Technical University of Denmark *Page 1 of 29 pages.*

Written examination: 17. December 2022

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

Duration: 4 hours

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 10 exercises. To answer the questions, you need to fill in the "multiple choice" form on exam.dtu.dk.

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

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

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

The exam paper contains 29 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. Also remember that there may be slight discrepancies between the result of the book's formulas and corresponding built-in functions in R.*

Exercise I

Let X and Y be independent random variables, where X has mean 2 and variance 2, and Y has mean -1 and variance 3.

Question I.1 (1)

What is the mean of $2X + Y$?

1 ☐ 0

2 ☐ 2

3* ☐ 3

4 ☐ 11

5 ☐ We have insufficient information to determine the mean of $2X + Y$.

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

Theorem 2.56.

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

Continue on page 3

Exercise II

Measurements of serum cholesterol (mg/100ml), x, and arterial calcium deposition (mg/100g dry weight of tissue), y, were made on twelve animals. The data was read into R:

```
y <- c(59, 52, 42, 59, 24, 24, 40, 32, 63, 55, 34, 24)
x <- c(298, 303, 270, 287, 236, 245, 265, 233, 286, 290, 264, 239)
```

Consider the following simple linear regression model.

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i \text{ where } \varepsilon_i \sim N(0, \sigma^2)$$

Question II.1 (2)

Calculate the coefficient of determination (R^2) and choose the correct answer below:

1 ☐ 0.9129

2 ☐ 0.8334%

3* ☐ 0.8334

4 ☐ 91.29%

5 ☐ 0.8168%

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

The coefficient of determination (R^2) measures how much variability in y is explained by the model. R^2 is bound between zero and one. One way to calculate it is by squaring the correlation coefficient r.

```
cor(x, y)^2 = 0.8334
```

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

Question II.2 (3)

The following line of code has been executed in R.

```
fit <- lm(y~x)
```

Which of the following commands can be used as part of the model validation, i.e., to check if the normality assumptions are fulfilled?

1 ☐
```
qqnorm(fit$fitted.values)
qqline(fit$fitted.values)
```

2* ☐
```
qqnorm(fit$residuals)
qqline(fit$residuals)
```

3 ☐
```
qqnorm(y)
qqline(y)
```

4 ☐
```
qqnorm(residuals)
qqline(residuals)
```

5 ☐
```
qqnorm(lm$residuals)
qqline(lm$residuals)
```

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

The model assumptions of a linear regression model are $\varepsilon_i \sim N(0, \sigma^2)$ and i.i.d., which means that the residulas have to be normally distributed with equal variance independent of location. The normality of the residulas can be assessed using a qq-plot (second answer).

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

Question II.3 (4)

The model summary for a simple linear regression model is shown below.

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2249 -3.4900 -0.8876  2.1968 10.9510
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -102.3218    20.5319  -4.984 0.000551 ***
## x              0.5398     0.0763   7.074  3.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.358 on 10 degrees of freedom
## Multiple R-squared:  0.8334,Adjusted R-squared:  0.8168
## F-statistic: 50.04 on 1 and 10 DF,  p-value: 3.401e-05
```

Which of the following expressions calculates the 95% confidence interval for the cholesterol slope ($\hat{\beta}_1$)?

1 ☐ $0.5398 \pm 1.9600 \cdot 0.0763$

2 ☐ $0.5398 \pm 2.1788 \cdot 0.0763$

3* ☐ $0.5398 \pm 2.2281 \cdot 0.0763$

4 ☐ $-102.3218 \pm 1.9600 \cdot 20.5319$

5 ☐ $-102.3218 \pm 2.2281 \cdot 20.5319$

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

The 95% confidence interval for a parameter of interest can be calculated using parameter estimate, critical quantile and standard error. These can be extracted from the displayed model summary. For this particular question the exact calculation is:

$\hat{\beta}_1 + t_{0.975,df=10} \cdot \hat{\sigma}_{\beta_1} = 0.5398 \pm 2.2281 \cdot 0.0763$

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

Continue on page 6

Exercise III

A person is considering to buy an electric car. In order to make a well-informed choice, he finds the range for a fully charged car (km) and battery size (kWh) for different car models, as given by the car manufactures.

Initially, the potential car owner considers the effectiveness of the electric cars, i.e. range per kWh. So he ran the following R-code (where `range1` is the range, and `battery` is the battery size given by the car manufacturer):

