Written examination: 22. June 2023

Course name and number: **Introduction to Mathematical Statistics (02403)**

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

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

Exercise	VII.5	VII.6	VIII.1	VIII.2	VIII.3	IX.1	IX.2	IX.3	X.1	X.2
Question	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Answer										

The exam paper contains 25 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

A researcher is interested in comparing the average weight gain (in grams) of three different groups of mice fed with three different diets. The data is provided below.

```
group1 <- c(27, 22, 18, 26, 24)
group2 <- c(32, 22, 32, 25, 25)
group3 <- c(29, 25, 30, 30, 24)
```

Question I.1 (1)

Perform a one-way ANOVA and test the usual null hypothesis of equal treatment means at significance level $\alpha = 0.05$. Is there a significant difference in weight gain among the three groups?

1 ☐ The p-value is 0.03. The difference between group means is not significant because the p-value is less than 0.05.

2 ☐ The p-value is 0.1879. The difference between group means is not significant because the p-value is greater than 0.05.

3 ☐ The p-value is 0.1879. The difference between group means is significant because the p-value is greater than 0.05.

4 ☐ The p-value is 0.3758. The difference between group means is significant because the p-value is greater than 0.05.

5 ☐ The p-value is 0.03. The difference between group means is significant because the p-value is less than 0.05.

Continue on page 3

Question I.2 (2)

