Assignment 1

Rinus van Grunsven (st.number 10755373), Florens Douwes (11254483)

24 September 2022

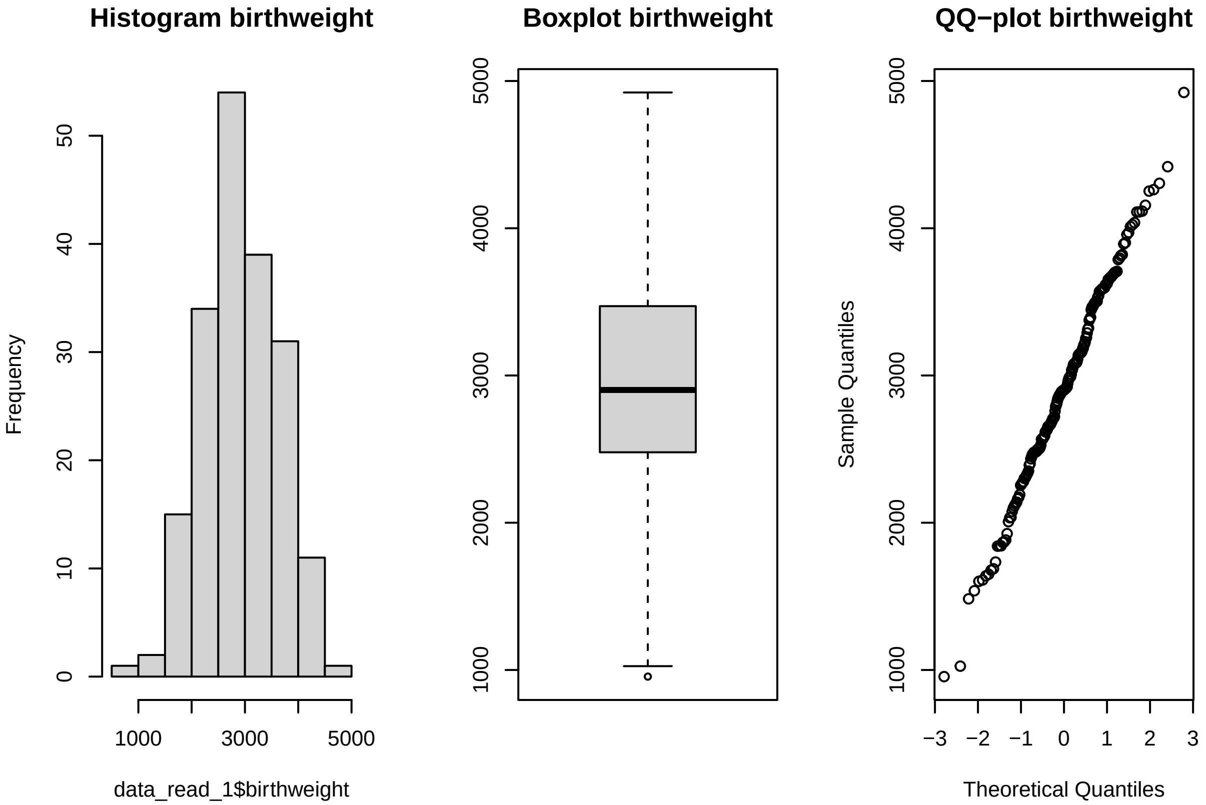
Exercise 1.1

The data that is being used in the exercise is retrieved via the following command:

```
data_read_1 = read.table("./birthweight.txt", header=TRUE)
```

a) Below are the histogram, boxplot, and QQ-plot of the data.

```
par(mfrow=c(1,3))
hist(data_read_1$birthweight, main="Histogram birthweight")
boxplot(data_read_1$birthweight, main="Boxplot birthweight")
qqnorm(data_read_1$birthweight, main="QQ-plot birthweight")
```



The figures above all show indications of a normal distribution: the histogram is symmetrical and bell-shaped, the boxplot has roughly equally distanced whiskers, and the qq-plot has an almost straight line from the bottom left to top right corner.