```
t.test(range1 / battery)

##
##  One Sample t-test
##
## data:  range1/battery
## t = 45.117, df = 34, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  6.170481 6.752586
## sample estimates:
## mean of x
##  6.461533
```

Question III.1 (5)

Initially the potential car owner wants to test the hypothesis

$$H_0: \quad \text{The mean effectiveness is 6 } km/kWh$$

What is the conclusion using significance level $\alpha = 0.05$ (all parts of the answer should be correct)?

1 ☐ The effectiveness is significantly different from $6km/kWh$ as the p-value from the output above is less than $2.2 \cdot 10^{-16}$

2 ☐ The effectiveness is not statistically different from $6km/kWh$ as $6 < 6.17$.

3 ☐ The effectiveness is less than $6km/kWh$ as $6 < 6.17$.

4* ☐ The effectiveness is larger than $6km/kWh$ as $6 < 6.17$.

5 ☐ The effectiveness is not significantly different from 6 as the p-value is less than $2.2 \cdot 10^{-16}$

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

We observe that 6 is outside the confidence interval. The estimated effictiveness is larger than 6, hence option 4 is correct. Also note that the p value from the t test tests if the effectiveness is zero, so options 1 and 5 are wrong.

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

Question III.2 (6)

Based on the analysis above, what is a 99% confidence interval for the effectiveness of electric cars?

1 ☐ $[5.67, 7.25]$

2 ☐ $[5.93, 6.99]$

3* ☐ $[6.07, 6.85]$

4 ☐ $[6.11, 6.81]$

5 ☐ $[6.17, 6.75]$

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

```
se <- 6.461533/45.117
6.461533+c(-1,1)*qt(0.995,df=34)*se

## [1] 6.070780 6.852286
```

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

The potential car owner decides to calculate a confidence interval for log effectiveness, as a help for the analysis the following R-code with output is given

```
mean(log(range1 / battery))

## [1] 1.857769

var(log(range1 / battery))

## [1] 0.01643549
```

Question III.3 (7)

What is the 95% confidence interval for log-effectiveness of electric cars?

1 ☐ $[1.66, 2.06]$

2 ☐ $[1.71, 2.00]$

3 ☐ $[1.75, 1.96]$

4* ☐ $[1.81, 1.90]$

5 ☐ $[1.85, 1.86]$

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

```
m1 <- 1.857769
v1 <- 0.01643549
m1 + c(-1, 1) * qt(0.975, 34) * sqrt(v1) / sqrt(35)

## [1] 1.813730 1.901808
```

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

A car magazine made an independent test on the same car models, and the potential car owner now wants to compare the effectiveness as given by the manufacturers with the ones found by the car magazine.

Question III.4 (8)

In the following R-code `range2` denote the range of a full battery as found by the car magazine. Which of the following pieces of code test if there is a significant difference between the effectiveness given by the car manufacturers and the effectiveness found by the test?

1 ☐
```
t.test(log(range1), log(range2), mu = 1, paired = TRUE)
```

2 ☐
```
t.test(log(range1 / battery), log(range2 / battery), mu = 1)
```

3 ☐
```
t.test(log(range1), log(range2))
```

4* ☐
```
t.test(log(range1), log(range2), paired = TRUE)
```

5 ☐
```
t.test(range1 / battery, range2 / battery, mu = 1)
```

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

First of all, it is a paired test since it is the same car models that are being compared. Option 1 tests the null hypothesis of the mean difference being 1 (incorrect), whereas option 1 tests the null hypothesis of the mean difference being 0 (correct).

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

Question III.5 (9)

If the standard deviation of the effectiveness of electric cars is assumed to be $0.8km/(kWh)$, how many cars should be tested to get a margin of error of 0.1?

1 ☐ 16

2 ☐ 61

3 ☐ 157

4* ☐ 246

5 ☐ 492

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