The experiment described above was repeated (same number of mice) by a second researcher who collected a different data set. Again, one-way ANOVA was used to test for significant difference between treatment means. The following ANOVA table was obtained. Please note that some elements have been replaced by question marks.

```
## Analysis of Variance Table

## Response: weight_gain
##           Df Sum Sq Mean Sq F value Pr(>F)
## group      2  78.53  39.267   1.069 0.3739
## Residuals  ? 440.80       ?
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Which of the following statements is correct?

1 ☐ `Df(Residuals)` = 14 and `Mean Sq(Residuals)` = 31.486.

2 ☐ `Df(Residuals)` = 15 and `Mean Sq(Residuals)` = 29.387.

3 ☐ `Df(Residuals)` = 12 and `Mean Sq(Residuals)` = 36.733.

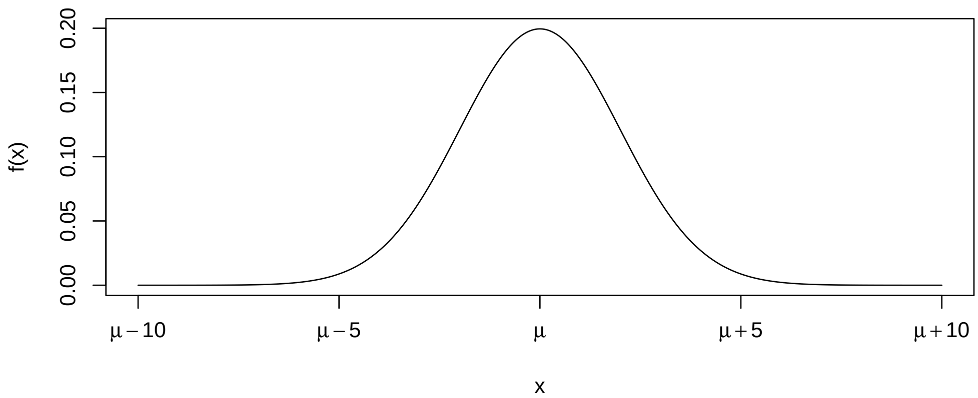
4 ☐ `Df(Residuals)` = 14 and `Mean Sq(Residuals)` = 2.805.

5 ☐ `Df(Residuals)` = 13 and `Mean Sq(Residuals)` = 3.021.

Continue on page 4

Exercise II

Let the random variable X be normal distributed with mean μ and standard deviation $\sigma = 2$, i.e. $X \sim N(\mu, 2^2)$, hence its' pdf is:



Question II.1 (3)

Let another random variable be defined by the function

$$Y_1 = a_1 + b_1 \cdot X + b_2 \cdot X$$

What is the mean and variance of Y_1?

1 ☐ $\mathrm{E}(Y_1) = a_1 + (b_1 + b_2)\mu$ and $\mathrm{V}(Y_1) = (b_1 + b_2)^2 \cdot 4$

2 ☐ $\mathrm{E}(Y_1) = a_1 + b_1 + b_2$ and $\mathrm{V}(Y_1) = b_1^2 + b_2^2$

3 ☐ $\mathrm{E}(Y_1) = a_1 + b_1 + b_2$ and $\mathrm{V}(Y_1) = b_1 + b_2$

4 ☐ $\mathrm{E}(Y_1) = 0$ and $\mathrm{V}(Y_1) = b_1^2 + b_2^2$

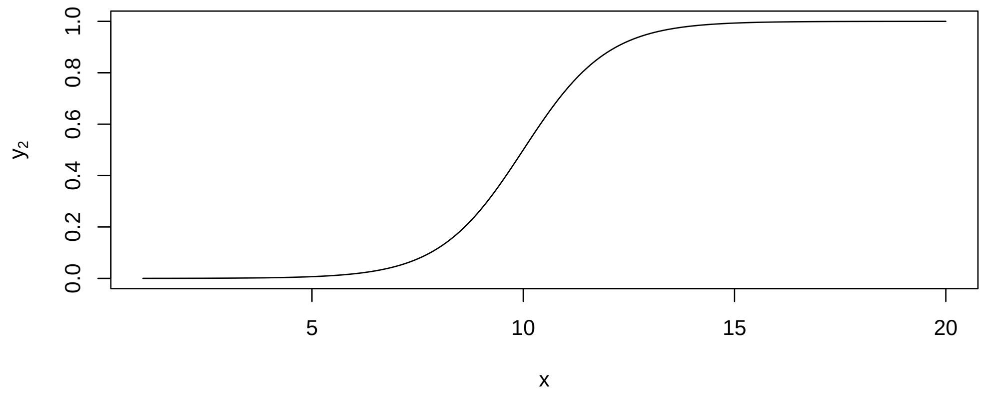
5 ☐ $\mathrm{E}(Y_1) = 0$ and $\mathrm{V}(Y_1) = b_1 + b_2$

Question II.2 (4)

Let another random variable be defined by the function

$$Y_2 = \frac{1}{1 + \exp(a_2 + b_2 \cdot X)}$$

where $a_2 = 10$ and $b_2 = -1$. A plot of this function is:



This function is called the logistic function (or Sigmoid function).

The notation $\mathrm{V}(Y_2|\mu = \mu_0)$ means the variance of Y_2 when μ is equal to μ_0. E.g. $\mathrm{V}(Y_2|\mu = 0)$ is the variance of Y_2 when μ is equal to 0.

Which one of the following statements is correct?

1 ☐ $\mathrm{V}(Y_2|\mu = 0) < \mathrm{V}(Y_2|\mu = 10) < \mathrm{V}(Y_2|\mu = 20)$

2 ☐ $\mathrm{V}(Y_2|\mu = 0) < \mathrm{V}(Y_2|\mu = 20) < \mathrm{V}(Y_2|\mu = 10)$

3 ☐ $\mathrm{V}(Y_2|\mu = 0) = \mathrm{V}(Y_2|\mu = 20) = \mathrm{V}(Y_2|\mu = 10)$

4 ☐ $\mathrm{V}(Y_2|\mu = 0) = \mathrm{V}(Y_2|\mu = 20) < \mathrm{V}(Y_2|\mu = 10)$

5 ☐ $\mathrm{V}(Y_2|\mu = 20) < \mathrm{V}(Y_2|\mu = 10) < \mathrm{V}(Y_2|\mu = 0)$

Continue on page 6

Exercise III

In a statistics class with 589 students, all students have taken an exam with two parts each lasting 2 hours. The instructors of the class are interested in evaluating whether the two exams parts have been equally difficult by comparing the mean scores in the two parts.

Question III.1 (5)

What test should the instructors apply in order to evaluate whether the mean scores of the two exam parts are equal?

1 ☐ A one-way ANOVA

2 ☐ An F-test with 2 and 589 degrees of freedom

3 ☐ A two-sample t-test assuming equal variances in the two groups

4 ☐ A two-sample t-test with a pooled variance

5 ☐ A paired t-test

Question III.2 (6)

One of the instructors gives the same course (and the same exam) at a different university, where 240 students are enrolled in the course and subsequently take the exam. Some summary statistics concerning the exam results are given in the below table:

	University A	University B
Students	589	240
Average score	736.4	769.9
Variance of score	169.1	402.7

When calculating the 90% confidence interval, not assuming equal variances in the two groups, for the difference in mean scores, the instructor has to use a quantile from a t-distribution. The instructor has to use which quantile of the t-distribution with how many degrees of freedom?

1 ☐ The 10% quantile of the t-distribution with 323.93 degrees of freedom

2 ☐ The 90% quantile of the t-distribution with 829 degrees of freedom

