Technical University of Denmark *Page 1 of 36 pages.*

Written examination: 27 May 2018

Course name and number: **Introduction to Statistics (02402)**

Aids and facilities allowed: All

The questions were answered by

_____ _____ _____

(student number) (signature) (table number)

This exam consists of 30 questions of the "multiple choice" type, which are divided between 11 exercises. To answer the questions, you need to fill in the "multiple choice" form (6 seperate pages) on CampusNet with the numbers of the answers that you believe to be correct.

5 points are given for a correct "multiple choice" answer, and -1 point is given for a wrong answer. ONLY the following 5 answer options are valid: 1, 2, 3, 4, or 5. If a question is left blank or an invalid answer is entered, 0 points are given for the question. Furthermore, if more than one answer option is selected for a single question, which is in fact technically possible in the online system, 0 points are given for the question. The number of points needed to obtain a specific mark or to pass the exam is ultimately determined during censoring.

The final answers should be given by filling in and submitting the form online via CampusNet. The table provided here is ONLY an emergency alternative. Remember to provide your student number if you do hand in on paper.

Exercise	I.1	I.2	I.3	I.4	II.1	II.2	II.3	III.1	IV.1	IV.2
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer	4	5	2	3	1	3	4	2	4	2

Exercise	IV.3	IV.4	IV.5	V.1	V.2	VI.1	VI.2	VI.3	VII.1	VII.2
Question	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Answer	3	2	5	1	4	2	2	3	3	4

Exercise	VII.3	VIII.1	IX.1	IX.2	X.1	X.2	X.3	XI.1	XI.2	XI.3
Question	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Answer	5	4	1	3	2	2	5	2	4	3

The exam paper contains 36 pages.

Continue on page 2

Multiple choice questions: *Note that in each question, one and only one of the answer options is correct. Furthermore, not all the suggested answers are necessarily meaningful. Always remember to round your own result to the number of decimals given in the answer options before you choose your answer.*

Exercise I

In order to investigate the download speed at a work station, download times (in seconds) were recorded for 53 files of different sizes (measured in MB). The download times are saved in the vector `time` in R, while the corresponding file sizes are saved in the vector `size`. Furthermore, the following code was executed in R:

```
logtime <- log(time)
model1 <- lm(logtime ~ size)
summary(model1)

##
## Call:
## lm(formula = logtime ~ size)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.571 -0.673  0.132  0.745  2.190
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.109      0.482   -0.23     0.82
## size           0.127      0.023    5.50  1.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.979 on 51 degrees of freedom
## Multiple R-squared:  0.372,Adjusted R-squared:  0.36
## F-statistic: 30.2 on 1 and 51 DF,  p-value: 1.25e-06
```

The statistical model given by `model1` is a linear regression model of the form

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \,,$$

where ε_i, $i = 1, \ldots, 53$, are assumed to be independent and identically $N(0, \sigma^2)$ distributed.

Question I.1 (1)

Identify the dependent variable and the explanatory variable in the model given by `model1`.

1 ☐ Download time is the dependent variable. File size is the explanatory variable.

2 ☐ File size is the dependent variable. The logarithm of download time is the explanatory variable.

3 ☐ Download time is the explanatory variable. File size is the dependent variable.

4* ☐ File size is the explanatory variable. The logarithm of download time is the dependent variable.

5 ☐ As (the logarithm of) download time depends on file size, both file size and the logarithm of download size are dependent variables. There is no explanatory variable in the model.

---------- FACIT-BEGIN ----------

The variable to the left of the tilde ($\sim$) in the R code corresponds to the dependent variable Y_i, while the variable to the right of the tilde corresponds to the explanatory variable x_i.

---------- FACIT-END ----------

Question I.2 (2)

Give estimates for the parameters of the model given by `model1`.

1 ☐ $\hat{\beta}_0 = 0.127$, $\hat{\beta}_1 = -0.109$, $\hat{\sigma} = 0.979$

2 ☐ $\hat{\beta}_0 = -0.109$, $\hat{\beta}_1 = 0.127$, $\hat{\sigma} = 0.979^2$

3 ☐ $\hat{\beta}_0 = -0.109$, $\hat{\beta}_1 = 0.127$, $\hat{\sigma} = 0.372$

4 ☐ $\hat{\beta}_0 = 0.127$, $\hat{\beta}_1 = -0.109$, $\hat{\sigma} = 0.979^2$

5* ☐ $\hat{\beta}_0 = -0.109$, $\hat{\beta}_1 = 0.127$, $\hat{\sigma} = 0.979$

---------- FACIT-BEGIN ----------

The estimates of the model intercept, $\hat{\beta}_0$, and slope, $\hat{\beta}_1$, are given in the column `Estimate` in the R output (in the rows `(Intercept)` and `size`, respectively). The estimated standard deviation of the error, $\hat{\sigma}$, is termed `Residual standard error` in the R output.

---------- FACIT-END ----------

Question I.3 (3)

The following code was also executed in R:

```r
mean(size)

## [1] 20.088

(53-1)*var(size)

## [1] 1807.6
```

Using the model given by `model1` as a starting point, give a 90% prediction interval for the logarithm of the download time for a file of size 17 MB.

1 ☐ $-0.109 + 0.127 \cdot 17 \pm 2.0076 \cdot 0.979 \cdot \sqrt{1 + \frac{1}{53} + \frac{(17-20.088)^2}{1807.6}}$

2* ☐ $17 \cdot 0.127 - 0.109 \pm 0.979 \cdot 1.6753 \cdot \sqrt{1 + \frac{1}{53} + \frac{(20.088-17)^2}{1807.6}}$

3 ☐ $-0.109 + 0.127 \cdot 17 \pm 1.6753 \cdot \sqrt{0.979} \cdot \sqrt{1 + \frac{1}{53} + \frac{(17-20.088)^2}{1807.6}}$

4 ☐ $-0.109 + 17 \cdot 0.127 \pm 1.6753 \cdot 0.979 \cdot \sqrt{\frac{1}{53} + \frac{(17-20.088)^2}{1807.6}}$

5 ☐ $17 \cdot 0.127 + 0.109 \pm 0.979 \cdot 2.0076 \cdot \sqrt{1 + \frac{1}{53} + \frac{(20.088-17)^2}{1807.6}}$

---------- FACIT-BEGIN ----------

Use Method 5.18, equation (5-60). Note that 1.6753 is the 0.95 quantile of the t-distribution with 51 degrees of freedom, `qt(0.95, df = 51)` in R.