```
(1.96 * 0.8 / 0.1) ^ 2
```

```
## [1] 245.8624
```

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

Continue on page 10

Exercise IV

In an apple plantation an experiment was carried out with different growing conditions. Trees were planted and divided into 4 groups, which each were grown in different conditions. After the growing season the trees were picked for apples and they were weighted for each tree. The values are shown in kg in the following table divided on each group:

Conditions A	Conditions B	Conditions C	Conditions D
14.9	14.3	14.6	14.0
15.9	16.1	12.7	15.1
16.2	14.9	13.6	13.3
15.9	16.4	13.7	17.3
15.9	15.7		12.9
	14.9		13.1
	15.7		

In order to investigate if the growing conditions led to significantly different mean weights of the apples, the data was analysed using an ANOVA. The results are summarised with the following ANOVA table:

```
anova(lm(Weights ~ Conditions))

## Analysis of Variance Table
##
## Response: Weights
##            Df Sum Sq Mean Sq F value  Pr(>F)
## Conditions  3 14.163   4.721  4.1852 0.02061 *
## Residuals  18 20.305   1.128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question IV.1 (10)

What is the total variance *SST* for this data?

1 ☐ 1.128

2 ☐ 5.85

3 ☐ 14.2

4* ☐ 34.5

5 ☐ 40.3

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

```
sum(val['Sum Sq'])

## [1] 34.46773

14.163 + 20.305

## [1] 34.468
```

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

Question IV.2 (11)

What is the correct statement based on the results from the ANOVA table (both conclusion and argument must be correct)?

1 ☐ At a significance level of 1% a significant difference in weight, caused by the growing conditions, is detected, since $1.128 < 4.721$.

2 ☐ At a significance level of 5% no significant difference in weight, caused by the growing conditions, is detected, since $4.1852 < 5$.

3 ☐ At a significance level of 5% no significant difference in weight, caused by the growing conditions, is detected, since $4.721 > 1.128$.

4* ☐ At a significance level of 5% a significant difference in weight, caused by the growing conditions, is detected, since $0.02061 < 0.05$.

5 ☐ None of the statements above are correct.

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

0.02061 is the p value for the test. Hence statement 4 is correct. The other options compare various wrong quantities (the 95% quantile in the $F(3, 18)$ distribution, which could have been compared to the F value, is 3.160.).

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

Continue on page 12

Exercise V

At DTU, 35% of new students are women. A group of six new students were randomly selected for a focus group.

Question V.1 (12)

What is the probability that at least two women are selected?

1 ☐ 0.243

2 ☐ 0.319

3 ☐ 0.328

4 ☐ 0.647

5* ☐ 0.681

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

```
1 - pbinom(1, 6, 0.35)
```

```
## [1] 0.6809201
```

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

Question V.2 (13)

If we let X be a random variable representing the number of women in the focus group, what is the variance of X?

1 ☐ 0.23

2 ☐ 0.74

3 ☐ 1.17

4* ☐ 1.37

5 ☐ 2.10

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

Theorem 2.21.

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

Question V.3 (14)

One of the students goes to DTU by bus. Suppose buses arrive at random at the bus stop with an average of one bus per fifteen minutes. What is the probability that the student must wait at least 20 minutes for the bus?

1 ☐ 1.8%

2* ☐ 26.4%

3 ☐ 35.1%

4 ☐ 52.8%

5 ☐ 73.7%

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

```
ppois(0, lambda = 20/15)

## [1] 0.2635971

## or
1 - pexp(20, rate = 1/15)

## [1] 0.2635971
```

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

Question V.4 (15)

It is considered an unlucky day if your bus waiting time is above the 90% quantile of the waiting time distribution. How much time must you wait in order for your day to become unlucky?

1 ☐ 13.5 min

2 ☐ 26.9 min

3* ☐ 34.5 min

4 ☐ 46.1 min

5 ☐ 135 min

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

```r
qexp(0.90, rate = 1/15)

## [1] 34.53878
```

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

Continue on page 15

Exercise VI

Question VI.1 (16)

We wish to simulate 50 random samples from a uniform distribution, where 0 and 100 defines the range of possible outcomes. Which of the following commands generates this?