3 ☐ The 90% quantile of the t-distribution with 323.93 degrees of freedom

4 ☐ The 95% quantile of the t-distribution with 829 degrees of freedom

5 ☐ The 95% quantile of the t-distribution with 323.93 degrees of freedom

Continue on page 7

Exercise IV

On April 14 1912 the passenger ship Titanic hit an iceberg and sank the following day. The table below shows the number of survivors and total number of passengers distributed on different passenger categories.

Class	1st	2nd	3rd	Crew	Total
Survived	202	117	178	212	709
Total	325	285	706	885	2201

Question IV.1 (7)

Based on the table above, what is a 95% confidence interval for the probability of survival (regardless of passenger category) given the data?

1 ☐ [0.66, 0.70]

2 ☐ [0.30, 0.34]

3 ☐ [0.45, 0.50]

4 ☐ [0.46, 0.49]

5 ☐ [0.66, 0.69]

Question IV.2 (8)

Is there a statistical significant difference in the survival probability between the crew and the 3rd class passengers, using a 5% significance level (both the argument and the conclusion should be correct)?

1 ☐ Yes, since the test statistics for the relevant test is -1.67

2 ☐ No, since the test statistics for the relevant test is 0.66

3 ☐ Yes, since the test statistics for the relevant test is 1.67

4 ☐ No, since the p-value for the relevant test is 0.41

5 ☐ No, since the p-value for the relevant test is 0.56

Continue on page 8

Question IV.3 (9)

Considering the entire table, what is the relevant observed test statistics (q), critical value (CV), and conclusion for a test of the hypothesis that the survival probability is the same across all classes, using significance level $\alpha = 0.05$?

1 ☐ q=187.1, CV=7.8, hence there is a significant difference

2 ☐ q=84.37, CV=15.5, hence there is a significant difference

3 ☐ q=84.37, CV=7.8, hence there is a significant difference

4 ☐ q=187.1, CV=15.5, hence there is not a significant difference

5 ☐ q=84.37, CV=7.8, hence there is not a significant difference

Question IV.4 (10)

We wish to test if the probability of survival of 1st class passengers differs by more than 20 procentage points compared to the average of all other passengers, which of the following statements regarding that is correct (using significance level $\alpha = 0.05$)?

1 ☐ Since $\hat{p}_{1st} - \hat{p}_{rest} = 0.35$ there is a significant difference and it is greater than 0.2

2 ☐ The relevant confidence interval is $[0.29, 0.41]$, and hence the survival probability of 1st class passengers is at least 20 procentage point higher than the survival probability of other passengers

3 ☐ 0.2 is not included in the relevant confidence interval, which is $[0.29, 0.41]$, and hence there is not a significant difference

4 ☐ The relevant confidence interval is $[0.33, 0.37]$, and hence the survival probability of 1st class passengers is at least 20 procentage points higher than the survival probability of other passengers

5 ☐ 0.2 is not included in the relevant confidence interval, which is $[0.33, 0.37]$, and hence there is not a significant difference

Continue on page 9

Exercise V

A school class with 20 children are collecting trash on a beach, it is assumed that the mean value of the collected trash is 1kg/child with a standard deviation of 0.2 kg/child.

Question V.1 (11)

If the amount of trash collected by each child is assumed independent, what is the standard deviation (σ) of all the collected trash then?

1 ☐ $\sigma = 4.0$ kg

2 ☐ $\sigma = 0.8$ kg

3 ☐ $\sigma = 0.18$ kg

4 ☐ $\sigma = 2.0$ kg

5 ☐ $\sigma = 0.89$ kg

Question V.2 (12)

It is decided that the children should go in pairs of two, the standard deviation is still assumed to be 0.2 kg/child, but it is now, in addition, assumed that the correlation between the amount trash collected by the chrildren in the same pair is 0.5. The pairs are assumed independent of each other. What is the standard deviation (σ_{pair}) of the total amount of collected trash?

1 ☐ $\sigma_{pair} = 1.10$ kg

2 ☐ $\sigma_{pair} = 0.84$ kg

3 ☐ $\sigma_{pair} = 3.60$ kg

4 ☐ $\sigma_{pair} = 1.8$ kg

5 ☐ $\sigma_{pair} = 1.89$ kg

