Assignment 1

Group 15: Ivona Bîrlad, Kaiyi Wang, Matteo Rapa

9/21/2022

Exercise 1.1 Birthweight

a) The normality of the data is firstly checked graphically by drawing the histogram and Q-Q plot(Figure 1). Furthermore, performing the Shapiro–Wilk test serves as a second assessment of normality.

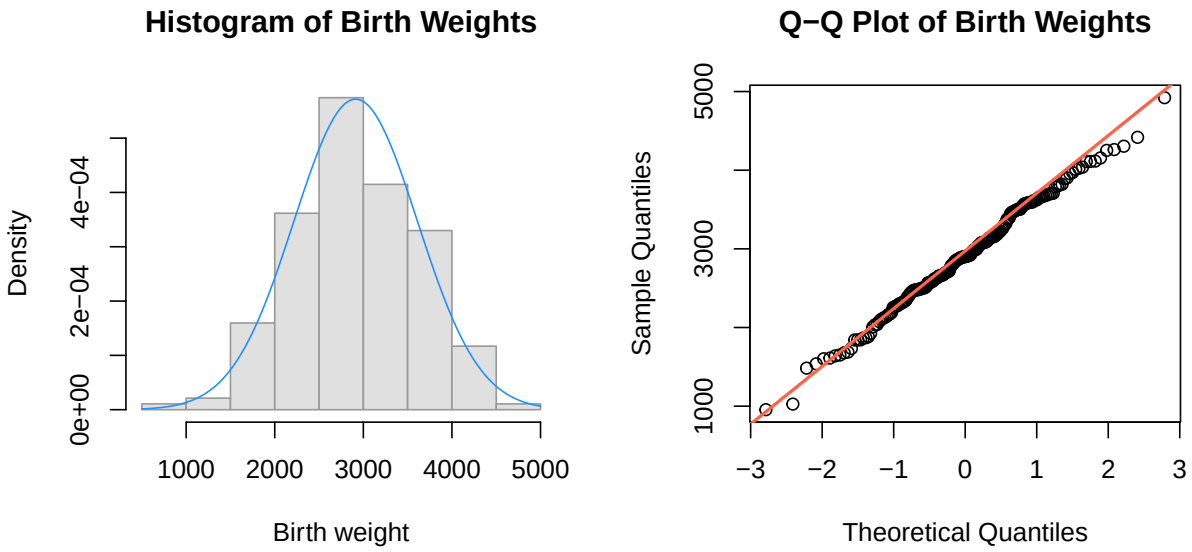

Figure 1: Data Visualization & Normality Exploration

Note: R code for plotting this figure are presented in Appendix 1.1

Comments: The histogram shows that the birth weight data approximately distributes in a bell curve shape. The Q-Q plot paints a similar picture, the theoretical quantiles laying on a straight line. The above plots indicate that the data appears to be normally distributed. To confirm this, a Shapiro test has been conducted as follows.

```
shapiro=shapiro.test(birthweight)
```

As the p-value is 0.9 and thus above $\alpha = 0.05$, H_0 (the data is normally distributed) is accepted. A justification has been made that the data is normally distributed both graphically and statistically.

```r
bw_mean = mean(birthweight); bw_mean
```

```
## [1] 2913.293
```

Comments: The point estimation for μ is 2913.293.

b) We compute the 90% t-confidence interval:

```r
bw_number = length(birthweight) # the sample size
bw_sd = sd(birthweight) # the sample standard deviation
bw_t = qt(0.95, df = bw_number-1) # the estimated t-value
c(bw_mean - bw_t*bw_sd/sqrt(bw_number),
  bw_mean + bw_t*bw_sd/sqrt(bw_number)) # calculate and display the CI
```

```
## [1] 2829.202 2997.384
```

Comments: The 90% confidence interval (CI) of μ is: [2829.202, 2997.384]. Thus, it can be said with 90% confidence that the interval [2829.202, 2997.384] contains the actual population birth weight μ.

c) We construct the following hypotheses:

$$\begin{cases} H_0 : \mu \leq 2800 \\ H_a : \mu > 2800 \end{cases}$$

```r
ttest = t.test(birthweight, mu = 2800, alt = "g") # One sample t-test
ttest[3]
```

```
## $p.value
## [1] 0.0136
```

Comments: The resulting p-value is 0.014 and thus smaller than $\alpha = 0.05$, so we can reject H_0. Therefore, there is enough statistically significant evidence to suggest that the true mean of birth weight is greater than 2800.

d) Comments: the R-output of the test from question b), indicates that we are 90% confident that the interval from 2829.202 to 2997.384 actually **does contain the true value of the population mean** μ . This CI is double-sided. In question c), the confidence interval in the t-test output shows that the [2829.202, inf] interval has 95% confidence to cover the **true mean difference** between the sample mean and 2800, which is entirely different from b).