Below a numeric summary of the data:

```
summary(data_read_1)
```

```
##   birthweight
##  Min.   : 955
##  1st Qu.:2479
##  Median :2902
##  Mean   :2913
##  3rd Qu.:3468
##  Max.   :4922
```

The mean, rounded to three decimals

```
m = round(mean(data_read_1$birthweight),2)
m
```

```
## [1] 2913.29
```

b)

```
mu=0.2
n = length(data_read_1$birthweight); n # calculates the length
```

```
## [1] 188
```

```
s = round(sd(data_read_1$birthweight), 3); s # calculates standard deviation
```

```
## [1] 697.5
```

```
t = round(qt(0.95,df=length(data_read_1$birthweight)-1),3); t   # calculates test statistic
```

```
## [1] 1.653
```

```
ci = round(c(m-t*s/sqrt(n),m+t*s/sqrt(n)), 3); ci # 90% confidence interval
```

```
## [1] 2829.201 2997.379
```

The length is: 188. This is calculated with the following formula:

```
n = length(data_read_1$birthweight)
n
```

```
## [1] 188
```

The standard deviation is: 698. This is calculated with the following formula:

```
s = round(sd(data_read_1$birthweight), 3)
s
```

```
## [1] 697.5
```

The test statistic is: 1.6530429. This is calculated with the following formula:

```
t = round(qt(0.95,df=length(data_read_1$birthweight)-1),3)
t
```

```
## [1] 1.653
```

The confidence interval is then calculated with the formula: c(m-t*s/sqrt(n),m+t*s/sqrt(n))

```
round(c(m-t*s/sqrt(n),m+t*s/sqrt(n)), 3)
```

```
## [1] 2829.201 2997.379
```

c) We can

TODO: latex for H0 and H1

H0: mean birthweight is equal or smaller than 2800 H1: mean birthweight is bigger than 2800

```
t.test(data_read_1$birthweight,mu=2800,alt="g")
```

```
##
##  One Sample t-test
##
## data:  data_read_1$birthweight
## t = 2.2271, df = 187, p-value = 0.01357
## alternative hypothesis: true mean is greater than 2800
## 95 percent confidence interval:
##  2829.202      Inf
## sample estimates:
## mean of x
##  2913.293
```

The p-value is smaller than 0,05. This means that there is enough statistical evidence to reject H0 and thus to state that H1 is true.

d) It's different because in b) we calculated a confidence interval with a 90% confidence level, while the T-test that was conducted at c) calculated a confidence interval with a 95% confidence level. A confidence interval is a range of values that has a upper and lower bound below and above the mean of your statistics. The population parameter that you are trying to find would have to fall in this range based on the confidence level. A confidence level is the degree of certainty that the population value would fall in this range. The more you are certain that the population parameter will fall in this range, the wider the range will get. So a 90% confidence level will have a smaller range and thus a smaller confidence interval than a 95% confidence level would have. It is one-sided because we wanted to check if the mean birthweight is bigger than 2800. Therefore, the upper bound of the confidence interval does not have a value but goes to infinity as it does not matter which value it is. The null hypothesis is rejected as long as the value is bigger than the lower bound.

Exercise 1.2

a)

The sample proportion $\hat{p}$ is the best point estimate of the population proportion p. This sample proportion can be calculated by dividing the number of successes x by the sample size n.

```
x = 140
n = 200
p_hat = 140 /200
p_hat
```

```
## [1] 0.7
```

b)

The 99%-confidence interval for the population proportion p can be derived as follows:

TODO latex code

with:

TODO latex code

```
p_hat = 0.7
q_hat = 1-p_hat
n = 200
alpha = 1 - 0.99
z_alpha2 = qnorm(1 - alpha/2)
margin_of_error = z_alpha2 * sqrt ((p_hat * q_hat) / n)
round(c(p_hat - margin_of_error, p_hat + margin_of_error), 3)
```

```
## [1] 0.617 0.783
```

