Assignment 1

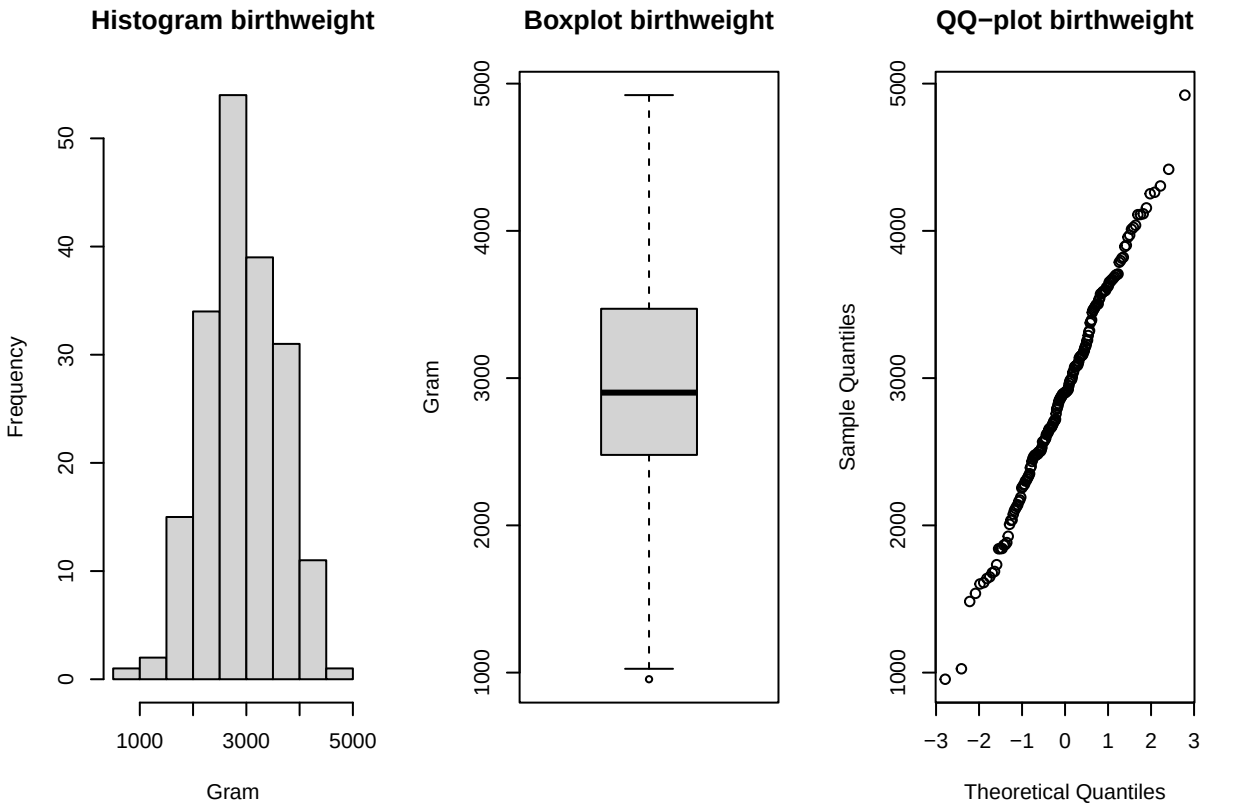
Group 26. Rinus van Grunsven (10755373), Florens Douwes (11254483), Imad Lotfi (14651610)

26 September 2022

Exercise 1.1

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

```
data_read_1 = read.table("./birthweight.txt", header=TRUE)
par(mfrow=c(1,3))
hist(data_read_1$birthweight, main="Histogram birthweight", xlab="Gram")
boxplot(data_read_1$birthweight, main="Boxplot birthweight", ylab="Gram")
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 Shapiro-Wilk test is a test for normality when the distribution P is unknown. That is the case right now. The hypotheses are: H0: P is a normal distribution. H1: P is not a normal distribution.

```
shapiro.test(data_read_1$birthweight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_read_1$birthweight
## W = 0.99595, p-value = 0.8995
```

The Shapiro-Wilk test is always a left-sided test (see lecture 3, slide 10) and H0 is rejected for small values of W. It now is 0.996, which is not small, so we will not reject H0. That means that the data is normally distributed.

The mean, rounded to three decimals

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

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

b)

We use the standard formula $\bar{X} \pm t_{a/2,n-1}\frac{s}{\sqrt{n}}$.

The length is:

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

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

The standard deviation is:

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

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

The test statistic is:

```
t = round(qt(0.95,df=n-1),3); t
```

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

The confidence interval is then calculated with:

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

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

c)

We start with the following null-hypothesis and alternative hypothesis:

$$H_0 : \text{mean birthweight} \leq 2800$$

$$H_0 : \text{mean birthweight} > 2800$$

Then, we can use a t-test. We are going to use a t-test since we concluded in question a) that the data looks to be normally distributed and the standard deviation is unknown (see slide 22, lecture 2).