1* ☐ `runif(50, 0, 100)`

2 ☐ `dnorm(50, 0, 100)`

3 ☐ `dunif(0, 100, 50)`

4 ☐ `runif(0, 100, 50)`

5 ☐ `rnorm(50, 0, 100)`

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

The `runif(n,min,max)` function generates `n` random numbers from a uniform distribution, with `min` and `max` as the boundaries.

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

Continue on page 16

Exercise VII

We are interested in studying the systolic blood pressure y in relation to weight x_1 (lb (pounds)) and age x_2 (years) in a group of males of approximately the same height.

Question VII.1 (17)

A multiple linear regression model of the following form has been established.

$$Y_i = \beta_0 + \beta_1 \cdot x_{1,i} + \beta_2 \cdot x_{2,i} + \varepsilon_i \text{ where } \varepsilon_i \sim N(0, \sigma^2)$$

The model summary is given below.

```
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6447 -2.0191 -0.0607  2.1331  6.0856
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.48354   26.36399   0.967 0.356537
## x1           0.62049    0.13513   4.592 0.000992 ***
## x2           0.04803    0.12948   0.371 ?
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.286 on 10 degrees of freedom
## Multiple R-squared:  0.7957,Adjusted R-squared:  0.7549
## F-statistic: 19.48 on 2 and 10 DF,  p-value: 0.0003555
```

Calculate the omitted p-value for the hypothesis $H_0 : \beta_2 = 0$. Which of the following answers is correct?

1 ☐ 0.0500

2 ☐ 0.3592

3 ☐ 0.3596

4* ☐ 0.7184

5 ☐ 0.7192

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

```
2*(1-pt(0.371, 10))
```

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

Question VII.2 (18)

Look at the model summary from the question above. How many observations (n) were measured in this dataset?

1 ☐ $n = 9$

2 ☐ $n = 10$

3 ☐ $n = 11$

4 ☐ $n = 12$

5* ☐ $n = 13$

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

The residual degrees of freedom are given by $n - 3$ because three parameters were estimated for the multiple linear regression model, hence $n = 13$.

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

Question VII.3 (19)

Look at the model summary above. Which of the following statements is true using significance level $\alpha = 0.05$ (both conclusion and argument must be correct)?

1* ☐ The p-value for weight (x_1) is less than 0.05, hence, there is a significant relationship between blood pressure and weight.

2 ☐ The p-value for weight (x_1) is less than 0.05, hence, there is not a significant relationship between blood pressure and weight.

3 ☐ The t-test statistic for weight (x_1) is greater than $t_{crit} = 1.96$, hence, there is not a significant relationship between blood pressure and weight.

4 ☐ The t-test statistic for weight (x_1) is greater than $t_{crit} = 1.96$, hence, there is a significant relationship between blood pressure and weight.

5 ☐ The t-test statistic for weight (x_1) is greater than $t_{crit} = 0.05$, hence, there is not a significant relationship between blood pressure and weight.

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

We reject the null hypothesis if the p-value is less than the significance level α, which is the case with respect to this question. When we reject we conclude that we have strong evidence against the null hypothesis ($H_0 : \beta_1 = 0$) and, hence, can state that there is a signficant relationship between x_1 and the response. In other words, the slope β_1 is significantly different from zero.

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

Continue on page 19

Exercise VIII

A fish farmer research group conducted an experiment to find out if there is a significant difference between five selected fish feeds (treatments), as well as a difference in the locations. The fish were farmed on three different locations, where at each location they were divided into groups, which each received one of the five different feeds. The weight of each group were measured after a period, here presented in kg:

	Location 1	Location 2	Location 3
Treatment 1	232.8	225.3	226.1
Treatment 2	201.6	214.9	205.3
Treatment 3	189.6	193.3	180.8
Treatment 4	184.5	209.3	179.0
Treatment 5	270.1	244.2	207.1