So to conclude, the 99%-confidence interval for p is [0.617, 0.783]

c)

In order to test the null hypothesis that the fraction is equal to 75%, we will use a binomial test. As null hypothesis, we take $H_0 : p = 0.75$ and as alternative hypothesis $H_1 : p \neq 0.75$.

```
binom.test(140, 200, p = 0.75)
```

```
##
##  Exact binomial test
##
## data:  140 and 200
## number of successes = 140, number of trials = 200, p-value = 0.103
## alternative hypothesis: true probability of success is not equal to 0.75
## 95 percent confidence interval:
##  0.6313501 0.7626104
## sample estimates:
## probability of success
##                    0.7
```

The resulting p-value is 0.103. In case that we chose a significance level (alpha) of 0.1, this implies that the p-value of 0.103 is greater than the significance level. This would lead us to the conclusion that we do NOT reject the null hypothesis.

In case the significance level would be chosen such that the p-value is smaller or equal to it, the null hypothesis can be rejected, which would be a strong conclusion.

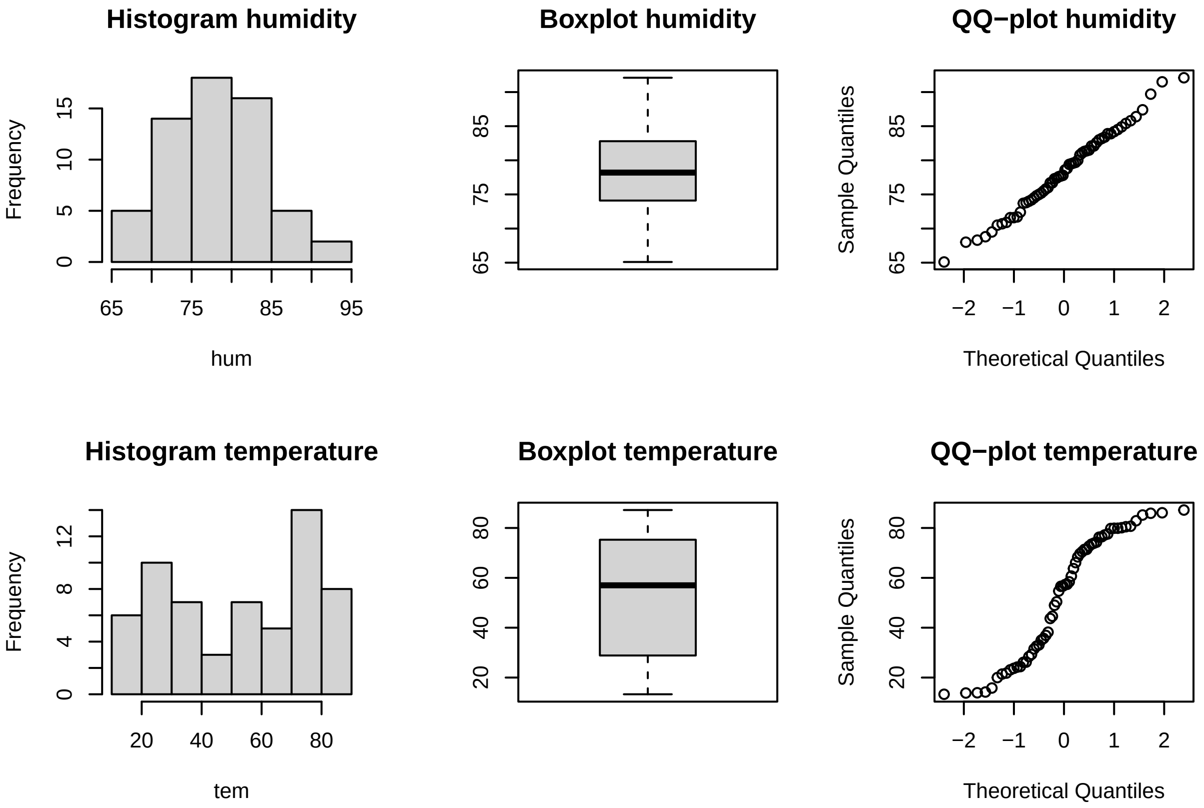
Exercise 1.3

a)