Continue on page 10

Question V.3 (13)

After they returned, one of the children had collected 21 items, of which 6 were made of plastic. She is now asked to pick 5 items at random to be discussed. What is the probability that 3 of those are made of plastic?

1 ☐ 0.103

2 ☐ 0.247

3 ☐ 0.119

4 ☐ 0.023

5 ☐ 0.052

Continue on page 11

Exercise VI

The quality assurance department at a candy factory has taken a random sample of 26 chocolate bars of a certain brand. Each chocolate bar in the sample is weighted, and it is found that the average weight is 200.3 grams and the observed standard deviation is 0.75 grams.

Question VI.1 (14)

What is the 95% confidence interval for the standard deviation?

1 ☐ $[0.346, 1.072]$

2 ☐ $[0.588, 1.035]$

3 ☐ $[0.611, 0.981]$

4 ☐ $[0.447, 1.053]$

5 ☐ $[0.462, 1.038]$

Question VI.2 (15)

The candy factory wants to test the null-hypothesis $\mathcal{H}_0 : \mu = 200$ grams (against a two-sided alternative) using a t-test. Which of the following statements is correct based on the hypothesis test? (Both the argument and the conclusion must be correct)

1 ☐ Using a significance level of 5%, the null-hypothesis is rejected since the test statistic is greater than $t_{0.975}(26)$

2 ☐ Using a significance level of 5%, the null-hypothesis is accepted since the test statistic is greater than $t_{0.975}(26)$

3 ☐ Using a significance level of 5%, the null-hypothesis is rejected since the test statistic is greater than $t_{0.975}(25)$

4 ☐ Using a significance level of 10%, the null-hypothesis is accepted since the test statistic is greater than $t_{0.95}(25)$

5 ☐ Using a significance level of 10%, the null-hypothesis is rejected since the test statistic is greater than $t_{0.95}(25)$

Continue on page 12

Question VI.3 (16)

To further investigate the mean weight of the chocolate bars, the candy factory is also planning another experiment. The quality assurance department wants to detect a difference in mean weight of 0.3 grams (against a two-sided alternative hypothesis) while using 0.75 grams as a guess of the standard deviation. Furthermore, the quality assurance department wants to keep both the Type I and the Type II error rates at (or below) 5%. What is the minimum number of chocolate bars to be included in the experiment in order to meet the criteria set by the department?

1 ☐ 10 or 12 depending on whether you apply the normal approximation

2 ☐ 68 or 70 depending on whether you apply the normal approximation

3 ☐ 82 or 84 depending on whether you apply the normal approximation

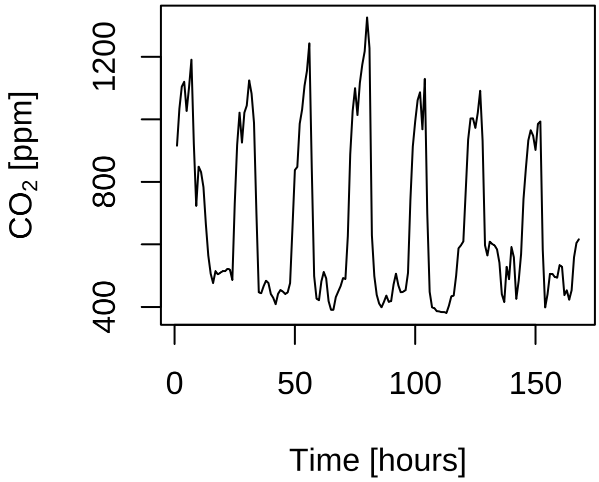
4 ☐ 97 or 98 depending on whether you apply the normal approximation

5 ☐ 162 or 164 depending on whether you apply the normal approximation

Continue on page 13

Exercise VII

CO_2 concentration is an important factor for well-being in the indoor environment, the figure below shows hourly CO_2 concentration [ppm] during a one week period in one room of a dwelling. The variance of the natural logarithm of the CO_2-concentration is 0.137.



As an initial analysis the CO_2 concentration is modeled as a function of time of day using the model

$$Y_i = \beta_0 + x_{1,i}\beta_1 + x_{2,i}\beta_2 + \epsilon_i,$$

where Y_i is the natural logarithm of CO_2 concentration at time i, $\epsilon_i \sim N(0, \sigma^2)$ and iid., and

$$x_{1,i} = \sin\left(2\pi\frac{h_i}{24}\right)$$
$$x_{2,i} = \cos\left(2\pi\frac{h_i}{24}\right),$$

where h_i is the hour of day for observation i.

The model is fitted and the result is reported below (some numbers are replaced by characters);

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59619 -0.09527  0.03135  0.12898  0.42424
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.43959    0.01468  t1       pv1
x1           0.40303    0.02076  t2       pv2
x2           0.20019    0.02076  t3       pv3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: Sig on 165 degrees of freedom
Multiple R-squared:  R2,  Adjusted R-squared:  0.7369
F-statistic: 234.9 on 2 and 165 DF,  p-value: < 2.2e-16
```

