## 6 sunflower    2385.
```

From qqplot assumption of normality holds. However the assumption of equal variances does not hold.

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] 5.113e-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.