Written examination: 22. May 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 11 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	I.2	I.3	II.1	II.2	III.1	III.2	III.3	IV.1	IV.2
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer										

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

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

The exam paper contains 23 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

We are considering a machine for producing certain items. When it's functioning properly, 3% of the items produced are defective. Assume that we will randomly select ten items produced on the machine and that we are interested in the number of defective items found.

Question I.1 (1)

What is the probability of finding no defect items?

1 ☐ 0.0009

2 ☐ 0.0582

3 ☐ 0.4900

4 ☐ 0.737

5 ☐ 0.9127

Question I.2 (2)

What is the number of defects, where there is 98% or higher probability of obtaining this number or fewer defects in the experiment?

1 ☐ 1

2 ☐ 2

3 ☐ 3

4 ☐ 5

5 ☐ 8

Question I.3 (3)

In another planned experiment the outcome is described by the random variable X. The probability density function for X is:

X	0	1	2	3
$f(x)$	0.1	0.3	0.4	0.2

The mean is $\mathrm{E}(X) = 1.7$. Which of the following expressions calculates the variance?

1 ☐ $\mathrm{V}(X) = 0.1 \cdot 0 + 0.3 \cdot 1 + 0.4 \cdot 2 + 0.2 \cdot 3$

2 ☐ $\mathrm{V}(X) = 0.1 \cdot 0 + 0.3 \cdot 1 + 0.4 \cdot 4 + 0.2 \cdot 9$

3 ☐ $\mathrm{V}(X) = 0.1 \cdot 2.89 + 0.3 \cdot 0.49 + 0.4 \cdot 0.09 + 0.2 \cdot 1.69$

4 ☐ $\mathrm{V}(X) = 0.1 \cdot 2.89 + 0.3 \cdot 7.29 + 0.4 \cdot 13.69 + 0.2 \cdot 22.09$

5 ☐ $\mathrm{V}(X) = 0.1 \cdot (-1.3) + 0.3 \cdot (-0.7) + 0.4 \cdot 0.3 + 0.2 \cdot 1.3$

Continue on page 4

Exercise II

The Danish energy company Ørsted made a survey in 2017 in different countries. The survey was about the opinion of people on climate change topics. For each country a randomly selected sample was obtained representative of the population in terms of age, gender, region and income.

One of the questions was: "How important do you think it is to create a world fully powered by renewable energy?".

Let the proportion who answers yes to the question in China be p_1. Similarly, let the proportion who answers yes to the question in Denmark be p_2.

In China $x_1 = 1920$ answered yes out of $n_1 = 2000$ people being asked, and in Denmark $x_2 = 1801$ answered yes out of $n_2 = 2024$ people being asked.

Question II.1 (4)

What is the estimate of the standard error of the estimated proportion who answered yes in Denmark?

1 ☐ $\hat{\sigma}_{\hat{p}_2} = 0.00696$

2 ☐ $\hat{\sigma}_{\hat{p}_2} = 0.0114$

3 ☐ $\hat{\sigma}_{\hat{p}_2} = 0.0136$

4 ☐ $\hat{\sigma}_{\hat{p}_2} = 0.0179$

5 ☐ $\hat{\sigma}_{\hat{p}_2} = 0.0834$

Question II.2 (5)

Given a significance level of $\alpha = 0.01$, what is the conclusion concerning the usual two-sample proportion test of the null hypothesis:

$$H_0 : p_1 = p_2$$

(both conclusion and argument must be correct)?

1 ☐ The null hypothesis is rejected because $0.96 \neq 0.89$, hence the two proportions are significantly different.

2 ☐ The null hypothesis is accepted because $0.96 - 0.89 > 0.01$, hence the two proportions are not significantly different.

3 ☐ The null hypothesis is rejected because $0 \notin [0.060, 0.081]$, hence the two proportions are significantly different.

4 ☐ The null hypothesis is accepted because $0 \notin [0.060, 0.081]$, hence the two proportions are not significantly different.

5 ☐ The null hypothesis is rejected because $0 \notin [0.049, 0.091]$, hence the two proportions are significantly different.

Continue on page 6

Exercise III

To compare two programs for training industrial workers to perform a skilled job, 20 workers were included in an experiment. Of these, 10 were selected at random and trained by "Method 1"; the remaining 10 workers were trained by "Method 2". After completion of training, all the participants were subjected to a time-and-motion test that records the speed of performance of the skilled job.