Below are the histogram, boxplot, and QQ-plot of the weather data.

```
weather = read.table(file="weather.txt", header=TRUE)
hum=weather$humidity
tem=weather$temperature

par(mfrow=c(2,3))
hist(hum, main="Histogram humidity")
boxplot(hum, main="Boxplot humidity")
qqnorm(hum, main="QQ-plot humidity")
hist(tem, main="Histogram temperature")
boxplot(tem, main="Boxplot temperature")
qqnorm(tem, main="QQ-plot temperature")
```



Histogram humidity

Boxplot humidity

QQ−plot humidity

Histogram temperature

Boxplot temperature

QQ−plot temperature

Next, a numeric summary of the humidity and temperature data:

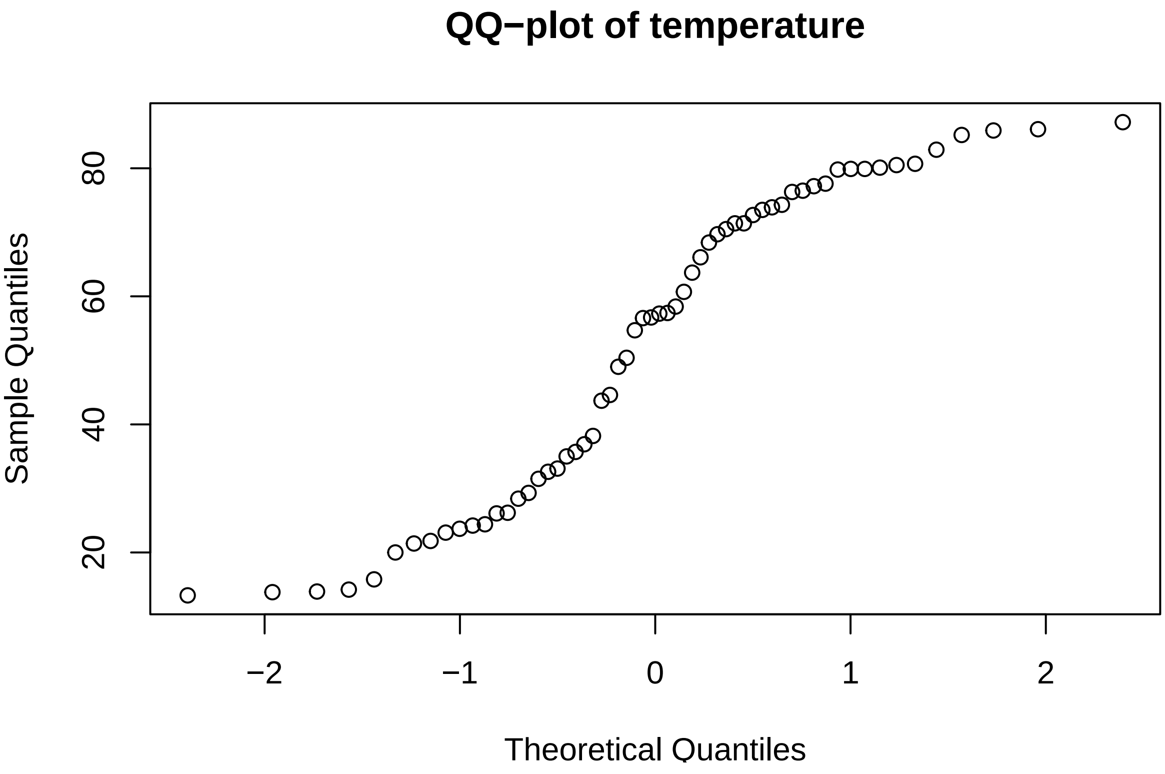
```
summary(weather)
```

```
##     humidity      temperature
##  Min.   :65.10   Min.   :13.30
##  1st Qu.:74.15   1st Qu.:29.07
##  Median :78.20   Median :57.00
##  Mean   :78.34   Mean   :52.73
##  3rd Qu.:82.70   3rd Qu.:74.80
##  Max.   :92.10   Max.   :87.20
```

b)