---------- FACIT-END ----------

Question I.4 (4)

Using the model given by `model1` as a starting point, one would like to investigate whether there is a significant linear relationship between the logarithm of download time and file size. Formulate the corresponding statistical null hypothesis that there is no association between the two variables.

1 ☐ $H_0 : \hat{\beta}_1 = 0$

2 ☐ $H_0 : \beta_1 \neq \beta_0$

3* ☐ $H_0 : \beta_1 = 0$

4 ☐ $H_0 : \beta_1 \neq 0$

5 ☐ $H_0 : \hat{\beta}_1 \neq 0$

---------- FACIT-BEGIN ----------

Under the null hypothesis $H_0 : \beta_1 = 0$, the model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ reduces to $Y_i = \beta_0 + \varepsilon$. The latter corresponds to a model for one sample (in which the logarithm of the download time, Y_i, does not depend on file size, x_i). Answer 1 is wrong since we are making a null hypothesis for the true mean and not our estimate.

---------- FACIT-END ----------

Continue on page 6

Exercise II

Train carriages in a mine are to be loaded with large pieces of blasted rock. Initially, a machine has sorted the pieces into two piles based on their size and weight. The weights of the pieces of rock are assumed to be independent and normally distributed, such that the weight of a randomly extracted piece from pile 1 can be represented by $X_1 \sim N(20, 5^2)$ kg and from pile 2 by $X_2 \sim N(50, 10^2)$ kg.

Question II.1 (5)

What is the probability that a randomly selected piece from pile 1 weighs more than 25 kg?

1* ☐ 15.9 %

2 ☐ 84.1 %

3 ☐ 42.1 %

4 ☐ 57.9 %

5 ☐ None of the values listed.

---------- FACIT-BEGIN ----------

Here,

$$P(X_1 > 25) = 1 - P(X_1 \leq 25) = 1 - F_{X_1}(25)$$

is to be computed, where F_{X_1} denotes the cumulative distribution function of the normal distribution with mean 20 and standard deviation 5. In R, the result may be found as:

```
1 - pnorm(q = 25, mean = 20, sd = 5)
```

```
## [1] 0.1586553
```

---------- FACIT-END ----------

Question II.2 (6)

Choose a correct statement:

There is a 20% probability of a randomly selected piece of rock from pile 2 being heavier than

1 ☐ 41.6 kg.

2 ☐ 52.5 kg.

3* ☐ 58.4 kg.

4 ☐ 67.4 kg.

5 ☐ 134 kg.

---------- FACIT-BEGIN ----------

Let F_{X_2} be the cumulative distribution function of the normal distribution with mean 50 and standard deviation 10. Here, we need to find x such that

$$P(X_2 > x) = 0.2\,,$$

which corresponds to finding x such that

$$P(X_2 \leq x) = F_{X_2}(x) = 0.8.$$

In R, x may be found as:

```
qnorm(0.8, mean = 50, sd = 10)
```

```
## [1] 58.41621
```

---------- FACIT-END ----------

Question II.3 (7)

If the train carriages are loaded too heavily, operations must stop. A manual crane must be used to remove pieces of rock from the overloaded carriage, and this procedure is very costly.

Two robotic cranes load the train carriages. One crane takes pieces from pile 1, and the other takes pieces from pile 2. If each crane takes 10 pieces from its pile, and all 20 pieces are loaded into the same (empty) carriage, what is the probability that the total load of this train carriage exceeds 800 kg?

1 ☐ The total weight is $Y \sim N(700, 15625)$, so $P(Y > 800) = 21.2\%$.

2 ☐ The total weight is $Y \sim N(700, 12500)$, so $P(Y > 800) = 18.6\%$.

3 ☐ The total weight is $Y \sim N(700, 2500)$, so $P(Y > 800) = 2.28\%$.

4* ☐ The total weight is $Y \sim N(700, 1250)$, so $P(Y > 800) = 0.234\%$.

5 ☐ The total weight is $Y \sim N(700, 225)$, so $P(Y > 800) = 1.31 \cdot 10^{-9}\%$.

---------- FACIT-BEGIN ----------

Let X_{1i} and X_{2i}, $i = 1, \ldots, 10$, with $X_{1i} \sim N(20, 5^2)$ and $X_{2i} \sim N(50, 10^2)$, be independent random variables which represent the weights of 10 randomly selected pieces of rock from pile 1 and pile 2, respectively. Then, the total load may be represented by

$$Y = \sum_{i=1}^{10} X_{1i} + \sum_{i=1}^{10} X_{2i} \, .$$

According to Theorem 2.40 and Theorem 2.56, Y is normally distributed with

$$\mathrm{E}(Y) = \sum_{i=1}^{10} \mathrm{E}(X_{1i}) + \sum_{i=1}^{10} \mathrm{E}(X_{2i}) = 10 \cdot 20 + 10 \cdot 50 = 700$$
$$\mathrm{V}(Y) = \sum_{i=1}^{10} \mathrm{V}(X_{1i}) + \sum_{i=1}^{10} \mathrm{V}(X_{2i}) = 10 \cdot 5^2 + 10 \cdot 10^2 = 1250$$

so

$$P(Y > 800) = 1 - P(Y \leq 800) = 1 - F_Y(800)$$

where F_Y is the cumulative distribution function for the $N(700, 1250)$ distribution. In R, the result may then be computed as:

```
1 - pnorm(800, mean = 700, sd = sqrt(1250))
```

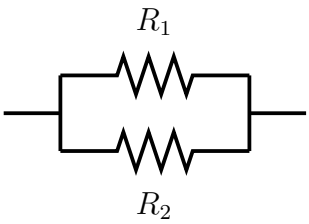
```
## [1] 0.002338867
```

---------- FACIT-END ----------

Continue on page 9

Exercise III

The resistances in the electrical circuit



are estimated to be $\hat{R}_1 = 2\ \Omega$ and $\hat{R}_2 = 3\ \Omega$, with an estimated standard deviation for the estimators (*standard error*) of $\hat{\sigma}_{\hat{R}_1} = 0.2$ and $\hat{\sigma}_{\hat{R}_2} = 0.5$, respectively. R_1 and R_2 can be assumed independent and normally distributed.

