Assignment 1

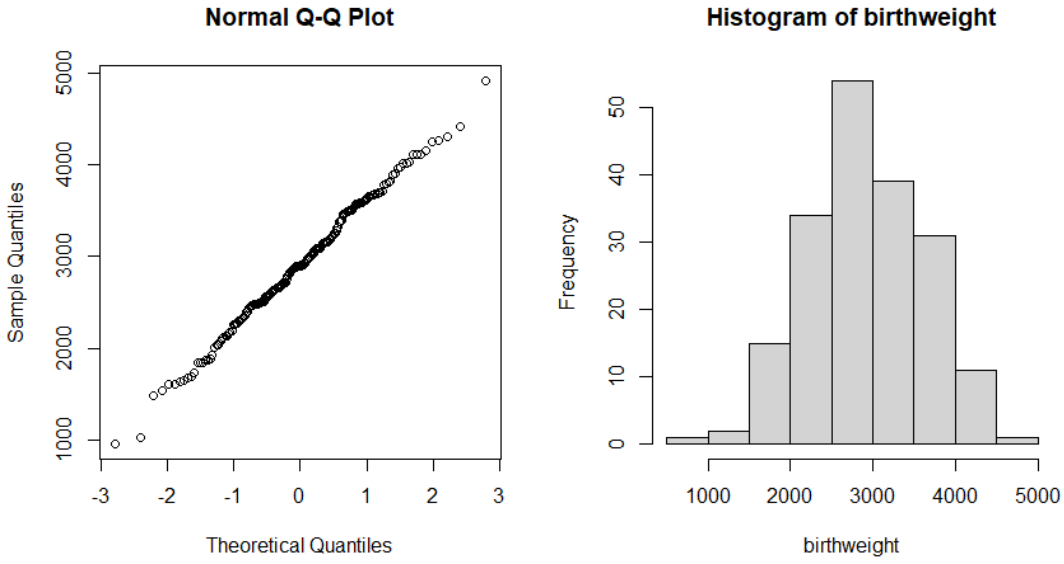
Teun Zwier [redacted], Jelle van der Schoot [redacted], Dragos Pop [redacted]
Group 54

September 2021

Exercise 1.1 - Birthweigths

a) To check the normality of the data we use both a graphical and an algorithmic check. Graphical Check = QQplot and histogram.

```
> birthweight = birthweight$birthweight;
> qqnorm(birthweight) ; hist(birthweight)
```



The QQplot is roughly a straight line, without large skew or heavy tails, showing the sample quantiles are close to the theoretical quantiles of a normal distribution. This suggests the data to be normally distributed. The histogram is roughly bell shaped, which too suggests the data is normally distributed.

Algorithmic Check = Shapiro Wilk Test

H0 : Birthweight data is a normal distribution.
H1 : Birthweight data is not a normal distribution.

```
> shapiro.test(birthweight)
   data:  birthweight
   W = 0.99595, p-value = 0.8995
```

The P-value is > 0.05. Therefore, the null hypothesis that the birthweight data is normally distributed is not rejected. To conclude, both the graphical and algorithmic check suggest that the data is normally distributed.

Point estimate of μ = the mean of the data.

```
> mean(birthweight)
[1] 2913.293
```

b) If the confidence interval is 90%, then 1- α = 0.9.
α = 0.1 and $\alpha/2$ = 0.05.
The standard deviation σ is unknown, so we estimate σ with s of the data and t-distribution is utilized with critical value $t_{0.05}$.

$$CI = \overline{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = \left[\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

```
> n = length(birthweight)  = 188
> m = mean(birthweight)    = 2913.29
> s = sd(birthweight)      = 697.50
> t = qt(0.95,df=n-1)      = 1.65
> c(m-t*s/sqrt(n),m+t*s/sqrt(n))
[1] 2829.202 2997.384
```

90% Confidence Interval = [2829.20, 2997,38] . We are 90% confident that mean μ is between 2829,20 and 2997,38.

c) Expert claim: the mean birthweight is greater than 2800. This claim is assumed to be true if its inverse is rejected, so the null hypothesis is that the mean birthweight is equal or less than 2800.
Null Hypothesis = H0 : $\mu \leq \mu 0 = 2800$
Alternative Hypothesis = H1 : $\mu > \mu 0$

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

One Sample t-test

data:  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 < 0.05, so the null hypothesis is rejected. Following this, the alternative hypothesis -- the expert claim that the mean birthweight > 2800 -- is assumed to be true.

d) It is different from the confidence interval found in b), because the confidence interval of c) doesn't have an upper bound but goes to infinity. The confidence interval found in b) gives a 90% confidence interval for the mean, while the t test gives a 95% confidence interval for the alternative hypothesis, namely mean > 2800. The second CI is one-sided,because only the lower bound of the confidence interval needs to have a limit. If the lower bound of the 95% CI > 2800, then any mean above the lower bound will satisfy the alternative hypothesis.