The following observations in minutes were obtained (the sample mean and standard deviation are included for each sample):

	1	2	3	4	5	6	7	8	9	10	Mean	Std. dev.
Method 1	11.9	22.5	12.4	16.5	12.6	17.2	9.8	15.0	17.1	14.1	14.9	3.6
Method 2	18.9	20.1	14.6	16.5	16.2	24.5	17.7	24.1	17.3	20.2	19.0	3.3

Question III.1 (6)

Assuming that the true variances of the two methods are the same, what is the estimated pooled standard deviation?

1 ☐ $s_{\text{pooled}} = \frac{3.6+3.3}{2}$

2 ☐ $s_{\text{pooled}} = \sqrt{\frac{3.6^2+3.3^2}{2}}$

3 ☐ $s_{\text{pooled}} = \frac{3.6^2+3.3^2}{2}$

4 ☐ $s_{\text{pooled}} = \sqrt{\frac{3.6+3.3}{2}}$

5 ☐ It is not possible to calculate the pooled standard deviation since the two samples variances are not equal, i.e. $s^2_{\text{Method1}} \neq s^2_{\text{Method2}}$.

Question III.2 (7)

We run a pooled t-test in R with the samples and obtain the output below:

```
##
##  Two Sample t-test
##
## data:  method1 and method2
## t = -2.6559, df = 18, p-value = 0.01609
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.3397176 -0.8562943
## sample estimates:
## mean of x mean of y
##  14.91241  19.01042
```

Which statement is correct (both of the two parts of the statement must be correct)?

1 ☐ We accept the null hypothesis of equal mean speed performances. Our risk of making a Type I error is 95%.

2 ☐ We reject the null hypothesis of equal mean speed performances. Our risk of making a Type I error is 95%.

3 ☐ We accept the null hypothesis of equal mean speed performances. Our risk of making a Type I error is 5%.

4 ☐ We reject the null hypothesis of equal mean speed performances. Our risk of making a Type I error is 5%.

5 ☐ We cannot apply the pooled t-test under the assumption of equal population variances.

Question III.3 (8)

We now want to plan a new experiment, where we control the power of the statistical test to differentiate the means, still with equally many observations in the two groups. We want to use a significance level of $\alpha = 1\%$ and we want to have a 98% probability for detecting a difference in means of 5 minutes.

Independent of the results in the questions above, we will use a guess of the population variance of 16.

What is the smallest number of observations one should take from each group in order to satisfy the requirements above?

1 ☐ At least 12

2 ☐ At least 18

3 ☐ At least 30

4 ☐ At least 45

5 ☐ At least 62

Continue on page 8

Exercise IV

A company assembles machines from various components. Assume that the lifetime of components in a machine can be modelled independently with the same exponential distribution.

Question IV.1 (9)

If the components mean lifetime is 3 years, which of the following R-codes calculates the probability that a randomly selected component lasts longer than one year?