The total resistance through the circuit is given by

$$R = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2}}$$

Question III.1 (8)

Using simulation, determine a 95% confidence interval for the total resistance R. The following R code must be used to get the specified result (and after this code has been executed, only one additional function call in R is needed to get the result):

```
set.seed(7643)
k <- 10000
R1 <- rnorm(k, mean = 2, sd = 0.2)
R2 <- rnorm(k, mean = 3, sd = 0.5)
R <- 1/(1/R1 + 1/R2)
```

1 ☐ [1.11, 1.29]

2* ☐ [0.96, 1.40]

3 ☐ [0.92, 1.47]

4 ☐ [0.82, 1.58]

5 ☐ [0.72, 1.68]

---------- FACIT-BEGIN ----------

We have used the parametric bootstrap simulation approach to error propagation as described in Method 4.7:

```r
set.seed(7643)
k <- 10000
R1 <- rnorm(k, mean = 2, sd = 0.2)
R2 <- rnorm(k, mean = 3, sd = 0.5)
R <- 1/(1/R1 + 1/R2)
quantile(R, c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.9647361 1.4016874
```

---------- FACIT-END ----------

Continue on page 11

Exercise IV

Students' choice of higher education has great political awareness. The table below is based on numbers from "Universiteternes Statistiske Beredskab", and contains the number of newly enrolled students by discipline for selected years (row and column sums are included in italics)

	Y2012	Y2016	Y2017	*Sum*
Hum	7966	7297	6691	*21954*
Soc	10173	10253	10006	*30432*
Hlth	2789	3137	3157	*9083*
TechNat	8551	10130	10339	*29020*
Sum	*29479*	*30817*	*30193*	*90489*

Question IV.1 (9)

Based on the numbers for 2017, one wants to test the hypothesis that the proportion of students newly enrolled in a technical or natural science bachelor education (TechNat) is 32.0%. The relevant test statistic for this hypothesis, which is assumed to be well approximated by a standard normal distribution, is

1 ☐ $(10253 - 0.68 \cdot 30817)/\sqrt{30817 \cdot 0.32 \cdot 0.68} = -130.70$

2 ☐ $(10339 - 0.32 \cdot 30193)/\sqrt{30193 \cdot 0.32 \cdot 0.32} = 12.18$

3 ☐ $(10130 - 0.32 \cdot 30817)/\sqrt{30817 \cdot 0.32 \cdot 0.68} = 3.28$

4* ☐ $(10339 - 0.32 \cdot 30193)/\sqrt{30193 \cdot 0.32 \cdot 0.68} = 8.36$

5 ☐ $(10339 - 0.68 \cdot 30193)/\sqrt{30193 \cdot 0.32 \cdot 0.32} = -183.30$

---------- FACIT-BEGIN ----------

Using equation (7-16) with $x = 10339$, $n = 30193$, and $p_0 = 0.32$:

```
(10339 - 0.32 * 30193) / sqrt(30193 * 0.32 * 0.68)
```

```
## [1] 8.355261
```

---------- FACIT-END ----------

Question IV.2 (10)

It is also of interest, whether there was a change from 2016 to 2017 in the proportion of students who were newly enrolled in the humanities (Hum).

Four different tests are performed in R using the following code:

```
  prop.test(x = 6691, n = 30193, p = 7297/30817, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  6691 out of 30193, null probability 7297/30817
## X-squared = 38.5, df = 1, p-value = 5.5e-10
## alternative hypothesis: true p is not equal to 0.23678
## 95 percent confidence interval:
##  0.21696 0.22633
## sample estimates:
##       p
## 0.22161
```

```
prop.test(x = c(7297, 6691), c(30817, 30193), correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(7297, 6691) out of c(30817, 30193)
## X-squared = 19.9, df = 1, p-value = 8.2e-06
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.0085083 0.0218461
## sample estimates:
##  prop 1  prop 2
## 0.23678 0.22161
```

```
binom.test(x = 6691, n = 30193, p = 7297/30817)
```

```
##
##  Exact binomial test
##
## data:  6691 and 30193
## number of successes = 6691, number of trials = 30193, p-value = 4.3e-10
## alternative hypothesis: true probability of success is not equal to 0.23678
## 95 percent confidence interval:
##  0.21693 0.22634
## sample estimates:
## probability of success
##                0.22161
```

```r
binom.test(x = 6691, n = 30193, p = (6691+7297)/(30193+30817))

##
##  Exact binomial test
##
## data:  6691 and 30193
## number of successes = 6691, number of trials = 30193, p-value = 0.0015
## alternative hypothesis: true probability of success is not equal to 0.22927
## 95 percent confidence interval:
##  0.21693 0.22634
## sample estimates:
## probability of success
##                0.22161
```

Which line of code performs the desired test?

1 ☐ `binom.test(x = 6691, n = 30193, p = 7297/30817)`

2* ☐ `prop.test(x = c(7297, 6691), c(30817, 30193), correct = FALSE)`

3 ☐ `binom.test(x = 6691, n = 30193, p = (6691+7297)/(30193+30817))`

4 ☐ `prop.test(x = 6691, n = 30193, p = 7297/30817, correct = FALSE)`

5 ☐ None of the four lines of code tests the relevant hypothesis

---------- FACIT-BEGIN ----------

The task is to compare proportions between two populations (Section 7.3) (and not, e.g., to test whether a proportion has a specific value).

---------- FACIT-END ----------

Question IV.3 (11)

Using a hypothesis test, it is also to be investigated whether the distribution of newly enrolled students across disciplines has changed over the three years for which data is given. The number of degrees of freedom in the relevant distribution of the test statistic is:

1 ☐ 3

2 ☐ 4

3* ☐ 6

4 ☐ 12

5 ☐ 20

---------- FACIT-BEGIN ----------

Comparison of distributions in different groups (Method 7.22). With four disciplines and three years, $(r-1)\cdot(c-1) = 3\cdot 2 = 6$.

---------- FACIT-END ----------

Question IV.4 (12)

Assuming independence between year and discipline, the expected number of newly enrolled students in TechNat in the year 2017 is estimated to be

1 ☐ 10339

2* ☐ $29020 \cdot 30193/90489 = 9683$

3 ☐ $(8551 + 10130 + 10339)/3 = 9673$

4 ☐ $10339 \cdot 29020/30193 = 9937$

5 ☐ $10339 \cdot 30193/29020 = 10757$

---------- FACIT-BEGIN ----------

As read in chapter 7.5.1 the expected number in a cell is calculated as:

$$\frac{\text{column total} \times \text{row total}}{\text{grand total}} = \frac{30193 \times 29020}{90489} = 9683$$

---------- FACIT-END ----------

Question IV.5 (13)

As the next step in testing whether the distribution across disciplines has changed over the years, the following is given: The test statistic is calculated to be 314.5. The significance level is set to $\alpha = 0.05$. In the distribution used to assess the test statistic, the 0.95 and 0.975 quantiles are, respectively, 12.59 and 14.45. What may be concluded? (Both the conclusion and reasoning must be correct).

1 ☐ The numbers provided above cannot be used to argue statistically, whether the distribution across disciplines has changed.