Exercise 1.2 - Kinderopvangtoeslag

a) We can define *x* as the number of parents receiving child care benefits and *n* as the sample size. The point estimation of p is: $\hat{p} = x/n = 140/200 = 0.7$

b) 99% confidence, therefore 1 - $\alpha = 0.99$, so $\alpha/2 = 0.005$. $\hat{q} = 1 - \hat{p} = 0.3$

$$\text{CI} = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.7 \pm z_{0.005}\sqrt{\frac{0.7*0.3}{200}}$$

```
> z = qnorm(0.995) =  2.575829
0.7 - z * sqrt((0.7*0.3)/200);0.7 + z * sqrt((0.7*0.3)/200)
[1] 0.6165336;[1] 0.7834664
```

99% Confidence Interval = [0.62, 0.78]

c) H0: $p = 0.75$
H1: $p \neq 0.75$

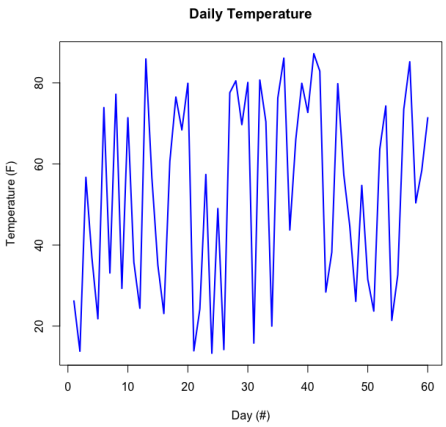
```
> binom.test(140, n=200, p=0.75)
[skipped output]
 p-value = 0.103
95 percent confidence interval:
 0.6313501 0.7626104
```

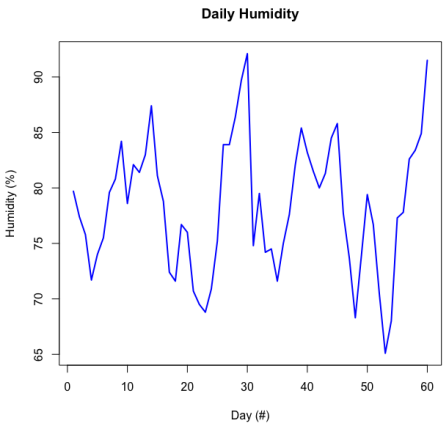
$\alpha = 0.1$ so p-value has to be < 0.1. P-value = 0.103, so for $\alpha = 0.1$, the null hypothesis is not rejected. For different α-values: the higher the α is, the smaller the confidence interval. At $\alpha = 0.125$, the upper bound of the CI < 0.75. So any $\alpha \geq 0.125$ will result in the null hypothesis being rejected.

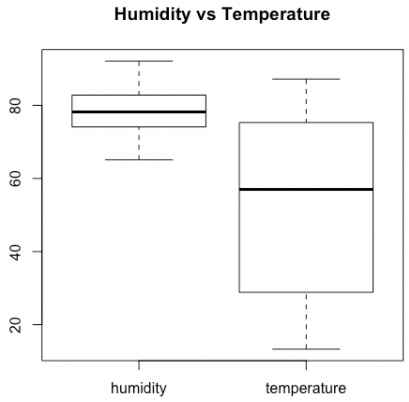
Exercise 1.3 - Weather

a) The dataset contains the humidity and temperature for a period of 60 days and can be summarized with `summary(), var(), plot() and boxplot()` as follows:

	Humidity (%)	Temperature (^{o}F)
Min	65.10	13.30
1st Qu.	74.15	29.07
Median	78.20	57.00
Mean	78.34	52.73
3rd Qu.	82.70	74.80
Max	92.10	87.20
Var	36.30	589.48





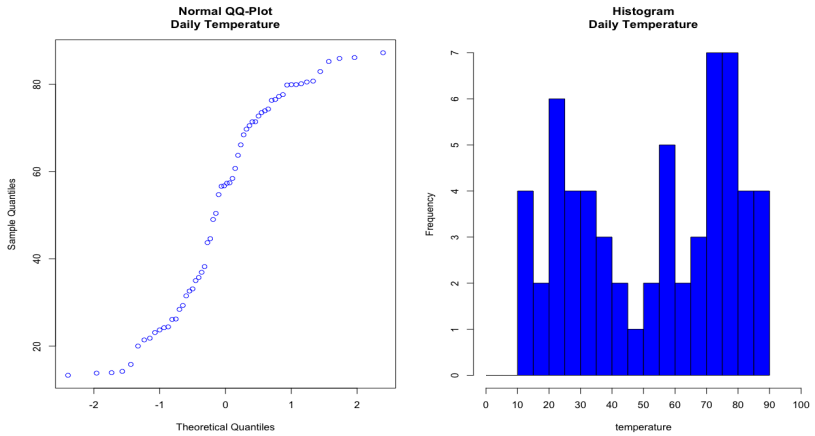


Looking at the boxplot on the right, one can observe the larger range in temperatures in comparison to humidity. Also, it is visible that the humidity is symmetric while the temperature is slightly negatively (left) skewed. Another thing worth mentioning is the fact that the humidity is much more constant over time compared to the temperature, as reflected in the line plots above.