In order to verify the claim that the mean birth weight is higher than 2800, a single-sided t-test is conducted for c). In the comparison of the means, the result is either greater or lower than a certain value, not two-tailed (H0: $\mu <= 2800$; H1: $\mu > 2800$ in this exercise). As a result, this is a one-sided CI.

Exercise 1.2 Kinderopvangtoeslag

a) We calculate the point estimate of the proportion (p) of the working parents receiving childcare benefits.

```
childcare_p = 140 / 200; childcare_p # point estimate for p
```

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

Comments: the point estimate for p is 0.7. From the data, 70% of the working parents are receiving this childcare benefit.

b) We compute the 99% confidence interval for the proportion p:

```
childcare_q = 1 - childcare_p # calculate q
childcare_n = 200 # sample size
childcare_z = qnorm(1 - 0.01/2) # calculate the z value
# calculate the 99% CI of childcare_p
round(c(childcare_p - childcare_z * sqrt(childcare_p * childcare_q / childcare_n),
  childcare_p + childcare_z * sqrt(childcare_p * childcare_q / childcare_n)), 3)
```

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

Comments: The 99% CI of p is: [0.617, 0.783]. The result shows that the CI of [0.617, 0.783] covers the true proportion of the childcare benefit with 99% confidence.

c) We formulate the following hypotheses for the two-tailed test:

$$\begin{cases} H_0 : p_0 = 75 \\ H_a : p_1 \neq 75 \end{cases}$$

Table 1 P-values by different significant levels

α	0.2	0.1	0.05	0.01
P-value	0.103	0.103	0.103	0.103

Note: R codes of the binomial tests are presented in Appendix 1.2

Comments: According to the binomial test at a significance level of $\alpha = 0.1$, the p-value is calculated as 0.103. Further tests have been conducted using different α (see Table 1). The table shows that the p-value doesn't change by different CI.

Firstly, this may relate to the calculation method of test statistics:

$$T = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

In this formula, α is not contained in the calculation. Secondly, α is the significance level of the test defined by the researchers. It is used to be compared with the p-value. Based on the real-world problem, different α could be selected. This may lead to different conclusions of the test, but not affect the p-value itself.

Exercise 1.3 Weather

a) We compute the summary statistics and summarize the data graphically by plotting the box and scatter plots of the variables.

Table 2 Summary of the Weather Data

Statistics	Humidity(%)	Temperature(°F)
Mean	78.343	52.725
Median	78.200	57.000
Min	65.100	13.300
Max	92.100	87.200
Standard deviation	6.025	24.279

```
par(mfrow= c(1,3))
# box plot of the humidity data
boxplot(humidity, main = "Humidity Boxplot", xlab = "humidity",
        ylab = "Percentage", col="dodgerblue", ylim =c(60, 95))
# box plot of the temperature data
boxplot(temperature, main = "Temperature Boxplot", xlab = "temperature",
        ylab ="fahrenheit", col="tomato", ylim =c(0, 100))
# scatter plot of humidity vs temperature data
plot(humidity, temperature, main = "Scatter Plot of Temperature vs Humidity")
```

Comments: According to the exploratory data analysis, the dataset contains 60 rows of humidity and temperature. No missing data has been found. The details of the relevant statistics have been summarized in Table 2. For the humidity data, the mean(±sd) is 78.343 (±6.025)%. The minimum humidity is 65.100% while the maximum is 92.100%. For the temperature data, the mean(±sd) is 52.725(±24.279) °F. The minimum temperature is 13.300°F while the temperature is 87.200°F.