A two-way ANOVA was applied and the following result was obtained (note that some values have been replaced by an 'X'):

```
anova(lm(Weight ~ Location + Treatment))

## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq F value   Pr(>F)
## Location    2  959.1  479.54  2.1576       X
## Treatment   4 6329.2 1582.31  7.1192       X
## Residuals   8 1778.1  222.26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question VIII.1 (20)

On a significance level of $\alpha = 5\%$ which conclusions can be drawn (both conclusions and arguments must be correct)?

1 ☐ There are significant effects of both Location and Treatment, since the relevant p-values are 1.55% and 1.67%, respectively.

2 ☐ There are no significant effects of neither Location nor Treatment, since the relevant p-values are 15.5% and 16.7%, respectively.

3 ☐ There is a significant effect of Location but no significant effect of Treatment, since the relevant p-values are 1.55% and 16.7%, respectively.

4 ☐ There are significant effects of both Location and Treatment, since the relevant p-values are 1.78% and 0.95%, respectively.

5* ☐ There are no significant effect of Location, but there is significant effect of Treatment, since the relevant p-values are 17.8% and 0.95%, respectively.

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

```
anova(lm(y ~ Location + Treatment))

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Location   2  959.1  479.54  2.1576 0.178073
## Treatment  4 6329.2 1582.31  7.1192 0.009538 **
## Residuals  8 1778.1  222.26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

1 - pf(2.1576, 2, 8)

## [1] 0.1780714

1 - pf(7.1192, 4, 8)

## [1] 0.009538452
```

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

Question VIII.2 (21)

The applied model is:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

where Y_{ij} represents the weight of the (i, j)'th group, i indicates the Location and j indicates the Treatment.

Which of the following statements is not correct about the model and the analysis above?

1 ☐ There are in total fifteen Y_{ij}'s and they are stochastic variables.

2 ☐ The errors are ε_{ij} and they are assumed i.i.d. for all i's and j's.

3 ☐ The tested hypotheses in the analysis are $H_0 : \alpha_i = 0$ and $H_0 : \beta_j = 0$ for all i's and j's, respectively.

4* ☐ The model predictions are $\mu + \alpha_i + \beta_j + \varepsilon_{ij}$.

5 ☐ σ^2 is estimated with the *MSE*.

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

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

Question VIII.3 (22)

If the test for significant effect of Treatment in the two-way ANOVA is carried out at a significance level of 5%, what significance level should be used for Bonferroni corrected post-hoc comparisons, when all possible Treatment comparisons are carried out?

1 ☐ 0.2%

2 ☐ 0.3125%

3* ☐ 0.5%

4 ☐ 1.875%

5 ☐ 2.5%

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