2 ☐ The distribution across disciplines has not changed significantly, as the test statistic is greater than the given 0.95 quantile.

3 ☐ The distribution across disciplines has not changed significantly, as the test statistic is greater than the given 0.975 quantile.

4 ☐ The distribution across disciplines has changed significantly, as, under the null hypothesis, there is a 95% probability of observing a test statistic greater than 12.59.

5* ☐ The distribution across disciplines has changed significantly, as the test statistic is greater than the given 0.95 quantile.

---------- FACIT-BEGIN ----------

Method 7.22. The test statistic is greater than the given 0.95 quantile, so the null hypothesis of no difference between groups (no change across years) is rejected, and the change is concluded to be significant.

---------- FACIT-END ----------

Continue on page 16

Exercise V

An experiment was conducted with the purpose of investigating the shelf life of a certain type of medicine. Altogether 26 identical, unopened bottles of medicine with the same production date were used for the experiment. Half of the bottles were stored at room temperature (21 °C), the other half at fridge temperature (5 °C). After 90 days the bottles were opened, and the content of active substance (in mg/ml) in each bottle was measured. The results were read into R in two vectors, `hightemp` (measurements from bottles stored at 21 °C) and `lowtemp` (measurements from bottles stored at 5 °C).

Furthermore, the following code was executed in R:

```
t.test(lowtemp, hightemp)

##
##  Welch Two Sample t-test
##
## data:  lowtemp and hightemp
## t = 5.3, df = 24, p-value = 2e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.4316 3.2592
## sample estimates:
## mean of x mean of y
##    7.4397    5.0943
```

Question V.1 (14)

What may be concluded when the significance level is set to $\alpha = 0.05$?

1* ☐ The mean content of the active substance is significantly larger after storage at fridge temperature than at room temperature. The difference is estimated to be 2.35 mg/ml.

2 ☐ It can be seen from the p-value that there is no significant difference between the mean content of the active substance in bottles stored at room temperature and at fridge temperature, respectively.

3 ☐ The mean content of the active substance is significantly larger after storage at fridge temperature than at room temperature. The difference is estimated to be 5.30 mg/ml.

4 ☐ The mean content of the active substance is significantly less after storage at fridge temperature than at room temperature. The difference is estimated to be 2.35 mg/ml.

5 ☐ The 95% confidence interval contains 2.35. Therefore, there is no significant difference between the mean content of the active substance in bottles which were stored at room temperature and fridge temperature, respectively.

---------- FACIT-BEGIN ----------

The p-value $2 \cdot 10^{-5}$ is much smaller than the significance level $\alpha = 0.05$ (or: 0 is not contained in the 95% confidence interval), so the difference is significant. The mean content of the active substance is estimated to be $7.4397 - 5.0943 = 2.35$ mg/ml larger in bottles which were stored at fridge temperature than in those which were stored at room temperature.

---------- FACIT-END ----------

Question V.2 (15)

A 99% confidence interval for the difference in the mean content of the active substance between bottles stored at fridge temperature and room temperature may be determined as follows:

1 ☐ $2.3454 \pm \frac{3.2592-2.3454}{2.7969} \cdot 2.0639 = [1.67, 3.02]$

2 ☐ $2.3454 \pm \frac{3.2592-2.3454}{2.4922} \cdot 2.7969 = [1.32, 3.37]$

3 ☐ $\left[1.4316 \cdot \frac{2.7969}{2.0639}, 3.2592 \cdot \frac{2.7969}{2.0639}\right] = [1.94, 4.42]$

4* ☐ $2.3454 \pm \frac{3.2592-2.3454}{2.0639} \cdot 2.7969 = [1.11, 3.58]$

5 ☐ $2.3454 \pm \frac{3.2592-2.3454}{2.0639} \cdot 2.4922 = [1.24, 3.45]$

---------- FACIT-BEGIN ----------

Using the notation from Method 3.47, it may be concluded from the R output that $\bar{x} - \bar{y} = 7.4397 - 5.0943 = 2.3454$, and that the degrees of freedom $\nu = 24$, so that $t_{0.975} = 2.0639$ and $t_{0.995} = 2.7969$ (`qt(0.975, df = 24)` and `qt(0.995, df = 24)`, respectively, in R).

Furthermore, according to the R output, $[1.4316, 3.2592]$ is a 95% confidence interval for the difference in mean content. The only thing in the equation we do not know is $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, but since we have one confidence interval given (95%), we can isolate this term from that equation. Thus, it follows from Method 3.47 that

$$\bar{x} - \bar{y} \pm t_{0.975} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.3454 \pm 2.0639 \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = [1.4316, 3.2592]$$

which may be used to conclude that

$$2.3454 + 2.0639 \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 3.2592 \Leftrightarrow \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \frac{3.2592 - 2.3454}{2.0639}\,.$$

Now, the 99% confidence interval can be computed using Method 3.47, as well:

$$\bar{x} - \bar{y} \pm t_{0.995} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.3454 \pm 2.7969 \cdot \frac{3.2592 - 2.3454}{2.0639} = [1.11, 3.58]\,.$$

---------- FACIT-END ----------

Continue on page 19

Exercise VI

In a production process things sometimes go wrong, and a component must be discarded after an inspection. From experience, it is known that there is a 20% probability of a component needing to be discarded. The assessment ("keep" or "discard") for a given component is independent of the assessments for the other components. One can assume that 20 components are produced per day.

Question VI.1 (16)

What is the probability that, on a randomly selected day, no components need to be discarded?

1 ☐ The number of components which need to be discarded is hypergeometrically distributed, and the probability is 0.

2* ☐ The number of components which need to be discarded is binomial distributed, and the probability is 0.0115.

3 ☐ The number of components which need to be discarded is binomial distributed, and the probability is 0.0576.

4 ☐ The number of components which need to be discarded is hypergeometrically distributed, and the probability is 0.0692.

5 ☐ The number of components which need to be discarded is binomial distributed, and the probability is 0.630.

---------- FACIT-BEGIN ----------

The number of components that get the assessment *discard* ("number of successes") out of 20 independently assessed components ("number of independent trials") is binomial distributed with probability $p = 0.2$ and size $n = 20$.

Let X be binomial distributed with probability $p = 0.2$ and size $n = 20$. Then $P(X = 0)$ (or, equivalently, $P(X \leq 0)$) may be computed in R as follows

```
dbinom(0, size = 20, p = 0.2)

## [1] 0.01152922

pbinom(0, size = 20, p = 0.2)

## [1] 0.01152922
```

---------- FACIT-END ----------

Question VI.2 (17)

A simulation of the number of discarded components per day has been carried out. Let X_i denote the number of discarded components on a randomly selected day i. A sample of $n = 20$ values has been simulated, and is denoted by x_i, $i = 1, \ldots, 20$. Which of the following statements is the only one that can be correct?