The humidity and temperature data are further visualized in box plots (see Figure 2). No outlier was identified in both data. A scatter plot investigating the correlation between the humidity and temperature data is also displayed. There is no clear pattern shown in the scatter plot, although a weak positive correlation is found in statistics (covariance=0.284)

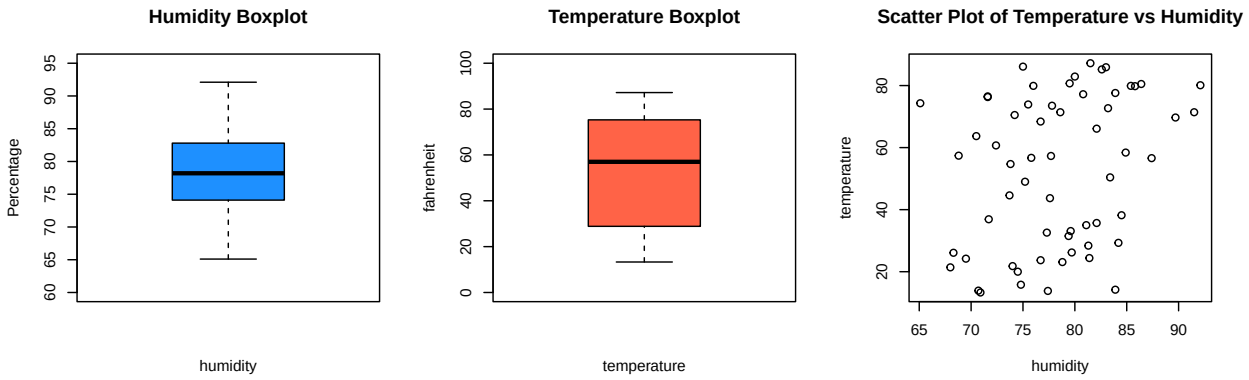

Figure 2: Exploratory Data Analysis of the Weather Data

b) In order to investigate the normality of the humidity and temperature data, histograms and Q-Q plots have been produced in Figure 3 and 4 (R codes of the figures are presented in Appendix 1.3)

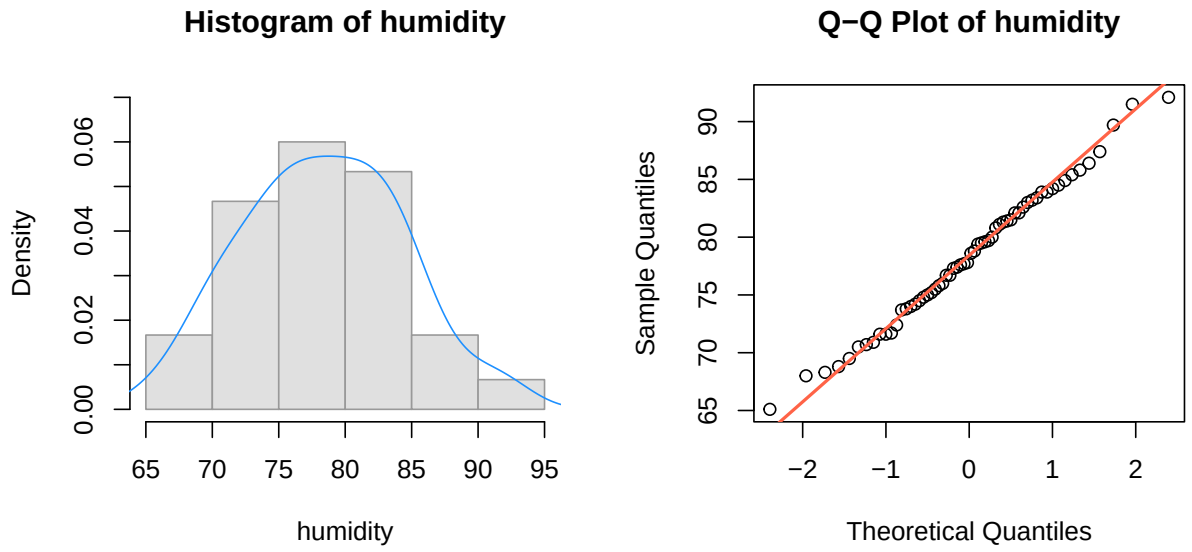

Figure 3: Normality Visualisation of Humidity Data

Comments: The histogram shows that the humidity data approximately distributes in a bell curve shape. The theoretical quantiles lay on a straight line in the Q-Q plot. The data visualizations indicate that the humidity data seems normally distributed.