```
k <- 5
k*(k-1)/2

## [1] 10
```

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

Continue on page 22

Exercise IX

The amount of gluten in oat flour is important, if the flour should be sold as gluten free. In an experiment, the amount of gluten has been measured in a sample of oat flour. The flour was mixed with a special mixer (called A), to try to make the gluten content as homogeneous as possible. 10 small test portions were taken from the flour after mixing and the gluten content was measured in each of them by ELISA tests. The results are stored in `glutenA`. The unit is ppm (parts per million).

Question IX.1 (23)

Which of the following R codes calculates a 95% non-parametric bootstrap confidence interval for the standard deviation of measured gluten content?

1 ☐
```
simsamples <- replicate(10000, sample(glutenA, replace = FALSE))
simmeans <- apply(simsamples, 2, mean)
quantile(simmeans, c(0.025, 0.975))
```

2 ☐
```
simsamples <- replicate(10000, sample(glutenA, replace = FALSE))
simmeans <- apply(simsamples, 2, sd)
quantile(simmeans, c(0.025, 0.975))
```

3 ☐
```
simsamples <- replicate(10000, sample(glutenA, replace = TRUE))
simmeans <- apply(simsamples, 2, sd)
quantile(simmeans, c(0.05, 0.95))
```

4* ☐
```
simsamples <- replicate(10000, sample(glutenA, replace = TRUE))
simmeans <- apply(simsamples, 2, sd)
quantile(simmeans, c(0.025, 0.975))
```

5 ☐
```
simsamples <- replicate(10000, sample(glutenA, replace = FALSE))
simmeans <- apply(simsamples, 2, sd)
quantile(simmeans, c(0.05, 0.95))
```

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

See Method 4.15 and Example 4.16.

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

Question IX.2 (24)

Let us now assume that the measured gluten content follows a normal distribution. Which of the following R codes calculates a 95% parametric bootstrap confidence interval for the standard deviation of the measured gluten content?

1 ☐
```
simsamples <- replicate(10000, rnorm(10,mean(glutenA),sd(glutenA)))
simsds <- apply(simsamples, 2, mean)
quantile(simsds, c(0.025, 0.975))
```

2* ☐
```
simsamples <- replicate(10000, rnorm(10,mean(glutenA),sd(glutenA)))
simsds <- apply(simsamples, 2, sd)
quantile(simsds, c(0.025, 0.975))
```

3 ☐
```
simsamples <- replicate(10000, rnorm(10,mean(glutenA),var(glutenA)))
simsds <- apply(simsamples, 2, sd)
quantile(simsds, c(0.025, 0.975))
```

4 ☐
```
simsamples <- replicate(10000, sample(glutenA,replace = FALSE))
simsds <- apply(simsamples, 2, sd)
quantile(simsds, c(0.025, 0.975))
```

5 ☐
```
simsamples <- replicate(10000, sample(glutenA,replace = TRUE))
simsds <- apply(simsamples, 2, sd)
quantile(simsds, c(0.025, 0.975))
```

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

See Method 4.7 and Example 4.8.

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

Continue on page 24

Question IX.3 (25)

A new portion of flour from the same lot of oat flour is now considered. This flour was mixed with another type of mixer, B. 10 small test portions were taken from the flour after mixing and the gluten content was measured in each of them by ELISA tests. The results are stored in `glutenB`.

We now want to compare the two different mixers, in terms of how homogeneous the gluten content is in the flour after mixing. We therefore want to assess the difference in the standard deviation of the gluten content between samples taken from flour mixed with mixer A and B. We still assume that the measured gluten content follows a normal distribution.

Which of the following R codes calculates a 95% parametric bootstrap confidence interval for the difference in the standard deviations between measurements taken from flour from mixer A and B?

1* ☐
```
simAsamples <- replicate(10000, rnorm(10,mean(glutenA),sd(glutenA)))
simBsamples <- replicate(10000, rnorm(10,mean(glutenB),sd(glutenB)))
simDifsds <- apply(simAsamples,2,sd) - apply(simBsamples,2,sd)
quantile(simDifsds, c(0.025, 0.975))
```

2 ☐
```
simsamples <- replicate(10000, rnorm(10,mean(glutenA)-mean(glutenB),
  sd(glutenA)-sd(glutenB)))
simDifsds <- apply(simsamples,2,sd)
quantile(simDifsds, c(0.025, 0.975))
```

3 ☐
```
simAsamples <- replicate(10000, rnorm(10,mean(glutenA),sd(glutenA)))
simBsamples <- replicate(10000, rnorm(10,mean(glutenB,sd(glutenB))))
simDifmeans <- apply(simAsamples,2,mean) - apply(simBsamples,2,mean)
quantile(simDifmeans, c(0.025, 0.975))
```

4 ☐
```
simAsamples <- replicate(10000, rnorm(10,mean(glutenA),var(glutenA)))
simBsamples <- replicate(10000, rnorm(10,mean(glutenB,var(glutenB))))
simDifsds <- apply(simAsamples,2,sd) - apply(simBsamples,2,sd)
quantile(simDifsds, c(0.025, 0.975))
```

5 ☐
```
simsamples <- replicate(10000, rnorm(10,mean(glutenA-glutenB),
  sd(glutenA-glutenB)))
simDifsds <- apply(simsamples,2,sd)
quantile(simDifs, c(0.025, 0.975))
```

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

See Method 4.10 and Example 4.11.

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

Question IX.4 (26)

The confidence interval from the previous question was found to be $[-4.43; 1.53]$. Which of the below statements are correct (both conclusion and argument must be correct)?

1 ☐ Since the confidence interval includes 0, we conclude that the standard deviations are significantly different. Thus, it seems from our experiment that mixer B is better than mixer A.

2 ☐ Since the confidence interval includes 0, we fail to reject that the standard deviations are equal. Thus, we conclude that the mean gluten content is equal in the two portions.

3 ☐ Since the confidence interval includes 0, we conclude that the mean gluten content is equal in the two portions.

4 ☐ Since the confidence interval includes 0, we conclude that the standard deviations are significantly different. Thus, it seems from our experiment that mixer A is better than mixer B.

5* ☐ Since the confidence interval includes 0, we fail to reject that the standard deviations are equal. Thus, we cannot conclude that one mixer is better than the other.

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

Since the confidence interval for the difference in standard deviations includes 0, the standard deviations are not significantly different. That rules out answer 1 and 4. Our confidence interval says something about the difference in the standard deviations and not the means, and that rules out answer 2 and 3. Since the interval for the difference in standard deviations includes 0, we cannot reject that they are equal. The standard deviation of the gluten content in the test portions is an indication of how well the flour is mixed (the lower the sd, the better it is mixed). Therefore answer 5 is correct.

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

Continue on page 26

Exercise X

A Palestinian university conducted a survey to investigate the reasons behind keeping webcam turned off during online learning. 1268 students responded. The following count data shows the responses to the statement "Teachers don't ask us to turn on the webcam".

Gender	Agree	Disagree	Indifferent	Row Total
Male	339	42	35	416
Female	746	61	45	852
Column Total	1085	103	80	1268

Question X.1 (27)

What is the 95% confidence interval for the proportion of "female" students based on the data given above?

1 ☐ $[0.285, 0.340]$

2 ☐ $[0.0.302, 0.353]$

3 ☐ $[0.561, 0.615]$

4* ☐ $[0.646, 0.698]$

5 ☐ $[0.659, 0.715]$

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

This answer is given by the formula

$$\hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{1-\alpha/2} \tag{1}$$

which can be calculated in R by