1 ☐ `1 - dexp(0, rate=1/3)`

2 ☐ `pexp(1, rate=3)`

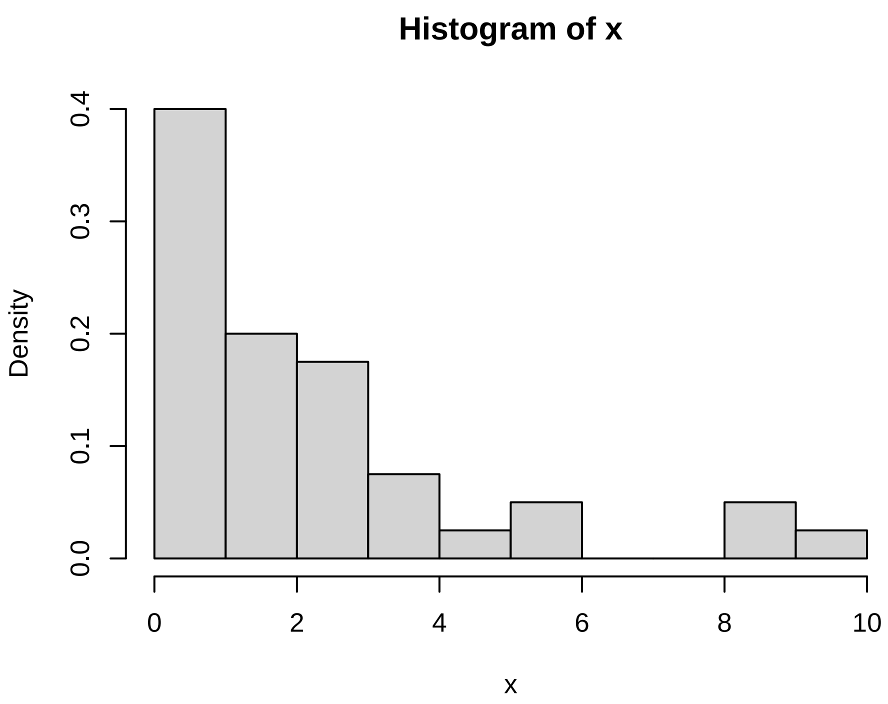
3 ☐ `1 - pexp(0, rate=1/3)`

4 ☐ `1 - pexp(1, rate=1/3)`

5 ☐ `dexp(0, rate=3)`

Question IV.2 (10)

An experiment was carried out by measuring when components stopped working when exposed to an accelerated ageing environment. A density histogram of the obtained sample is plotted below:



What is the number of observations in the sample?

1 ☐ $n = 80$

2 ☐ $n = 160$

3 ☐ $n = 320$

4 ☐ $n = 480$

5 ☐ This cannot be known with the provided information.

Question IV.3 (11)

A parametric bootstrap 95% confidence interval for the mean was calculated with the R-code below. The obtained sample was loaded in the vector `x`:

```
# Set the number of simulations:
k <- 100000
# Simulate k samples
simSamples <- replicate(k, rexp(length(x), 1/mean(x)))
# Compute the simulated means
simMean <- apply(simSamples, 2, mean)
# Quantiles for the confidence interval
quantile(simMean, c(0.025, 0.975))

##     2.5%    97.5%
## 1.561813 2.914021
```

Based on this analysis what is the conclusion of a test of the null hypothesis

$$H_0 : \mu = 2$$

on significance level $\alpha = 0.05$ (both the argument and the conclusion must be correct)?

1 ☐ The null hypothesis is accepted, since $2 \in [1.56, 2.91]$, hence we conclude that the mean might be 2.

2 ☐ The null hypothesis is accepted, since $2 \in [1.56, 2.91]$, hence we conclude that the mean is 2.

3 ☐ The null hypothesis is rejected, since $2 \in [1.56, 2.91]$, hence we conclude that the mean might be 2.

4 ☐ The null hypothesis is rejected, since $2 \in [1.56, 2.91]$, hence we conclude that the mean is 2.

5 ☐ The null hypothesis is rejected, since $2 \in [1.56, 2.91]$, hence we conclude that the mean is different from 2.

Continue on page 10

Exercise V

This exercise is about calculating standard deviation and variance of functions of random variables.

Question V.1 (12)

Using simulation, what is the standard deviation of Y approximately when

$$Y = e^{X_1} + X_2^4 + X_1 \cdot X_2$$

where X_1 and X_2 are independent and both standard normal distributed?

1 ☐ $\sigma_Y \approx 3.3$

2 ☐ $\sigma_Y \approx 10$

3 ☐ $\sigma_Y \approx 100$

4 ☐ $\sigma_Y \approx 920$

5 ☐ $\sigma_Y \approx 9800$

Question V.2 (13)

Let Y be defined by

$$Y = X_1^3 + 5X_2$$

The two random variables X_1 and X_2 are independent and have standard deviations σ_1 and σ_2, respectively. Let x_1 and x_2 be observations of X_1 and X_2, respectively

What is the linear approximation to the variance of Y, derived using the propagation of error method?

1 ☐ $\text{V}(Y) \approx 9x_1^4\sigma_1 + 25\sigma_2$

2 ☐ $\text{V}(Y) \approx 3x_1^2\sigma_1^2 + 5\sigma_2^2$

3 ☐ $\text{V}(Y) \approx 9x_1^4\sigma_1^2 + 25\sigma_2^2$

4 ☐ $\text{V}(Y) \approx 9x_1^2\sigma_1 + 25x_2\sigma_2$

5 ☐ $\text{V}(Y) \approx 3x_1^4\sigma_1 + 5x_2\sigma_2$

Continue on page 11

Exercise VI

12 observations of manganese (Mn) at six different concentrations were analysed using *inductively coupled plasma atomic emission spectroscopy* (ICP-AES).