Question VII.1 (17)

What is the total number of observations used for the estimation?

1 ☐ 165

2 ☐ 166

3 ☐ 164

4 ☐ 167

5 ☐ 168

Question VII.2 (18)

What is the order of the p-values (`pv1`, `pv2`, and `pv3`) in the R-summary above?

1 ☐ pv2<pv3<pv1

2 ☐ pv1<pv2<pv3

3 ☐ pv1<pv3<pv2

4 ☐ pv3<pv1<pv2

5 ☐ pv1<pv2=pv3

Continue on page 15

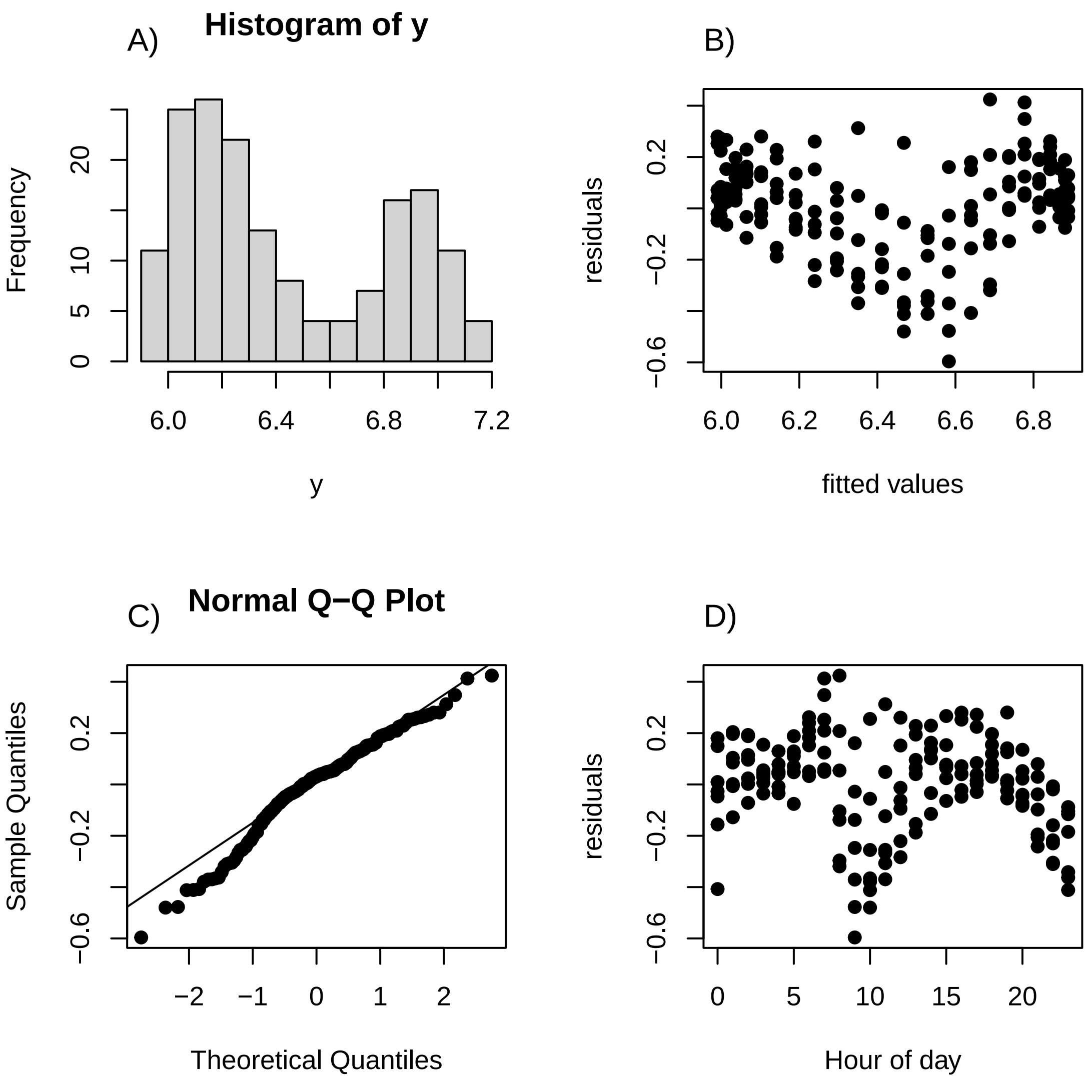
As part of the model validation the figure below is created. The plot show