```
n <- 1268
p <- 852/n
p + c(-1, 1) * sqrt(p * (1 - p) / n) * qnorm(0.975)

## [1] 0.6460817 0.6977669
```

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

Question X.2 (28)

What is the 95% confidence interval for the difference in the proportion of male and female students' responses selecting "agree" ($p_{\text{agree when female}} - p_{\text{agree when male}}$)?

1 ☐ $[0.003, 0.194]$

2 ☐ $[0.013, 0.113]$

3* ☐ $[0.017, 0.104]$

4 ☐ $[0.284, 0.357]$

5 ☐ $[0.336, 0.414]$

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

The CI is calculated by

$$\hat{p}_1 - \hat{p}_2 \pm \sqrt{\hat{p}_1(1-\hat{p}_1/n_1 + \sqrt{\hat{p}_2(1-\hat{p}_2}/n_2} \tag{2}$$

The result can be found in R by

```
p1 <- 746/852
p2 <- 339/416
p1-p2 + c(-1,1) * sqrt(p1*(1-p1)/852+p2*(1-p2)/416)*qnorm(0.975)

## [1] 0.01727775 0.10408826
```

which is answer no 3.

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

Question X.3 (29)

What is the expected number of students with gender "male" and "agree" under the null-hypothesis of independence between gender and agreement?

1 ☐ 111.22

2 ☐ 227.78

3 ☐ 233.08

4 ☐ 266.20

5* ☐ 355.96

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

The expected number under the null hypothesis for each cell is found as

$$\text{"column total"} \cdot \frac{\text{"row total"}}{\text{"total"}},$$

for table cell (1,1), which is the number of male students who responded "agree". So, the answer is

$$e_{11} = 1085 \cdot \frac{416}{1268} = 355.96.$$

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

Question X.4 (30)

The null-hypothesis of independence between gender and agreement with the statement is to be tested by χ^2-test.

What is the relevant critical value to use for testing whether there is a significant gender difference in the students' responses to the statement (with significance level $\alpha = 0.05$)?

1 ☐ 3.841

2* ☐ 5.991

3 ☐ 7.815

4 ☐ 9.210

5 ☐ 12.59

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

This is a so-called null hypothesis of homogeneity in a 2 x 3 frequency table (r x c) table, see Method 7.22). The critical value for the χ^2-test is based on the χ^2 distribution with $(r-1)(c-1) = (2-1)(3-1) = 2$ degrees of freedom. Hence the correct answer is: $\chi^2_{0.95}(2) = 5.991$.

```
qchisq(0.95, 2)

## [1] 5.991465
```

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

Continue on page 29

SÆTTET ER SLUT. God jul!