Manganese concentrations are measured in ppb (parts per billion). The data was read into R by:

```
# Manganese concentrations
x <- c(0, 0, 2, 2, 4, 4, 6, 6, 8, 8, 10, 10)
# ICP-AES values
y <- c(114, 14, 870, 1141, 2087, 2212, 3353, 2633, 3970, 4299, 4950, 5207)
```

The linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ where } \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d. for } i = 1, \ldots, 12.$$

was set up, where Y_i is the ICP-AES value and x_i the manganese concentration of the i'th observation.

Note, that in the remaining of the exercise the normal distribution and i.i.d. assumptions of the errors are implicit (hence not written with the model).

Question VI.1 (14)

What is the estimate of β_1?

1 ☐ 49.2

2 ☐ 504.3

3 ☐ 511.0

4 ☐ 520.7

5 ☐ 2570

Question VI.2 (15)

Researchers would like to know the uncertainty in ICP-AES value for a new observation with manganese concentration 5 ppb. What is the 95% prediction interval for this concentration?

1 ☐ [2087, 3054]

2 ☐ [2437, 2705]

3 ☐ [465, 544]

4 ☐ [2388, 2656]

5 ☐ [2038, 3005]

Question VI.3 (16)

We wish to test the hypothesis $H_0 : \beta_0 = 0$, as this would indicate whether the expected ICP-AES value is 0 for a manganese concentration of 0 ppb.

Which of the following statements is correct?

1 ☐ We accept the null hypothesis, since p-value is 0.006.

2 ☐ We reject the null hypothesis, since p-value is 0.006.

3 ☐ We accept the null hypothesis, since $|1 - \hat{\beta}_0|$ is less than the standard deviation.

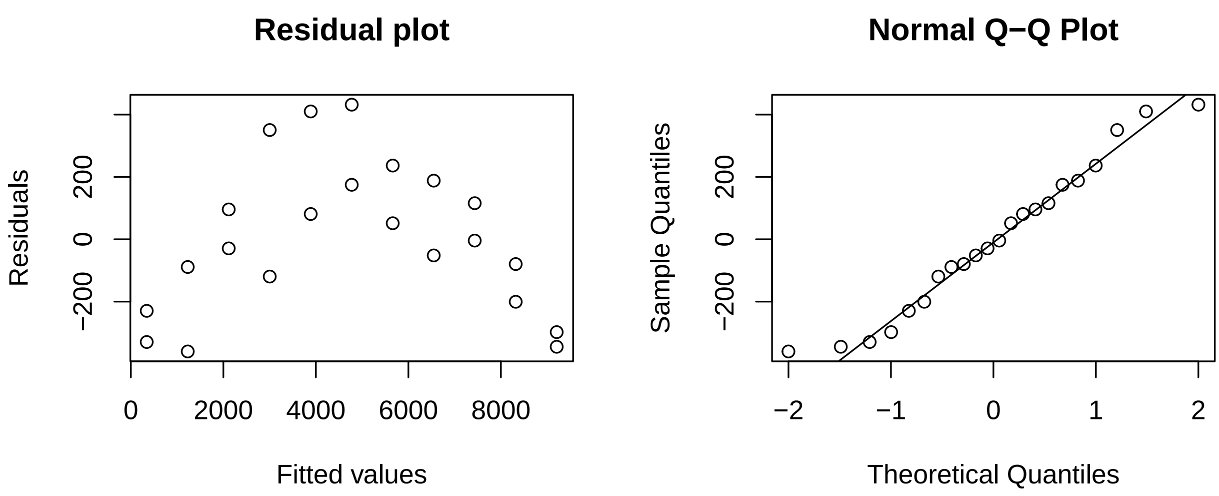
4 ☐ We accept the null hypothesis, since p-value is 0.655.

5 ☐ We reject the null hypothesis, since p-value is 0.655.

Question VI.4 (17)

Subsequently, we receive additional data with higher manganese concentrations (up to 20 ppb). A new linear regression $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ was carried out.

The plots below show a residual plot and a normal q-q plot of the residuals:



Which of the following statements is the correct interpretation of the two plots?

1 ☐ The residual plot looks questionable. This indicates a problem with the normality assumption.

2 ☐ The residual plot looks questionable. This indicates a problem with the linear dependence assumption.