A) Histogram of y (log-CO_2 concentration)

B) Residuals as a function of the fitted values using the model

C) Normal quantile-quantile plot of the residuals from the model

D) Residuals from the model as a function of hour of day



Continue on page 16

Question VII.3 (19)

Based on the plots in the figure, which of the following statements is correct (both the statement and figure reference should be correct)?

1 ☐ Based on figure A we should consider log-transforming the outcome

2 ☐ The residuals seems to be independent (figure C)

3 ☐ The normality assumption is clearly violated (figure A)

4 ☐ The residuals seems to be normally distributed (figure B)

5 ☐ There are still systematic effects related to time of day (figure D)

Regardless of the conclusions from the previous question, it is decided to continue the investigation based on the developed model. As an aid for the next questions the following relations are given

$$\sum_{h=1}^{24} \sin\left(2\pi \frac{h}{24}\right) = \sum_{h=1}^{24} \cos\left(2\pi \frac{h}{24}\right) = 0$$

$$\sum_{h=1}^{24} \sin\left(2\pi \frac{h}{24}\right) \cos\left(2\pi \frac{h}{24}\right) = 0$$

$$\sum_{h=1}^{24} \sin^2\left(2\pi \frac{h}{24}\right) = \sum_{h=1}^{24} \cos^2\left(2\pi \frac{h}{24}\right) = 12$$

and as stated the observations are made over 7 full days.

Question VII.4 (20)

Referring to the summary table above, what is `Sig`?

1 ☐ 0.27

2 ☐ 0.033

3 ☐ 0.021

4 ☐ 0.19

5 ☐ 0.056

Continue on page 17

Question VII.5 (21)

Now let $\hat{\sigma}$ denote the estimated standard deviation, what is the usual 95% confidence interval of log-CO_2 concentration at noon ($h = 12$)?

1 ☐ $6.04 \pm 0.58\hat{\sigma}$

2 ☐ $6.24 \pm 1.97\hat{\sigma}$

3 ☐ $6.24 \pm 0.26\hat{\sigma}$

4 ☐ $6.04 \pm 0.15\hat{\sigma}$

5 ☐ $6.04 \pm 0.85\hat{\sigma}$

Question VII.6 (22)

If `x1` and `x2` was removed from the model (so a constant mean model), what would the standard error related to the estimate of β_0 then be (hint: the variance of the outcomes is given above)?

1 ☐ 0.0147

2 ☐ 0.00990

3 ☐ 0.00734

4 ☐ 0.0208

5 ☐ 0.0106

Continue on page 18

Exercise VIII

A researcher is interested in investigating the effects of fertilizer and watering frequency on plant growth. 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 plant growth when applying the i'th fertilizer and j'th watering frequency ("Daily", "Twice a week" or "Weekly").

Question VIII.1 (23)

Considering the statistical model above, which of the following statements regarding α_i is correct (note the statements are on the underlying model not on statistical tests)?

1 ☐ α_i denotes the effect size for watering frequency. $\alpha_i \neq 0$ implies that expected plant growth depends on watering frequency.

2 ☐ α_i denotes the effect size for watering frequency. This term should be omitted when plant growth depends on fertilizer type.

3 ☐ α_i denotes the effect size for fertilizer. $\alpha_i \neq 0$ implies that expected plant growth depends on fertilizer type.

4 ☐ α_i denotes the effect size for fertilizer. This term should be omitted when plant growth depends on fertilizer type.

5 ☐ α_i denotes the mean of the i-th fertilizer.

Question VIII.2 (24)

A two-way ANOVA was carried out. The resulting ANOVA table is shown below. Please note that the p-values have been replaced by question marks.

```
## Analysis of Variance Table

## Response: Plant_Growth
##                    Df Sum Sq Mean Sq F value  Pr(>F)
## Fertilizer          1 8.4017  8.4017  78.766       ?
## Watering_Frequency  2 4.0133  2.0067  18.812       ?
## Residuals           2 0.2133  0.1067
```

Calculate the critical F-value for fertilizer and test the hypothesis of equal plant growth among fertilizers ($\alpha = 0.05$). Which of the following statements is the correct one?