b) The Normal QQ-plot and Histogram are useful tools to investigate the normality of data:

```
qqnorm(temperature,main="Normal QQ-Plot\nDaily Temperature",col="blue")
hist(temperature,main="Histogram\nDaily Temperature",col="blue",xaxp=c(0,100,10),breaks=seq(0,90,5),xlim=c(0,100))
```

As it can be seen in the graphs below, the temperature is slightly light tailed but could be normally distributed when looking at the Normal QQ-plot, but the histogram doesn't support this hypothesis: there is no clear bell curve present, which would be expected if the data was drawn from a normally distributed population. Therefore, the temperature data does not appear to be normally distributed.



c) Same as in 1.1a, we use an estimate of σ and the t-distribution with the critical value $t_{0.05}$.

If the confidence interval is 90%, then 1- α = 0.9. α = 0.1 and α/2 = 0.05.
The standard deviation σ is unknown, so we estimate σ with s of the data and t-distribution is utilized with critical value $t_{0.05}$.

$$\text{CI} = \overline{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = \left[\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

```
n=length(temperature)
m=mean(temperature)
s=sd(temperature)
t=qt(0.95,df=n-1)
c(m-t*s/sqrt(n),m+t*s/sqrt(n))
```

This results in the interval [47.5, 58.0]: we can be 90% confident that the population mean is between 47.5 and 58.0.

d) The minimum sample size must satisfy $n \geq \frac{(z_{a/2})^2 s^2}{E^2}$. An error of at most E is the same as having a CI length of at most 2E. The confidence interval can have at most length 2%, so our E = 1. Because the standard deviation σ is unknown, we estimate it by s and calculate the z-score using qnorm().

```
s=sd(humidity)
z = qnorm(0.975)
```

We can then calculate the minimum number of samples using:

```
min_length = (z^2 * s^2) / E^2
```

This results in a minimum sample size of 139.46, so we should include at least 140 days to have a confidence interval of at most length 2%.

Exercise 1.4 - Jane Austen

a) As the four novels can be viewed as independent samples, a contingency table for homogeneity is most appropriate.

b) In order to check if Austen was consistent in her different novels, the following hypotheses were designed:
H0: Austen's novels are homogenous (i.e. probabilities of having a, an, this, that, with, without are the same for each novel)
H1: Austen's novels are not homogenous
First, the table of expected values was generated:

```
> chisq.test(data[,-4])$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
```

Given that all the expected values $E_{i,j} = \frac{O_{i.} O_{.j}}{n}$ in the table are above 5, the condition (at least 80% of the $E_{i,j}$ should be at least 5) is satisfied. Accordingly, the distribution of the test statistic $X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$ is approximately the chi-squared-distribution with (3 − 1)(6 − 1)=10 degrees of freedom, hence, the Chi-squared test can be applied. The p-value following the Chi-squared test is 0.16, which is above the significance level of 0.05.

```
> chisq.test(data[,-4])
X-squared = 14.274, df = 10, p-value = 0.1609
```

This indicates that there is not enough evidence to reject H0. Consequently, there is not enough evidence to say that Austen was inconsistent in her novels.

c) In order to check if the admirer was successful in imitating Austen's style, the following hypotheses were designed:
H0: Austen's and admirer's novels are homogenous (i.e. probabilities of having a, an, this, that, with, without is the same for each novel)
H1: Austen's and admirer's novels are not homogenous
First, the table of expected values was generated:

```
> chisq.test(data)$expected
              Sense      Emma     Sand1     Sand2
a       161.08809 185.54649 85.604405 84.761011
an       25.23817  29.07015 13.411909 13.279772
this     31.46982  36.24796 16.723491 16.558728
that     87.55465 100.84829 46.527732 46.069331
with     62.93964  72.49592 33.446982 33.117455
without  13.70962  15.79119  7.285481  7.213703
```

Given that all the expected values $E_{i,j} = \frac{O_{i.} O_{.j}}{n}$ in the table are above 5, the condition (at least 80% of the $E_{i,j}$ should be at least 5) is satisfied. Accordingly, the distribution of the test statistic $X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$ is approximately the chi-squared-distribution with (4 − 1)(6 − 1)=15 degrees of freedom, hence, the Chi-squared test can be applied. The p-value following the Chi-squared test is 0.12, which is above the significance level of 0.05.

```
> chisq.test(data)
X-squared = 21.528, df = 15, p-value = 0.1208
```

This indicates that there is not enough evidence to reject H0. Consequently, there is not enough evidence to say that the admirer was not successful in imitating Austen's style.