Assignment 1

Group 15: Ivona Bîrlad, Kaiyi Wang, Matteo Rapa

9/18/2022

Exercise 1.1 Birthweight

a)

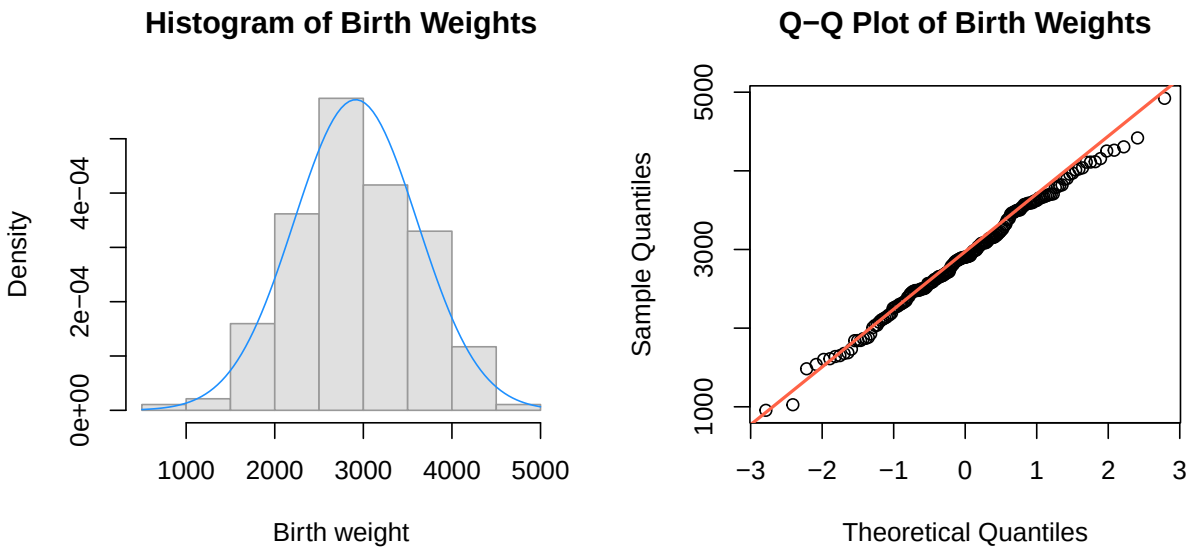

Figure 1: Data Visualization & Normality Exploration

(Note: R codes of this figure are presented in Appendix 1.1)

Comments: The histogram shows that the birth weight data approximately distributes in a bell curve shape. And from the Q-Q plot, the theoretical quantiles lay on a straight line. The data visualizations indicate that the data seems normally distributed. To confirm this, a shapiro test has been conducted as follows.

```
shapiro.test(birthweight) # shapiro test for the normality of the data
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  birthweight
## W = 0.99595, p-value = 0.8995
```

As the p-value is bigger than 0.05, H0 (the data is not normally distributed) is rejected. A justification has been made that the data is normally distributed both graphically and statistically.

```
bw_mean = mean(birthweight); bw_mean # the sample mean is for point estimation
```

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

Comments: The point estimation for μ is 2913.293.

b)

```
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: [2829.20, 2997.38]. We could interpret the result that the CI [2829.20, 2997.38] has 90% confidence in containing the actual population birth weight μ.

c)

```
t.test(birthweight, mu = 2800, alt = "g") # One sample t-test
```

```
##
##  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
```

Comments: P-value is smaller than 0.05, and here we reject H0. The true mean of birth weight is significantly greater than 2800 statistically ($\alpha = 0.05$).

d)

Comments: the R-output of the test from question b), indicates that we are 90% confident that the interval from 2829.20 to 2997.38 actually **does contain the true value of the population mean** μ . This CI is double side. 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 totally different from b).

In order to verify the claim that the mean birth weight is bigger than 2800, a single-sided t-test is conducted for c). In the comparison of the means, the result is either bigger or smaller than a certain value, not on both sides (H0: $\mu <= 2800$; H1: $\mu > 2800$ in this exercise). As a result, we have got a one-side CI.

Exercise 1.2 Kinderopvangtoeslag

a)

```
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)

```
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
c(childcare_p - childcare_z * sqrt(childcare_p * childcare_q / childcare_n),
  childcare_p + childcare_z * sqrt(childcare_p * childcare_q / childcare_n))
```

```
## [1] 0.6165336 0.7834664
```

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)

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, the p-value is calculated as 0.103. Further tests have been conducted using different α (see Table 1). The table shows that 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 significant level of the test defined by the researchers. It is used to be compared with the p-value. Based on the actual real world problem, different α could be selected. This may lead to different conclusion of the test, but not affecting p-value itself.