1 ☐ $F_{crit} = 38.51$. We reject the null hypothesis of equal plant growth among fertilizers because $F_{obs} > F_{crit}$

2 ☐ $F_{crit} = 19$. We accept the null hypothesis of equal plant growth among fertilizers because $F_{obs} < F_{crit}$

3 ☐ $F_{crit} = 18.51$. We accept the null hypothesis of equal plant growth among fertilizers because $F_{obs} > F_{crit}$

4 ☐ $F_{crit} = 19$. We reject the null hypothesis of equal plant growth among fertilizers because $F_{obs} > F_{crit}$

5 ☐ $F_{crit} = 18.51$. We reject the null hypothesis of equal plant growth among fertilizers because $F_{obs} > F_{crit}$

Question VIII.3 (25)

Which of the following commands can be used to assess if the assumption of normality is fullfilled?

1 ☐

```
lm1 <- lm(Plant_Growth~Fertilizer+Watering_Frequency, data)
qqnorm(lm1$residuals)
qqline(lm1$residuals)
```

2 ☐

```
lm1 <- lm(Plant_Growth~Fertilizer+Watering_Frequency, data)
lm1 <- anova(lm1)
qqnorm(lm1$residuals)
qqline(lm1$residuals)
```

3 ☐

```
qqnorm(data$Plant_Growth)
qqline(data$Plant_Growth)
```

4 ☐

```
qqnorm(data$Plant_Growth[data$Fertilizer=="Type 1"])
qqline(data$Plant_Growth[data$Fertilizer=="Type 1"])
```

5 ☐

```
qqnorm(rnorm(length(data)))
qqline(rnorm(length(data)))
```

Continue on page 20

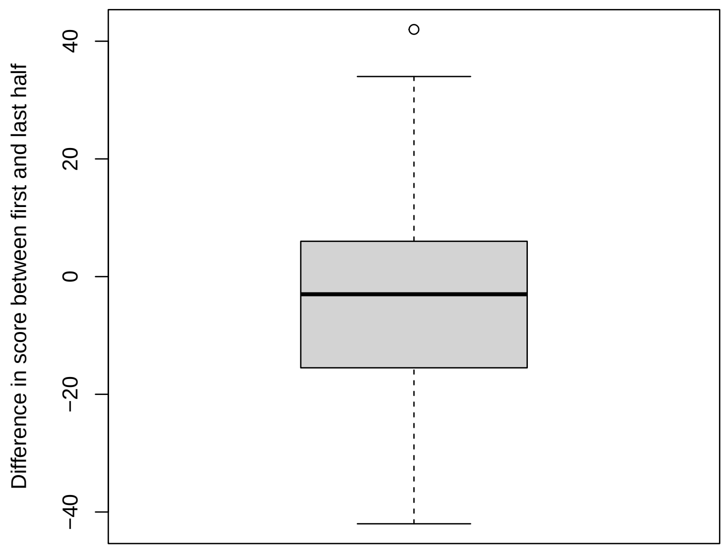
Exercise IX

After a multiple choice exam in the introductory statistics course at DTU the teachers wanted to investigate the scores of different groups.

One question they would like to answer were: Were the students better at answering the first half of the exam (i.e. Question 1 to 15) than the last half (Question 16 to 30).

Let `xfirst` be a vector with students' scores in the first half of the exam and similarly `xlast` the students' scores for the second half of the exam. The observed differences in score between the last and the first half for all passed students is calculated and showed with a boxplot by:

```
x <- xlast - xfirst
boxplot(x, ylab="Difference in score between first and last half")
```



Continue on page 21

Question IX.1 (26)

Which one of the following conclusions is wrong based on the information presented by the box-plot?

1 ☐ More than half of the students in the sample had a negative difference in scores.

2 ☐ More than 20% of the students in the sample had a positive difference in scores.

3 ☐ At least one student in the sample had a difference higher than 40 points in scores.

4 ☐ 60% of the students in the sample had a positive difference in scores.

5 ☐ No student in the sample had a difference in scores higher than 50 points.

Question IX.2 (27)

The teachers want to test the null hypothesis

$$H_0 : \mu = 0,$$

where μ is the mean of the difference in scores between first and last part. They want to test without making any assumption of the distribution of the population where the sample was taken from.