1 ☐ The expected number of components to be discarded during one day is $\mu_X = 4.5$. The sample mean of the simulated values was $\hat{\mu}_X = 4.3$.

2* ☐ The expected number of components to be discarded during one day is $\mu_X = 4$. The sample mean of the simulated values was $\hat{\mu}_X = 3.7$.

3 ☐ The expected number of components to be discarded during one day is $\mu_X = 5$. The sample mean of the simulated values was $\hat{\mu}_X = 4.9$.

4 ☐ The expected number of components to be discarded during one day is $\mu_X = 2$. The sample mean of the simulated values was $\hat{\mu}_X = 2.2$.

5 ☐ The expected number of components to be discarded during one day is $\mu_X = 4.25$. The sample mean of the simulated values was $\hat{\mu}_X = 4.25$.

---------- FACIT-BEGIN ----------

Again, let X be binomial distributed with probability $p = 0.2$ and size $n = 20$. See theorem 2.21. The expected number of components to be discarded during one day, μ_X, may then be computed as

$$\mu_X = \mathrm{E}(X) = np = 20 \cdot 0.2 = 4\,,$$

so the answer option with $\mu_X = 4$ is the only one that can be correct.

---------- FACIT-END ----------

Question VI.3 (18)

An experiment is planned in order to estimate the proportion of components that need to be discarded.

How many days does the experiment need to run in order to obtain a 95% confidence interval for the proportion of components to be discarded, which has an expected width of 5 percentage points?

1 ☐ $n = (1.96 \cdot 0.2/0.05)^2 = 61.5$, i.e. 4 days.

2 ☐ $n = 0.16 \cdot (1.96/0.05)^2 = 246$, i.e. 13 days.

3* ☐ $n = 0.16 \cdot (1.96/0.025)^2 = 983$, i.e. 50 days.

4 ☐ $n = 0.2 \cdot (1.96/0.025)^2 = 1229$, i.e 62 days.

5 ☐ $n = (1.96 \cdot 0.2/0.025)^2 = 3934$, i.e. 197 days.

---------- FACIT-BEGIN ----------

Use Method 7.13 with $p = 0.2$, remembering that the ME is half the expected width of the confidence interval, and using the information that 20 components are produced per day.

---------- FACIT-END ----------

Continue on page 22

Exercise VII

A factory which produces percussion instruments and accessories produces drumsticks as well. For the purpose of quality control, the lengths (in cm) of 20 drumsticks selected at random were measured. These lengths were read into the vector `length` in R. The lengths can be assumed to be independent and normally distributed with mean μ and variance σ^2.

Question VII.1 (19)

The following commands were executed in R:

```
sum(length)

## [1] 793.1

var(length)

## [1] 0.01099
```

Give an expression for a 95% confidence interval for the mean length of the drumsticks.

1 ☐ $\frac{793.1}{20} \pm 2.093 \cdot \frac{0.01099}{\sqrt{20}}$

2 ☐ $793.1 \pm 2.086 \cdot \frac{0.01099}{\sqrt{20}}$

3* ☐ $\frac{793.1}{20} \pm 2.093 \cdot \frac{\sqrt{0.01099}}{\sqrt{20}}$

4 ☐ $\frac{793.1}{20} \pm 2.086 \cdot \frac{\sqrt{0.01099}}{\sqrt{20}}$

5 ☐ $\frac{793.1}{20} \pm 2.093 \cdot \frac{0.01099}{\sqrt{19}}$

---------- FACIT-BEGIN ----------

Use Method 3.9 / (3-11) with $n = 20$, $\bar{x} = 793.1/20$, $s = \sqrt{0.01099}$. The 0.975 quantile for the t-distribution with 19 degrees of freedom is:

```
qt(0.975, df = 19)

## [1] 2.093024
```

---------- FACIT-END ----------

Question VII.2 (20)

A t-test is performed in order to investigate whether the mean length of the drumsticks may be assumed to be 39.60 cm. The observed t-test statistic is calculated to be $t_{\text{obs}} = 2.38$. Which of the following is the only correct conclusion?

1 ☐ The mean length is significantly different from 39.60 cm when the significance level $\alpha = 0.01$ is used.

2 ☐ As the p-value is greater than 0.05, the null hypothesis that the mean length is 39.60 cm is rejected, no matter whether the significance level is set to $\alpha = 0.05$ or $\alpha = 0.01$.

3 ☐ As the p-value is less than 0.05, the null hypothesis that the mean length is 39.60 cm is accepted when the significance level is set to $\alpha = 0.05$.

4* ☐ The mean length is significantly different from 39.60 cm when the significance level is set to $\alpha = 0.05$.

5 ☐ As the p-value is greater than 0.05, the null hypothesis that the mean length is 39.60 cm is accepted, no matter whether the significance level $\alpha = 0.05$ or $\alpha = 0.01$ is used.

---------- FACIT-BEGIN ----------

The p-value is

$$p = 2 \cdot P(T > |t_{\text{obs}}|) = 2 \cdot P(T > 2.38) = 2 \cdot (1 - P(T \leq 2.38)) = 0.028$$

where $P(T \leq 2.38)$ is computed in R as `pt(2.38, df = 19)`.

As the p-value lies in between 0.01 and 0.05, the null hypothesis is rejected at significance level $\alpha = 0.05$ (but not at significance level $\alpha = 0.01$).

---------- FACIT-END ----------

Question VII.3 (21)

In relation to subsequent quality control, plans are made to select a new sample. The sample size must be sufficiently large to detect a difference of 0.5 mm between the mean length of the drumsticks and the desired mean length 39.60 cm with power 90%, at significance level $\alpha = 0.01$. Here, 0.11 is to be used as a guess for the population standard deviation. Use the `power.t.test` function in R to determine the necessary sample size.

1 ☐ $n = 20$

2 ☐ $n = 22$

3 ☐ $n = 146$

4 ☐ $n = 53$

5* ☐ $n = 76$

---------- FACIT-BEGIN ----------

```
power.t.test(power = 0.9, delta = 0.05, sd = 0.11,
             sig.level = 0.01, type = "one.sample")

##
##      One-sample t test power calculation
##
##               n = 75.36328
##           delta = 0.05
##              sd = 0.11
##       sig.level = 0.01
##           power = 0.9
##     alternative = two.sided
```

See example 3.67 for examples of power.t.test.