3 ☐ The residual plot looks (reasonably) fine, but the q-q plot looks questionable. This indicates a problem with the linear dependence assumption.

4 ☐ We see no linear tendency in the residual plot. This is evidence for the null hypothesis of no significant effect of concentration on ICP-AES value.

5 ☐ Neither the residual plot nor the q-q plot are related to the validity of the model or the associated null hypotheses.

Question VI.5 (18)

Finally, the curve-linear model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ was fitted to the new data. The new data was stored in the data.frame `Mangan2`. The result is:

```
x2 <- Mangan2$x^2
fit <- lm(y ~ x + x2, data = Mangan2)
summary(fit)

##
## Call:
## lm(formula = y ~ x + x2, data = Mangan2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -251.95  -63.95   17.22   69.20  218.05
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.6469    74.8293   0.169    0.868
## x           553.4783    17.4075  31.795  < 2e-16 ***
## x2           -5.5157     0.8383  -6.580 2.68e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 138.9 on 19 degrees of freedom
## Multiple R-squared:  0.9979,Adjusted R-squared:  0.9977
## F-statistic:  4500 on 2 and 19 DF,  p-value: < 2.2e-16
```

Which of the following statements is correct, given a significance level of $\alpha = 1\%$?

1 ☐ There is no significant model improvement of including the quadratic term, since the p-value for β_1 is less than the p-value for β_2.

2 ☐ The value of ICP-AES increases in average by $(-5.5)^2 = 30.25$, when the concentration increases by 1 ppb.

3 ☐ The value of ICP-AES increases in average by 553.5, when the concentration increases by 1 ppb.

4 ☐ Neither β_1 nor β_2 are significantly different from zero, since their associated p-values are below 0.01.

5 ☐ The model is able to explain more than 99% of the observed variation in data.

Continue on page 15

Exercise VII

For the transition to a CO_2 emission free energy system fossil fuels must be phased out. For example heating with natural gas must be replaced with other sources. District heating (DH), if based on renewable sources, is often a good alternative.

When it is decided whether DH should be established in a new area an information meeting is held, where the DH company informs about the cost and possibility for connecting to the DH grid and how it compares with other heating alternatives.

A survey was carried out among the house owners participating in an information meeting to determine if the meeting changed their opinion to join the district heating.

Their opinions whether they will connect to the DH were collected anonymously before and after the meeting. Their answers were:

	Before	After	Sum
Yes	18	28	46
No	22	14	36
Not decided	25	17	42
Sum	65	59	124

The usual null hypothesis, that the proportion was the same before and after, is:

$$H_0 : p_{i,1} = p_{i,2}, \text{ for all rows } i = 1, 2, 3.$$

Question VII.1 (19)

What is the expected number of people under the null hypothesis answering 'Not decided' after the information meeting?

1 ☐ $17/124$

2 ☐ $25 \cdot 59/124$

3 ☐ $42 \cdot 17/59$

4 ☐ $59 \cdot 42/124$

5 ☐ $17 \cdot 25/59$

Question VII.2 (20)

The following R code was run:

```r
chisq.test(matrix(c(18, 28, 22, 14, 25, 17), ncol = 2, byrow = TRUE),
           correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  matrix(c(18, 28, 22, 14, 25, 17), ncol = 2, byrow = TRUE)
## X-squared = 5.1973, df = 2, p-value = 0.07437
```

What is the correct conclusion regarding of the null hypothesis tested at a significance level $\alpha = 0.05$?

1 ☐ The null hypothesis is rejected since the p-value is above the significance level.

2 ☐ The null hypothesis is rejected since the p-value is below the significance level.

3 ☐ The null hypothesis is accepted since the p-value is above the significance level.

4 ☐ The null hypothesis is accepted since the p-value is below the significance level.

5 ☐ None of the conclusions stated above are correct as sufficient information is not available to make a decision.

Question VII.3 (21)

What is the critical level, i.e. a χ^2 quantile, for testing the null hypothesis at a significance level of $\alpha = 0.01$?

1 ☐ 9.21

2 ☐ 2.32

3 ☐ 1.96

4 ☐ 0.196

5 ☐ 0.103

Continue on page 17

Exercise VIII

A brochure, inviting subscriptions for a new diet program, states that the participants are expected to lose 23 pounds in five weeks. Let X denote the weight loss. From data of five-week weight losses of $n = 56$ participants the sample mean and the standard deviation were found to be $\bar{x} = 21.5$ and $s = 9.8$ pounds, respectively.