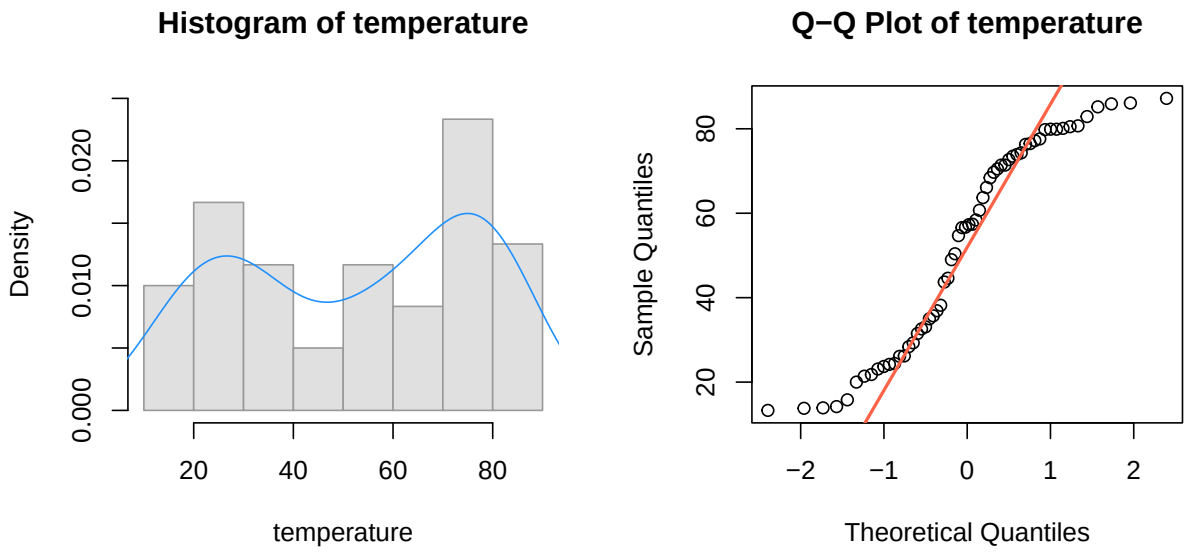

Figure 4: Normality Visualisation of the Temperature Data

Comments: The histogram shows that the temperature data approximately distributes in a saddle curve shape. From the Q-Q plot, the theoretical quantiles lay on an 'S' shape, which is definitely not straight-like. The data visualizations indicate that the temperature data does not distribute normally.

c) We compute the give a 90% confidence interval for the mean temperature:

```
# the sample mean is used for point estimate for the data
tem_mean = mean(temperature) # sample mean
tem_number = length(temperature) # sample size
tem_sd = sd(temperature) # sample sd
tem_t = qt(0.95, df = tem_number-1) # calculate t value
round(c(tem_mean - tem_t*tem_sd/sqrt(tem_number),
  tem_mean + tem_t*tem_sd/sqrt(tem_number)), 3) # calculate temperature CI
```

```
## [1] 47.487 57.963
```

Comments: the R-output of the test from b), indicates that we are 90% confident that the interval from 47.487 to 57.963 actually does contain the true value of the temperature mean μ.

d) We derive the minimum sample size for the mean humidity to fall within the 95% confidence interval by using the following formula:

$$n \geq \frac{t^2_{\alpha/2} s^2}{E^2}$$

Given that the confidence interval has at most length 2%, this means that 2E = 0.02 and, thus, the margin of error is E = 0.01.

```
hum_mean = mean(humidity) # sample mean
hum_number = length(humidity) # sample size
hum_sd = sd(humidity) # sample sd
hum_t = qt(0.975, df = hum_number-1) # calculate t value
E = 0.01*100 #Ensuring the maximum margin of error is in percentage form
round(hum_t^2 *  hum_sd^2 / (E^2) , 3)# calculate minimum sample size
```

```
## [1] 145.36
```

Comments: According to the calculation, we may have to include 146 humidity samples that the CI has at most length of 2%, with 95% confidence.

Note for the calculation: the measurement of humidity is "percentage %". To satisfy the CI length is 2%, the unit in calculation should be consistent with the original data.