```
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 H_0 and thus to state that H_1 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.

```r
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:

$$\hat{p} - E < p < \hat{p} + E \text{ where } E = z_{a/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where p = population proportion, $\hat{p}$ = sample proportion, n = number of sample values, E = margin of error, and $z_{a/2}$ is the z-score.

Using the following R code we can get the results of this formula:

```r
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 $H_1 : p \neq 0.75$.

```r
binom.test(x, n, p = 0.75)
```

```
##
##  Exact binomial test
##
## data:  x and n
## 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.

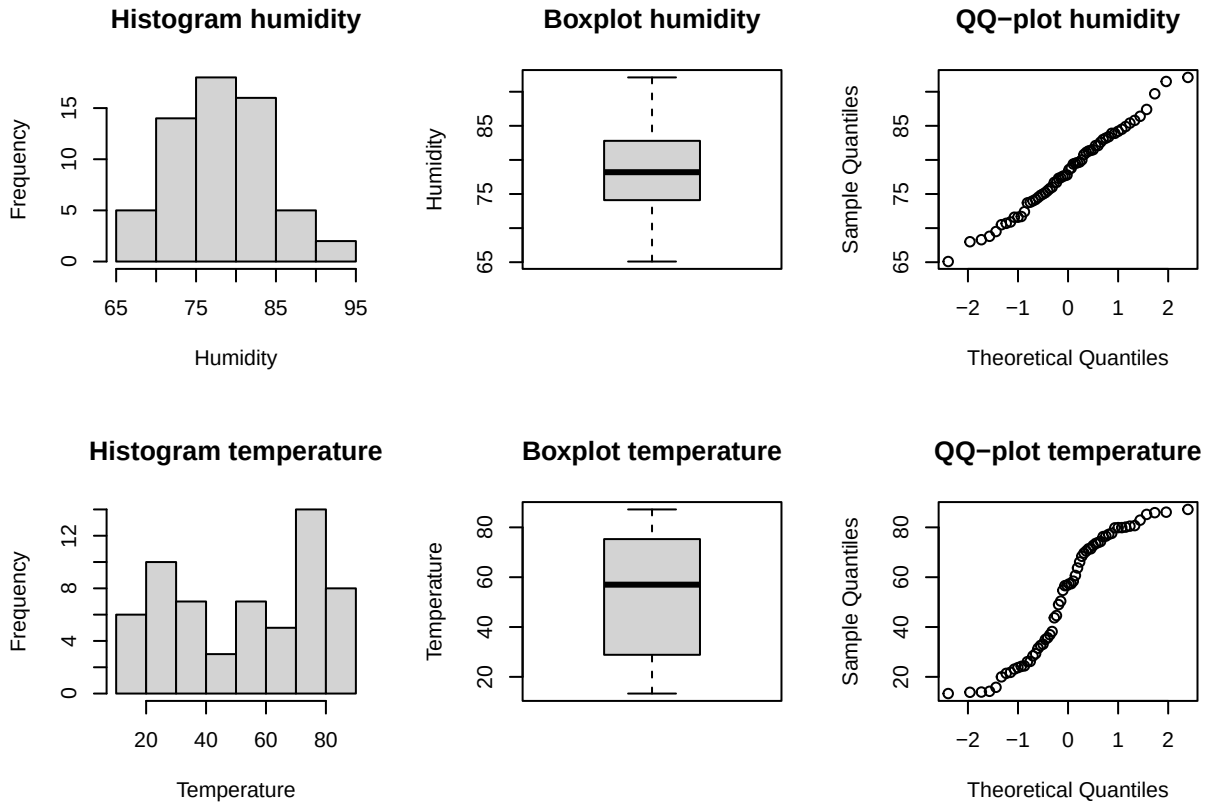
Exercise 1.3

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", xlab="Humidity")
boxplot(hum, main="Boxplot humidity", ylab="Humidity")
qqnorm(hum, main="QQ-plot humidity")
hist(tem, main="Histogram temperature", xlab="Temperature")
boxplot(tem, main="Boxplot temperature", ylab="Temperature")
qqnorm(tem, main="QQ-plot temperature")
```