To investigate the claim of weight loss the hypothesis

$$H_0 : \mu = 23$$

must be tested with the obtained data.

Question VIII.1 (22)

Which of the following statements is correct, when applying significance level $\alpha = 0.05$ (both argument and conclusion must be correct)?

1 ☐ The 95% confidence interval is $[18.88, 24.12]$. The hypothesized weight loss of 23 pounds is contained in the interval, hence the statement can be substantiated.

2 ☐ The 95% confidence interval is $[18.88, 24.12]$. The hypothesized weight loss of 23 pounds is contained in the interval, hence the statement can NOT be substantiated.

3 ☐ The 95% confidence interval is $[19.30, 23.69]$. The hypothesized weight loss of 23 pounds is contained in the interval, hence the statement can be substantiated.

4 ☐ The 95% confidence interval is $[19.30, 23.69]$. The hypothesized weight loss of 23 pounds is contained in the interval, hence the statement can NOT be substantiated.

5 ☐ The 95% confidence interval is $[20.88, 22.12]$. The hypothesized weight loss of 23 pounds is NOT contained in the interval, hence the statement can be substantiated.

Question VIII.2 (23)

What is the value of the test statistic used for testing the hypothesis?

1 ☐ $t_{\text{obs}} = 1.135$

2 ☐ $t_{\text{obs}} = -1.135$

3 ☐ $t_{\text{obs}} = -1.145$

4 ☐ $t_{\text{obs}} = 16.42$

5 ☐ $t_{\text{obs}} = -16.42$

Continue on page 18

Exercise IX

More than two million visitors browse the DTU website every month, which enables DTU to attract researchers, students, and others. The website has in average seven visitors per minute. It's assumed that the rate is constant.

Question IX.1 (24)

What is the probability that there are two or more visitors on the website in a randomly selected one-minute period?

1 ☐ 0.09

2 ☐ 0.64

3 ☐ 0.77

4 ☐ 0.97

5 ☐ 0.99

Question IX.2 (25)

What is the probability that there are no visitors in a randomly selected 30-second period?

1 ☐ 0.03

2 ☐ 0.07

3 ☐ 0.18

4 ☐ 0.43

5 ☐ 0.78

Continue on page 19

Exercise X

The following measurements have been carried out across 3 groups. The values are given in the table below:

Group 1	Group 2	Group 3
1.89	3.15	1.54
2.35	2.16	2.02
1.68	2.40	2.01
2.11	2.59	2.11

The data can be read into R by:

```
y <- c(1.89, 2.35, 1.68, 2.11, 3.15, 2.16, 2.40, 2.59, 1.54, 2.02, 2.01, 2.11)
```

Question X.1 (26)

The overall mean $\bar{y}$ and the means within each group $\bar{y}_i$ (for $i = 1, 2, 3$) are given below:

$\bar{y}$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$
2.17	2.01	2.58	1.92

We perform a one-way analysis of variance (ANOVA). What are the total sum of squares (SST), treatment sum of squares (SS(Tr)) and sum of squared errors (SSE)?

1 ☐ $SST = 0.18,\ SS(Tr) = 0.51,\ SSE = 0.11$

2 ☐ $SST = 0.31,\ SS(Tr) = 0.42,\ SSE = 0.88$

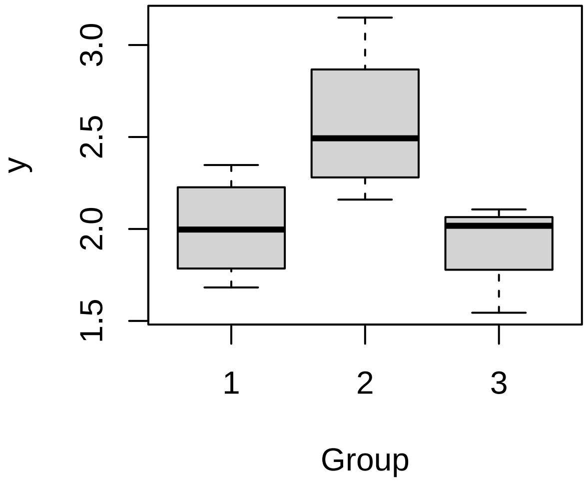
3 ☐ $SST = 2.50,\ SS(Tr) = 2.12,\ SSE = 0.38$

4 ☐ $SST = 1.99,\ SS(Tr) = 1.01,\ SSE = 0.98$

5 ☐ $SST = 4.12,\ SS(Tr) = 0.75,\ SSE = 3.37$

Question X.2 (27)

The data stated above has been visualized using a boxplot:



Which of the following statements is true?