---------- FACIT-END ----------

Continue on page 25

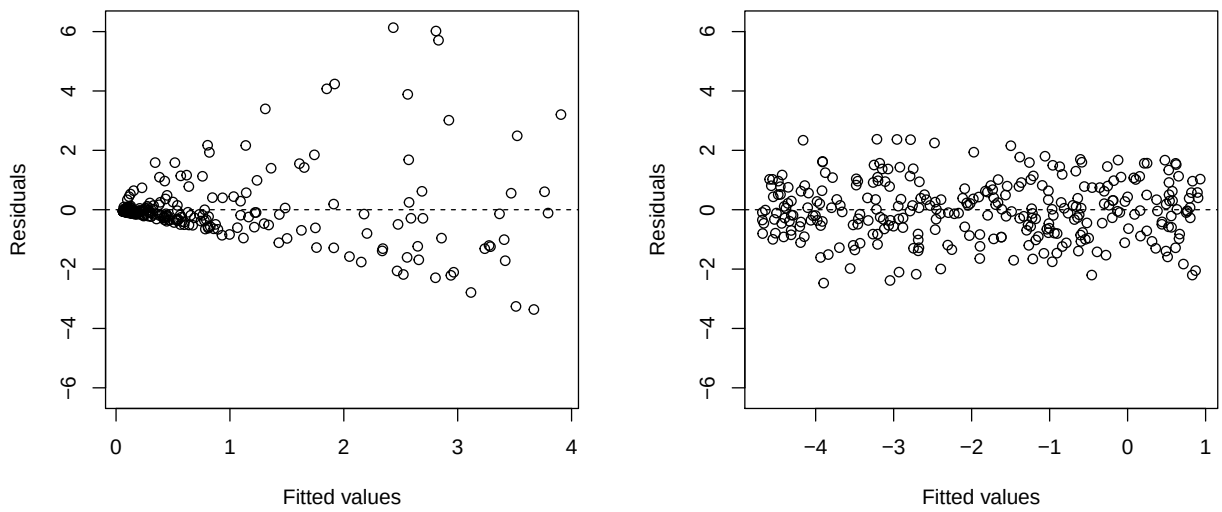
Exercise VIII

Two quantitative variables have been read into R as `x` and `y`. One would like to describe the relationship between these two variables using a linear regression model. To this end, two different linear regression models have been estimated in R, see the R code below.

```
logx = log(x)
logy = log(y)
model1 <- lm(y ~ x)
model2 <- lm(logy ~ logx)
```

Question VIII.1 (22)

Plots of the residuals against the fitted values for `model1` (left) and `model2` (right), respectively, are shown below. On the basis of these plots, would one prefer to analyse the data using the statistical model given by `model1` or the one given by `model2`? (Both the conclusion and reasoning must be valid).



1 ☐ There are clear linear associations between the residuals and the fitted values in the plot to the left, while no linear association is seen in the plot to the right. Thus, one would prefer to use `model1`.

2 ☐ The assumption of variance homogeneity is clearly not satisfied for `model2`, while the assumption seems reasonable for `model1`. Thus, one would prefer to use `model1`.

3 ☐ In the plot to the left, a lot of the fitted values lie in the interval [0,1], while they are better spread out over the whole x axis in the plot to the right. Thus, one would prefer to use `model2`.

4* ☐ The assumption of variance homogeneity is clearly not satisfied for `model1`, while the assumption seems reasonable for `model2`. Thus, one would prefer to use `model2`.

5 ☐ The residuals in the figure to the right are obviously uniformly distributed in an interval around 0 (thus not normally distributed), while the residuals in the figure to the left might well be normally distributed. Thus, one would prefer to use `model1`.

---------- FACIT-BEGIN ----------

In the plot to the left, it is clear that the variance of the residuals increases with the fitted values, revealing that the assumption of variance homogeneity does not hold for `model1`. In the plot to the right, the variance of the residuals seems to be quite constant across the fitted values, indicating that the assumption of variance homogenity should be ok for `model2`.

---------- FACIT-END ----------

Continue on page 27

Exercise IX

Many factors affect the indoor climate of a building. One of the most common measures for the quality of the indoor climate is the level of CO_2. If there is insufficient ventilation, the CO_2 level becomes too high, which, among other things, decreases peoples' ability to concentrate. In new buildings with classrooms, the CO_2 level may not exceed 1000 ppm - in outdoor air there is around 400 ppm (before the industrial revolution it was around 280 ppm!).

In a study of the indoor climate in classrooms, samples of the CO_2 level were taken from two different classrooms. Both samples consist of one-hour average values measured over a period of 2 months. Only values where people were present in the classroom have been included in the samples. The observations for room 1 and room 2, respectively, were loaded into R in the vectors `room1CO2` and `room2CO2`.

Question IX.1 (23)

The following code was run in R:

```
length(room1CO2)

## [1] 304

length(room2CO2)

## [1] 252

sum((room1CO2 - mean(room1CO2))^2)

## [1] 131606104

(length(room2CO2)-1)*var(room2CO2)

## [1] 12775276
```

Determine the sample standard deviation for room 1 (s_1) and room 2 (s_2), respectively.

1* ☐ $s_1 = 659.0475$ and $s_2 = 225.6048$

2 ☐ $s_1 = 434343.6$ and $s_2 = 50897.51$

3 ☐ $s_1 = 657.9626$ and $s_2 = 225.1567$

4 ☐ $s_1 = 11471.97$ and $s_2 = 3574.252$

5 ☐ None of the four answers above can be correct.

---------- FACIT-BEGIN ----------

See definition 1.11.

$$s_1 = \sqrt{\frac{131606104}{304-1}} = 659.0475$$
$$s_2 = \sqrt{\frac{12775276}{252-1}} = 225.6048$$

---------- FACIT-END ----------

Question IX.2 (24)

In addition, the following has been run in R:

```
Q3 <- function(x){ quantile(x, 0.75) }

simSamples1 <- replicate(10000, sample(room1CO2, replace = TRUE))
simSamples2 <- replicate(10000, sample(room2CO2, replace = TRUE))

simQ3s1 <- apply(simSamples1, 2, Q3)
simQ3s2 <- apply(simSamples2, 2, Q3)
simQ3sdiff <- simQ3s1 - simQ3s2

quantile(simQ3s1, c(0, 0.025, 0.05, 0.95, 0.975, 1))

##       0%     2.5%       5%      95%    97.5%     100%
## 1417.896 1562.332 1583.146 1833.104 1838.021 1953.792

quantile(simQ3s2, c(0, 0.025, 0.05, 0.95, 0.975, 1))

##       0%     2.5%       5%      95%    97.5%     100%
## 772.0833 827.5000 831.1042 916.4583 920.5000 966.6667

quantile(simQ3sdiff, c(0, 0.025, 0.05, 0.95, 0.975, 1))

##        0%      2.5%        5%       95%     97.5%      100%
##  534.6458  685.8297  712.4562  976.1042  991.6250 1093.4167
```

Use this R output to determine a 95% confidence interval for the difference between the 0.75 quantiles for the CO_2 level in room 1 and 2.