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

```
summary(weather)
```

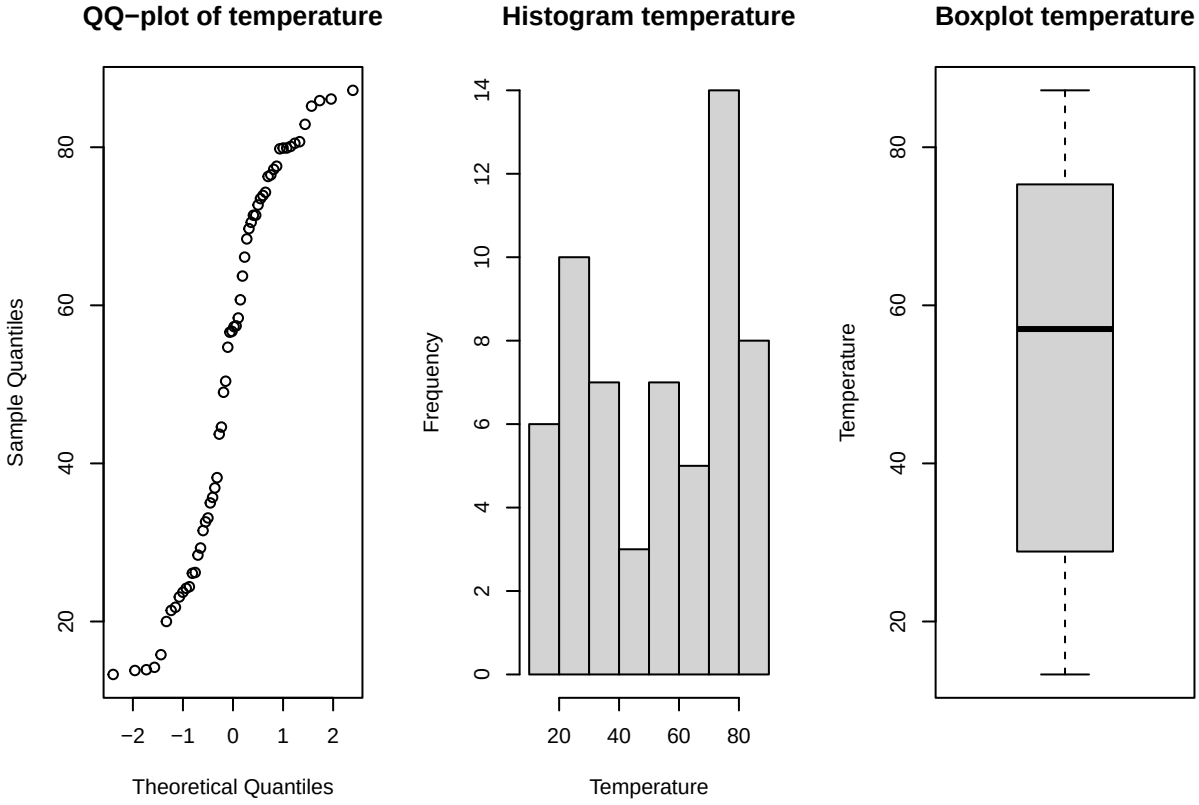
```
##     humidity      temperature
##  Min.   :65.10   Min.   :13.30
```

```
##  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 plots of the temperature:

```
par(mfrow=c(1,3))
qqnorm(tem, main="QQ-plot of temperature")
hist(tem, main="Histogram temperature", xlab="Temperature")
boxplot(tem, main="Boxplot temperature", ylab="Temperature")
```



With the QQ-plot we can see that the temperature data does not follow a pattern seen with normal distributed samples; it is not a straight line. It looks like the data is too spread out, meaning fat tails in the normal distribution. In the histogram this can be seen more clearly, data is spreak out around the 23 degree mark and the 70 degree mark. And as last, the boxplot is skewed.

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.

```r
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. This means that if this experiment would be repeated multiple times, in 90% of the cases the mean would be between those values. It would not mean that we know with a confidence level of 90% that the mean is between those values.

d)

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

The confidence interval margin must then be below 2. We can rewrite the standard formula (lecture 2 slide 14):

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

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 interpretation is that Sense and Emma are 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.

```
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 0.1609 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.

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 process is also described in the book Mathematical Statistics and Data Analysis by John A. Rice. We can do that with the following R code, where we sum the other three authors word count's with the imitator's:

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

```
##         Combined Sand2
## a            434    83
## an            62    19
## this          86    15
## that         240    41
## with         163    39
## without       40     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 = 6.7959, df = 5, p-value = 0.2363
```

The p-value is 0.2363 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 both columns. As a result of this, we conclude that the admirer was successful in imitating Austen's writing style.