We'll revisit the QQ plot of the temperature:

```
qqnorm(tem, main="QQ-plot of temperature")
```



With this plot we can see that the temperature data does not follow a pattern seen with normal distributed samples. It is not a straight line in the Q-Q plot. It looks like the data is too spread out, meaning fat tails in the normal distribution.

c)

The mean is unknown, as is the standard deviation. Therefore we will estimate the confidence interval using the t-distribution.

The formula to calculate the confidence interval then is $\bar{X} \pm t_{a/2,n-1}\frac{s}{\sqrt{n}}$.

With the following R code we can calculate the 90% confidence interval.

```
alpha = 0.10
n = length(weather$temperature)
m = mean(weather$temperature)
s = sd(weather$temperature)

t = qt(1 - alpha / 2, df=n - 1)
c(m - t * s / sqrt(n), m + t * s / sqrt(n))
```

```
## [1] 47.48704 57.96296
```

This means that the true mean is (with a confidence interval of 90%) between 47.49 and 57.96.

d)

This means that the margin of error should be a maximum 2% of humidity (that is, a value of 2).

n must satisfy the following formula:

$$t_{a/2,n-1}\frac{s}{\sqrt{n}} \leq E = \sqrt{n} \geq \frac{t_{a/2,n-1^s}}{E} \approx \frac{(z_{a/2})^2 s^2}{E^2} = \frac{(1.64)^2(24.28)^2}{4} = 396.40$$

Which would mean we would need at least 397 samples.

Exercise 1.4

The following data is being used for this exercise:

```r
data_read_4 = read.table("./austen.txt", header=TRUE)
data_read_4
```

```
##         Sense Emma Sand1 Sand2
## a         147  186   101    83
## an         26   25    11    19
## this       32   38    16    15
## that       98  105    37    41
## with       59   76    28    39
## without    20   10    10     4
```

a) It is unclear by whom Sense and Emma are written. For the other two novels it is not: Sand1 is written by Austen and Sand2 is written by an admirer. Our guess is that Sense and Emma ar written by Austen as well and we will use this assumption while conducting the other exercises also. Furthermore, we presumed that the test would be performed in order to check to what extend the novels are all written in the same way. According to us, a test for homogeneity is the most suitable option to use. This test is able to check if the distribution of words is evenly spread across columns when comparing rows with each other. The distribution should be the same across rows in case the writing style is the same in all the novels.

b) We start with making a matrix of only the novels that Austen wrote. The matrix is shown below.

```r
Sense = data_read_4$Sense
Emma = data_read_4$Emma
Sand1 = data_read_4$Sand1
x=as.data.frame(matrix(c(Sense, Emma, Sand1),ncol=3,nrow=6))
dimnames(x)=list(c("a","an", "this", "that", "with", "without"),c("Sense","Emma", "Sand1"))
x
```

```
##         Sense Emma Sand1
## a         147  186   101
## an         26   25    11
## this       32   38    16
## that       98  105    37
## with       59   76    28
## without    20   10    10
```

Now that we only have the data of Austens novels, we can run a chi-squared test to see if Austen herself was consistent in her writing style or not. We will use the following hypotheses: H0: the distribution of words is the same for each column H1: the distribution of words is not the same for each column This is performed in R with the command below

```
z = chisq.test(x)
z
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 14.274, df = 10, p-value = 0.1609
```

The p-value is 0.1308 and this is greater than 0.05 (or 5%). This means that there is not enough statistical evidence to reject H0, and thus we conclude that the distribution of words is the same for each column. As a result of this, we conclude that Austen was consistent in her writing style.