1 ☐ The black lines within the boxes indicate the mean of each sample.

2 ☐ The medians are approximately 2.0, 2.5 and 2.0 for groups 1, 2 and 3, respectively.

3 ☐ The box width is defined as the difference between the upper and lower quartiles, i.e. the difference between the 95th and 5th percentiles.

4 ☐ The box width is defined as the difference between the upper and lower quartiles, i.e. the difference between the 90th and 10th percentiles.

5 ☐ The whiskers of the boxplot define the Interquartile Range, i.e. $IQR = Q3 - Q1$.

Continue on page 21

Exercise XI

As part of a cooperative study on the nutritional quality of oats, 6 varieties of oat kernels with their hulls removed are subjected to a mineral analysis. The plants are grown under four different treatments using a randomized block design and measurements of protein by percent of dry weight are recorded at harvest time.

A two-way ANOVA model for this data is

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

where Y_{ij} is the relative protein content of the i'th oat kernel variant at the j'th treatment and α_i and β_j represent effect sizes corresponding to oat variant and treatment, respectively.

The result of fitting the model is given in the ANOVA table below (note, that some values have been replaced by question marks):

```
## Analysis of Variance Table
## Response: protein
##            Df   Sum Sq  Mean Sq   F value   Pr(>F)
## oat        5    2.2060  0.44120   4.2367    0.01333 *
## treat      ?    0.2554  0.08513   ?         0.50410
## Residuals  15   1.5620  0.10414
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Futhermore, the estimated effect sizes are:

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimated effect	-0.09	0.59	0.37	-0.22	0.53	0.35	0.24	0.37	0.30	0.09

Question XI.1 (28)

The overall mean $\hat{\mu} = \bar{y} = 5.6$ has been estimated from the data. Given the results in the ANOVA table and estimated effect sizes, what is the expected (predicted) protein content for oat kernels of variant 2 at treatment levels 1 and 4, when only significant effects are taken into account at significance level 5%?

1 ☐ $\hat{y}_{21} = 2 \cdot 2.2060$ and $\hat{y}_{24} = 2 \cdot 2.2060$

2 ☐ $\hat{y}_{21} = 2 \cdot 2.2060 + 1 \cdot 0.2554$ and $\hat{y}_{24} = 2 \cdot 2.2060 + 4 \cdot 0.2554$

3 ☐ $\hat{y}_{21} = 0.59$ and $\hat{y}_{24} = 0.59$

4 ☐ $\hat{y}_{21} = 5.6 + 0.24$ and $\hat{y}_{24} = 5.6 + 0.09$

5 ☐ $\hat{y}_{21} = 5.6 + 0.59$ and $\hat{y}_{24} = 5.6 + 0.59$

Question XI.2 (29)

Some of the elements in the ANOVA table above have been replaced by question marks. Which of the following statements is correct for the treatment?

1 ☐ Degrees of freedom is 4 and the observed test statistic is 0.8175.

2 ☐ Degrees of freedom is 3 and the observed test statistic is 0.8175.

3 ☐ Degrees of freedom is 4 and the observed test statistic is 0.4087.

4 ☐ Degrees of freedom is 3 and the observed test statistic is 0.4087.

5 ☐ Degrees of freedom is 4 and the observed test statistic is 1.960.

Question XI.3 (30)

When testing the null hypothesis $H_{0,\text{Oat}} : \alpha_i = 0$, where $i = 1, 2, ..., k$, which implication stated below is correct on a significance level $\alpha = 0.05$?

1 ☐ The probability of making a type I error is 1.33%.

2 ☐ The probability of making a type I error is 95%.

3 ☐ The probability of making a type II error is 98.67%.

4 ☐ Given that the null hypothesis is false, the probability that the test statistic is higher than the observed test statistic is 98.67%.

5 ☐ Given that the null hypothesis is true, the probability that the test statistic is higher than the observed test statistic is 1.33%.

Continue on page 23

The exam is finished. Enjoy the summer!