1 ☐ $[916, 1833]$

2 ☐ $[828, 921]$

3* ☐ $[686, 992]$

4 ☐ $[1556 - 828, 1838 - 921] = [728, 917]$

5 ☐ $[828, 1838]$

---------- FACIT-BEGIN ----------

The 95% bootstrap confidence interval is determined by the 0.025 and 0.975 quantiles of the simulated differences (`simQ3sdiff`).

---------- FACIT-END ----------

Continue on page 30

Exercise X

The table below shows the average yield (measured in hkg/acres) for 5 crops (Crop 1-5) in Denmark in the years 2014-2017.

	2014	2015	2016	2017	*Average*
Crop 1	79	80	73	83	*78.75*
Crop 2	46	48	47	52	*48.25*
Crop 3	64	63	57	66	*62.50*
Crop 4	66	68	62	68	*66.00*
Crop 5	57	60	55	58	*57.50*
Average	*62.40*	*63.80*	*58.80*	*65.40*	*62.60*

In addition to the row and column averages given in the table (in italics), it is given that $SS(\text{Year}) = 118.8$ and $SST = 2172.8$.

In this exercise, the 20 average crop yields in the table are considered to be observations from 20 different randomly selected fields. It is assumed that within each year, there is no difference between the expected yields from the five crops. The analysis must therefore be carried out as if the same crop was sown on all the fields (and the information about crop type should not be used in the exercise). We use a model of the form

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

with $\sum \alpha_i = 0$, where $\varepsilon_{ij} \sim N(0, \sigma^2)$ and independent.

Question X.1 (25)

Let α_1 describe the effect of year 2014 on the expected yield. Give the estimate of α_1.

1 ☐ $\hat{\alpha}_1 = 78.75 - 62.40 = 16.35$

2* ☐ $\hat{\alpha}_1 = 62.40 - 62.60 = -0.20$

3 ☐ $\hat{\alpha}_1 = 78.75$

4 ☐ $\hat{\alpha}_1 = 62.60$

5 ☐ $\hat{\alpha}_1 = 62.40$

---------- FACIT-BEGIN ----------

As stated in equation 8-4 $\hat{\alpha}_1$ is computed as the average yield in the year 2014 (62.40) minus the overall average yield (62.60).

---------- FACIT-END ----------

Question X.2 (26)

Set the significance level to $\alpha = 0.05$. Give the critical value for the usual test used to investigate whether the expected crop yield differs between years.

1 ☐ 3.73

2* ☐ 3.24

3 ☐ 26.30

4 ☐ 3.49

5 ☐ 2.96

---------- FACIT-BEGIN ----------

Theorem 8.6. The F-test statistic is evaluated using the F distribution with

$$(k-1, n-k) = (4-1, 20-4) = (3, 16)$$

degrees of freedom, and the critical value is the 0.95 quantile of this distribution:

```
qf(0.95, df1 = 3, df2 = 16)
```

```
## [1] 3.238872
```

---------- FACIT-END ----------

Question X.3 (27)

Give the estimate of σ^2.

1 ☐ $\hat{\sigma}^2 = 2054$

2 ☐ $\hat{\sigma}^2 = 3.058$

3 ☐ $\hat{\sigma}^2 = 36.7$

4 ☐ $\hat{\sigma}^2 = 29.7$

5* ☐ $\hat{\sigma}^2 = 128.375$

---------- FACIT-BEGIN ----------

As can be read in chapter 8.2.2, MSE and MST are both central estimators for the the variance, but that MSE holds true both if the null-hypothesis is rejected or not, so in this case we will use this as the estimate.

$$SSE = SST - SS(\text{Year}) = 2172.8 - 118.8 = 2054$$

and

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-k} = \frac{2054}{16} = 128.375$$

---------- FACIT-END ----------

Continue on page 33

Exercise XI

The data set used in this exercise is the same as in the previous exercise. However, here, it is to be used that data describes the average yield (measured in hkg/acres) for 5 different crops (Crop 1-5) in Denmark in the years 2014-2017.

	2014	2015	2016	2017	*Average*
Crop 1	79	80	73	83	*78.75*
Crop 2	46	48	47	52	*48.25*
Crop 3	64	63	57	66	*62.50*
Crop 4	66	68	62	68	*66.00*
Crop 5	57	60	55	58	*57.50*
Average	*62.40*	*63.80*	*58.80*	*65.40*	*62.60*

In addition to the row and column averages given in the table (in italics), it is given that $SS(\text{Crop}) = 2017.3$, $SS(\text{Year}) = 118.8$ and $SST = 2172.8$. The average crop yields are assumed to be realizations of normally distributed random variables.

Question XI.1 (28)

A key question is whether there has been a development in the yield over time. To investigate this, it is tested whether there is a statistically significant difference between the four years, when the variation between the different crops is taken into account. Give the p-value and the conclusion at significance level $\alpha = 0.05$ for the usual test.

1 ☐ There is no significant difference in yield between the years, since $p = 0.82 > 0.05$.

2* ☐ There is a significant difference in yield between the years, since $p = 0.00045 < 0.05$.

3 ☐ There is no significant difference in yield between the years, since $p = 2 \cdot 10^{-10} < 0.05$.

4 ☐ There is a significant difference in yield between the years, since $p = 0.82 > 0.05$.

5 ☐ There is a significant difference in yield between the years, since $p = 8 \cdot 10^{-7} < 0.05$.

---------- FACIT-BEGIN ----------

Use Theorem 8.22. First, find *SSE*:

$$SSE = SST - SS(\text{Crop}) - SS(\text{Year}) = 2172.8 - 2017.3 - 118.8 = 36.7\,.$$

Then, the F-test statistic may be computed by

$$F_{\text{Year}} = \frac{SS(\text{Year})/(k-1)}{SSE/((k-1)(l-1))} = \frac{118.8/(4-1)}{36.7/((4-1)(5-1))} = \frac{118.8/3}{36.7/12}\,,$$

and the p-value by

```r
1-pf((118.8/3)/(36.7/12), df1 = 3, df2 = 12)
```

```
## [1] 0.0004526046
```

---------- FACIT-END ----------

Question XI.2 (29)

It was decided in advance that the difference between the yields in 2014 and 2017 should be seen as an indicator of the overall development. Set the significance level to $\alpha = 0.05$ and test the hypothesis that there is no difference between the two years. What is the conclusion? (Both the conclusion and reasoning must be correct).