The following code was run:

```
k <- 10000
simsamples <- replicate(k, sample(x, replace = TRUE))

quantile(apply(simsamples, 2, mean), c(0.05, 0.95))

##    5%   95%
## -6.54 -1.21

quantile(apply(simsamples, 2, mean), c(0.025, 0.975))

##  2.5% 97.5%
## -7.05 -0.77

quantile(apply(simsamples, 2, mean), c(0.005, 0.995))

##  0.5% 99.5%
## -8.09  0.23
```

Which one of the following answers is correct?

1 ☐ On a significance level $\alpha = 0.1$ a significant difference in scores between the first and last half is detected.

2 ☐ On a significance level $\alpha = 0.025$ a significant difference in scores between the first and last half is detected.

3 ☐ On a significance level $\alpha = 0.01$ a significant difference in scores between the first and last half is detected.

4 ☐ No conclusion can be made, the calculations don't meet the requirements, since in the calculations a normal distribution is assumed.

5 ☐ None of the answers above are correct.

Question IX.3 (28)

The teachers wanted to investigate if the difference in score between the first and the second part of the exam differs according to the total score for a student. In order to investigate this the students were divided into two groups: one group that had a low total score and another group that had a high total score.

The score differences for low scoring students were stored in `xlow` and for high scoring students in `xhigh`.

The following code was executed:

```
k <- 10000
sim.xlow.samples <- replicate(k, sample(xlow, replace = TRUE))
sim.xhigh.samples <- replicate(k, sample(xhigh, replace = TRUE))

sim.xlow.means <- apply(sim.xlow.samples, 2, mean)
sim.xhigh.means <- apply(sim.xhigh.samples, 2, mean)
sim.dif.means <- apply(sim.xhigh.samples, 2, mean) -
    apply(sim.xlow.samples, 2, mean)

quantile(sim.xlow.means, c(0.025, 0.975))

##  2.5% 97.5%
## -9.23 -2.94

quantile(sim.xhigh.means, c(0.025, 0.975))

##  2.5% 97.5%
## -3.11  1.92

quantile(sim.dif.means, c(0.025, 0.975))

##  2.5% 97.5%
##  1.41  9.56
```

Which of the following conclusions is correct about the difference in mean of the two groups at significance level $\alpha = 0.05$ (both conclusion and argument must be correct)?

1 ☐ A significant difference between the two groups is not detected, since their one-sample confidence intervals overlap.

2 ☐ A significant difference between the two groups is detected, since their one-sample confidence intervals overlap.

3 ☐ A significant difference between the two groups is detected, since the one-sample confidence interval of one group includes zero, but that of the other groups' does not.

4 ☐ A significant difference between the two groups is detected, since the confidence interval for the difference in mean doesn't include zero.

5 ☐ None of the above conclusions are correct.

Continue on page 24

Exercise X

In a production of muesli, raisins are added to the other ingredients in a specific amount, that is well known. The mix is put into boxes ready for sale, hence the expected value of raisins in each box is well known (e.g. if 10 kg of raisins is distributed in 1000 boxes then the expected value is 10 g in each box). The production engineer is however concerned about the variation in the amount of raisins between ready-made boxes. Therefore a sample of 10 boxes is taken out for inspection, and the average sum of squared deviation between the known mean and the individual observed amount of raisins is calculated. It is assumed that the amount of raisins in each box is iid. normally distributed.

The average sum of squared deviations is defined by

$$\frac{1}{N_1}\sum_{i=1}^{N_1}(y_i-\mu)^2,$$

where, in this case, $N_1 = 10$, y_i is the amount of raisins in each package, and μ is the known mean.

Question X.1 (29)

What is the probability that the observed average sum of squared deviation from the (known) mean is less than half the true variance?

1 ☐ 0.69

2 ☐ 0.11

3 ☐ 0.5

4 ☐ 0.17

5 ☐ 0.31

Continue on page 25

Question X.2 (30)

Consider another experiment now examining 20 packages of muesli. It is assumed that the variance is the same in the two experiments.

If the assumption is correct, what is the probability that the observed sum of squared deviations:

$$\sum_{i=1}^{N}(y_i - \mu)^2$$

in experiment 1 (i.e. $N = N_1 = 10$) is greater than the observed sum of squared deviations in experiment 2 (i.e. $N = N_2 = 20$)?

1 ☐ 0.090

2 ☐ 0.023

3 ☐ 0.5

4 ☐ 0.66

5 ☐ 0.097

The exam is finished. Enjoy the summer!