Exercise 1.3 Weather

a)

Table 2 Summary of the Weather Data

Statistics	Humidity(%)	Temperature(°F)
Mean	78.34	52.73
Median	78.20	57.00
Min	65.10	13.30
Max	92.10	87.20
Standard deviation	6.03	24.28

```
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 78.34 (±6.03)%. The minimum

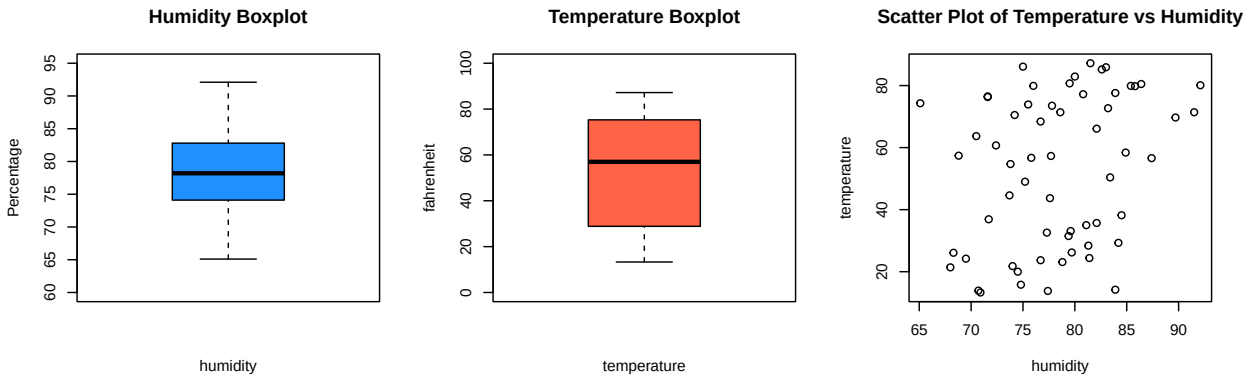

Figure 2: Exploratory Data Analysis of the Weather Data

humidity is 65.10% while the maximum is 92.10%. For the temperature data, the mean(±sd) is 52.73(±24.28) °F. The The minimum temperature is 13.30°F while the temperature is 87.20°F.

The humidity and temperature data are further visualized in box plots (see Figure 2). No outlier 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)

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)

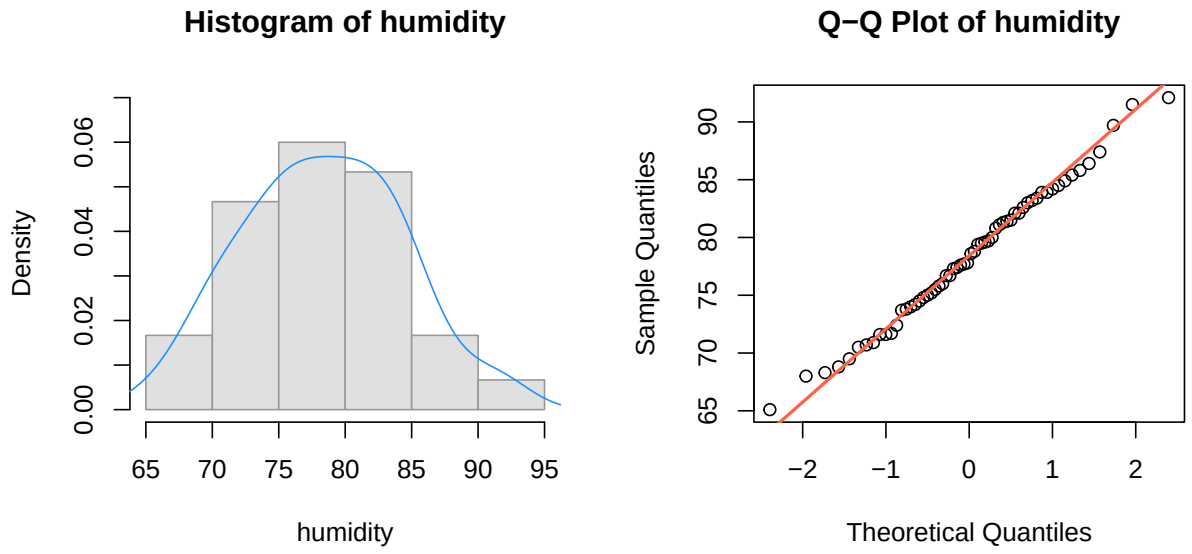

Figure 3: Normality Visualisation of Humidity Data

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.

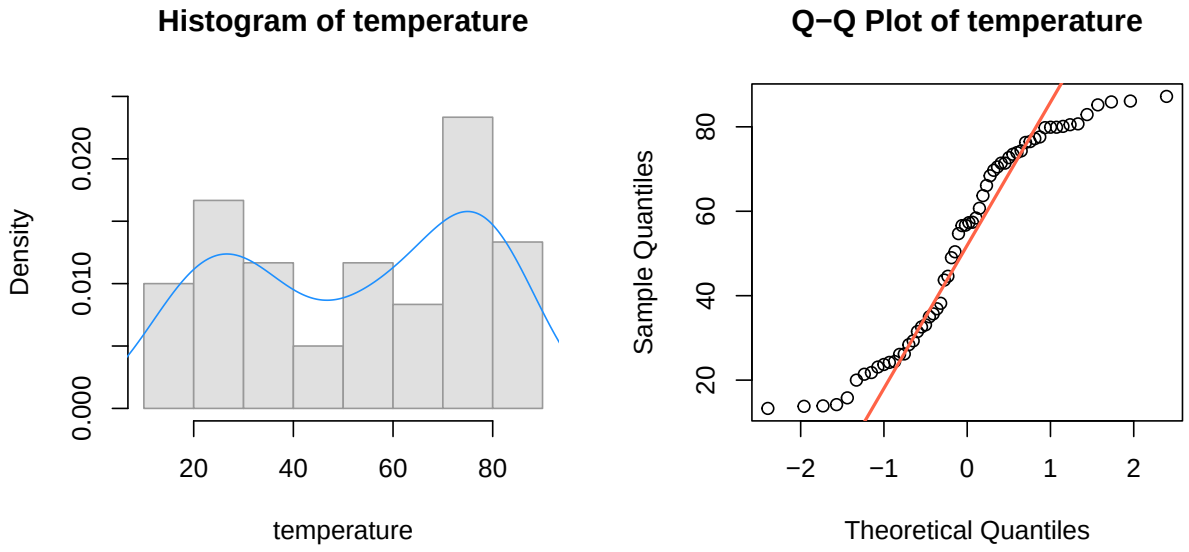

Figure 4: Normality Visualisation of the Temperature Data

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 a 'S' shape, which is definitely not a straight like. The data visualizations indicate that the temperature data does not distributed normally.

c)

```
# 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
c(tem_mean - tem_t*tem_sd/sqrt(tem_number),
  tem_mean + tem_t*tem_sd/sqrt(tem_number)) # calculate temperature CI
```

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

Comments: the R-output of the test from b), indicates that we are 90% confident that the interval from 47.49 to 57.96 actually does contain the true value of the temperature mean μ.

d)

```
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
hum_t^2 *  hum_sd^2 / (1^2) # 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 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

```
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: Yes we could use contingency table test in this exercise. * 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. * When conducting the statistical tests, no warning message is shown in R. This indicates that at least 80% of the Eij 's are least 5, which fullfill the test condition.

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

```
##
##  Pearson's Chi-squared test
##
## data:  austen[, 1:3]
## X-squared = 14.274, df = 10, p-value = 0.1609
```

Comments: The h0 couldn't be rejected at 95% significant level (p = 0.161). The distributions of the selected words over rows are equal from different work written by Jane Austen. This may indicate that Austen herself was consistent in her different novels.

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

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

Comments: The h0 still couldn't 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

1.1

```
#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]
```

```
# 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)
```

1.2

```
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)
```