1 ☐ There is a significant difference in yield between the two years, since $\frac{65.4-62.4}{62.6} < 0.05$.

2 ☐ There is no significant difference in yield between the two years, since $\frac{118.8}{2017.3} > 0.05$.

3 ☐ There is no significant difference in yield between the two years, since $\frac{3}{\sqrt{36.7}} < 2.18$.

4* ☐ There is a significant difference in yield between the two years, since $\frac{-3}{1.11} < -2.18$.

5 ☐ There is a significant difference in yield between the two years, since $\frac{3}{36.7} > 0.05$.

---------- FACIT-BEGIN ----------

Use Method 8.10 modified for two-way ANOVA. With 5 observations in each of the two years, and $MSE = SSE/12$, the relevant t-test statistic becomes

$$t_{\text{obs}} = \frac{62.40 - 65.40}{\sqrt{\frac{36.7}{12} \cdot \frac{2}{5}}} = \frac{-3}{1.11} = -2.712\,.$$

The negative/smaller of the two critical values for the test is:

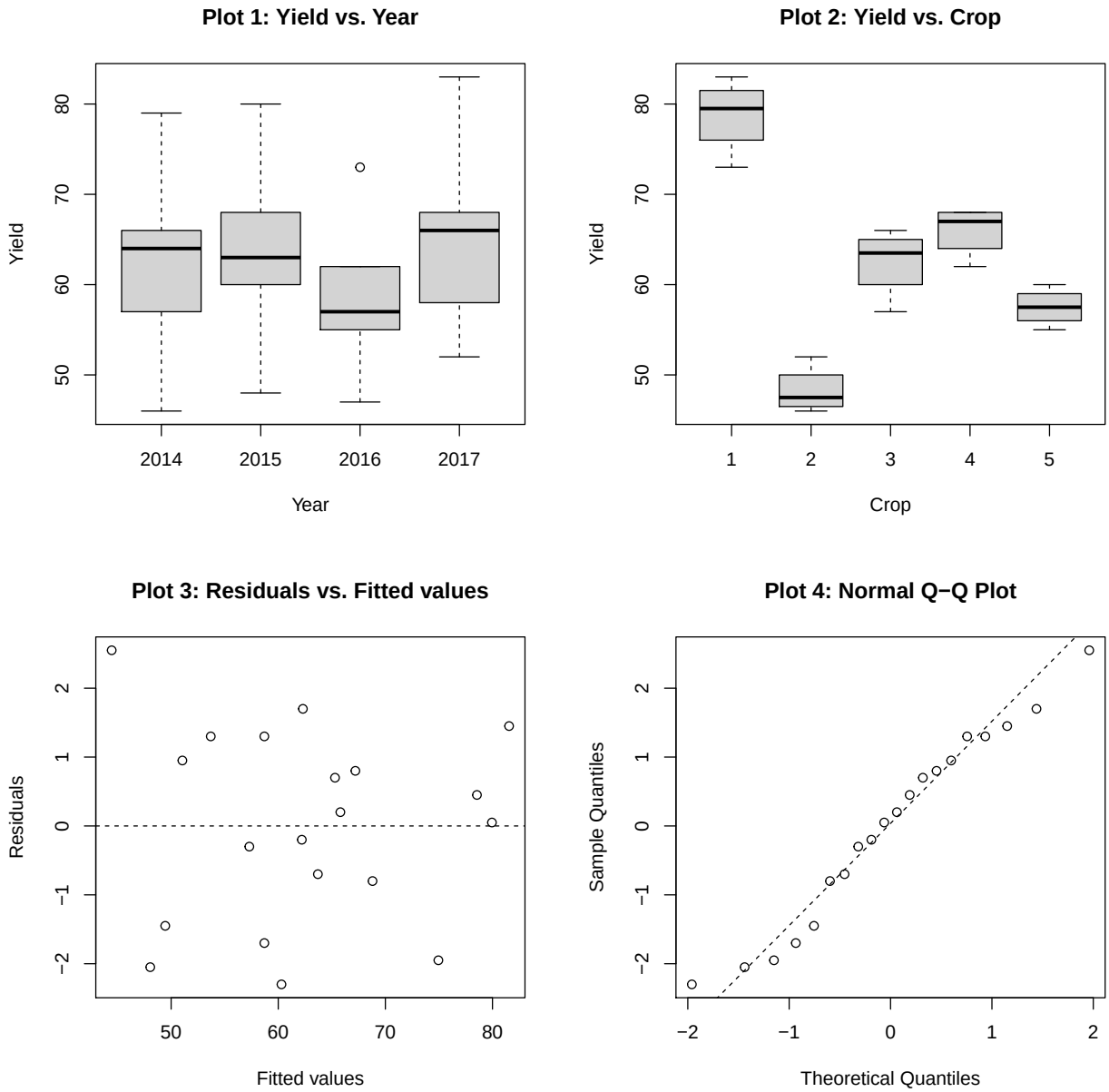
```r
qt(0.025, df = 12)
```

```
## [1] -2.178813
```

Since the test-statistic is more extreme than the critical value there is a significant different.

---------- FACIT-END ----------

As part of the model validation, different plots are given in the figure below.



Question XI.3 (30)

Based on the figure, which of the following statements is correct?

1 ☐ As the variation within years is much greater than the variation within crops (Plot 1 and Plot 2), the assumption of variance homogeneity is clearly not fulfilled.

2 ☐ The normal distribution assumption is clearly not fulfilled (Plot 3), while the assumption of variance homogeneity is clearly fulfilled (Plot 4).

3* ☐ The normal distribution assumption appears to be fulfilled (Plot 4), and the same applies to the assumption of variance homogeneity (Plot 3).

4 ☐ The yield cannot be different for the five crops, as the residuals vary randomly around the horizontal line through $y = 0$ (Plot 2 and Plot 3).

5 ☐ There cannot be any difference between the yields in the different years (Plot 1), but the normal distribution assumption is clearly fulfilled (Plot 3).

---------- FACIT-BEGIN ----------

A plot of the residuals against the fitted values may be used to investigate the assumption of variance homogeneity for a two-way ANOVA model, while a normal QQ-plot may be used to investigate the assumption of normality.

---------- FACIT-END ----------

The exam is finished. Have a great summer!