Exercise 1.4 Jane Austen

a) We are given the following data:

```
austen = read.table(file = "austen.txt", header = TRUE); austen # display data
```

```
##         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
```

Comments: For this dataset, we could use a contingency table test homogeneity since:

- The data is presented in a 6 * 4 table, and no missing data has been found.
- In order to conduct quantitative studies on literary styles, the distribution of different words(rows) on different chapters(columns) could be investigated through contingency table test.
- Using the contingency table to test homogeneity is most appropriate here since we are interested in understanding how the distribution over rows varies from column to column
- When conducting the statistical tests, no warning message is shown in R. This indicates that at least 80% of the Eij 's are at least 5, which fulfils the test condition.

b) We check for the consistency of Austen's writing using the Chi Square test:

```r
# The first three columns are Jane Austen's novels
chisq.test(austen[,1:3])[3]
```

```
## $p.value
## [1] 0.161
```

Comments: The H_0 cannot be rejected at 95% significant level (p = 0.161). The distributions of the selected word counts over rows are the same among different works written by Jane Austen. This may indicate that Austen herself was consistent in her different novels.

c) We use a Chi^2 test over the entire dataset to verify if the writing style of Austen's admirer was consistent with her own:

```r
# a chi-sq test for all columns including Jane Austin and the admirer's novels
chisq.test(austen)[3]
```

```
## $p.value
## [1] 0.121
```

Comments: The H_0 still cannot be rejected, as p-value > 0.05. The distributions of the selected words over rows are equal from different work written by Jane Austen and her admirer. This may indicate that the admirer has successfully imitated Austen's writing style.

Appendix The Supplementary R Codes

```r
#load data
raw_data_birthweight = read.table(file = "birthweight.txt", header = TRUE)
#check if the data is correctly loaded
head(raw_data_birthweight)
birthweight = raw_data_birthweight[, 1]
```

```r
# histogram
par(mfrow= c(1,2))
hist(birthweight, main = "Histogram of Birth Weights", xlab = "Birth weight",
     col="gray88", border="gray60", freq = FALSE)
# Add a density curve to the histogram
curve(dnorm(x,mean=mean(birthweight),sd=sd(birthweight)), add=TRUE,col="dodgerblue")
# QQ Plot
qqnorm(birthweight, main = "Q-Q Plot of Birth Weights")
qqline(birthweight, col = "tomato", lwd = 2)
```

1.1

1.2

```
# binomial tests of different significance level
bi_0.9 = binom.test(140, 200, p = 0.75,conf.level = 0.9); bi_0.9[3]
bi_0.95 = binom.test(140, 200, p = 0.75,conf.level = 0.95); bi_0.95[3]
bi_0.99 = binom.test(140, 200, p = 0.75,conf.level = 0.99); bi_0.99[3]
bi_0.8 = binom.test(140, 200, p = 0.75,conf.level = 0.8); bi_0.8[3]
```

1.3

```
raw_data_weather = read.table(file = "weather.txt", header = TRUE) #load data
head(raw_data_weather) #check if the data is correctly loaded
humidity = raw_data_weather$humidity # assign variables
temperature = raw_data_weather$temperature # assign variables
```

```
summary(humidity) # summary humidity data
summary(temperature) # summary temperature data
length(temperature) # get the sample size
sd(humidity) # compute the sd
sd(temperature) # compute the sd
```

```
# histogram
par(mfrow= c(1,2))
hist(humidity, main = "Histogram of humidity", xlab = "humidity",
     col="gray88", border="gray60", freq = FALSE, ylim = c(0, 0.07))
lines(x = density(x = humidity), col = "dodgerblue")
# QQ Plot
qqnorm(humidity, main = "Q-Q Plot of humidity")
qqline(humidity, col = "tomato", lwd = 2)
```

```
# histogram
par(mfrow= c(1,2))
hist(temperature, main = "Histogram of temperature", xlab = "temperature",
     col="gray88", border="gray60", freq = FALSE, ylim = c(0, 0.025))
lines(x = density(x = temperature), col = "dodgerblue")
# QQ Plot
qqnorm(temperature, main = "Q-Q Plot of temperature")
qqline(temperature, col = "tomato", lwd = 2)
```