Shall we include the parts below?

```
z$expected
```

```
##             Sense      Emma     Sand1
## a       161.74439 186.30244 85.953171
## an       23.10634  26.61463 12.279024
## this     32.05073  36.91707 17.032195
## that     89.44390 103.02439 47.531707
## with     60.74732  69.97073 32.281951
## without  14.90732  17.17073  7.921951
```

```
z$observed
```

```
##         Sense Emma Sand1
## a         147  186   101
## an         26   25    11
## this       32   38    16
## that       98  105    37
## with       59   76    28
## without    20   10    10
```

```
X2=sum((z$observed-z$expected)^2/z$expected)
X2
```

```
## [1] 14.27373
```

And this also?

```
1-pchisq(X2,df=(6-1)*(3-1)) # same is the same as the p-value we got from the Chi-squared test
```

```
## [1] 0.1608682
```

```
residuals(z)
```

```
##               Sense       Emma      Sand1
## a       -1.159343692 -0.0221579  1.6229833
## an       0.601979480 -0.3129781 -0.3650034
## this    -0.008961083  0.1782319 -0.2501074
## that     0.904691184  0.1946396 -1.5275909
## with    -0.224185835  0.7207861 -0.7536363
## without  1.319006371 -1.7304899  0.7383115
```

c) Since we are going to use all data, we have to add the data from the novel written by the admirer to our current matrix. This is done with the code below.

```
Sand2 = data_read_4$Sand2
x=as.data.frame(matrix(c(Sense, Emma, Sand1, Sand2),ncol=4,nrow=6))
dimnames(x)=list(c("a","an", "this", "that", "with", "without"),c("Sense","Emma", "Sand1", "Sa
x
```

```
##         Sense Emma Sand1 Sand2
## a         147  186   101    83
## an         26   25    11    19
## this       32   38    16    15
## that       98  105    37    41
## with       59   76    28    39
## without    20   10    10     4
```

The matrix now consists of data from all novels. We will again use a chi-squared test in order to test whether the admirer was successful in imitating Austen's style. We run a test for homogeneity since we test if the admirer was successful in imitating Austen's writing style. The hypotheses will therefore be: H0: the distribution of words is the same for each column H1: the distribution of words is not the same for each column The chi-squared test is performed in R with the command below

```
z = chisq.test(x)
z
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 21.528, df = 15, p-value = 0.1208
```

The p-value is 0.1208 and this is greater than 0.05 (or 5%). This means that there is not enough statistical evidence to reject H0, and thus we conclude that the distribution of words is the same for each column. As a result of this, we conclude that the admirer was successful in imitating Austen's writing style.

Shall we include the code below as well?

We also ran a second test in order to confirm if our conclusion stated above. For this second test we only took one novel of Austen and compared this with the one from the admirer. A new matrix was made for this, see below.

```
sand2 = data_read_4$Sand2
x=as.data.frame(matrix(c(Sand1, Sand2),ncol=2,nrow=6))
dimnames(x)=list(c("a","an", "this", "that", "with", "without"),c("Sand1", "Sand2"))
x
```

```
##         Sand1 Sand2
## a         101    83
## an         11    19
## this       16    15
## that       37    41
## with       28    39
## without    10     4
```

A chi-squared test was conducted on these numbers. The result is shown below.

```
z = chisq.test(x)
z
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 8.4993, df = 5, p-value = 0.1308
```

The p-value is 0.1308 and this is again greater than 0.05 (or 5%). This means that there is not enough statistical evidence to reject H0, and thus we conclude that the distribution of words is the same for each column. As a result of this, we conclude that the admirer was successful in imitating Austen's writing style.

Shall we include the code below as well?