*Written examination*: 14. May 2023

*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 14 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 | III.1 | IV.1 | V.1 | V.2 | VI.1 | VI.2 | VI.3 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Question** | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| **Answer** | | | | | | | | | | |

| **Exercise** | VII.1 | VII.2 | VIII.1 | VIII.2 | VIII.3 | IX.1 | X.1 | X.2 | X.3 | XI.1 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Question** | (11) | (12) | (13) | (14) | (15) | (16) | (17) | (18) | (19) | (20) |
| **Answer** | | | | | | | | | | |

| **Exercise** | XI.2 | XII.1 | XII.2 | XIII.1 | XIII.2 | XIV.1 | XIV.2 | XIV.3 | XIV.4 | XIV.5 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Question** | (21) | (22) | (23) | (24) | (25) | (26) | (27) | (28) | (29) | (30) |
| **Answer** | | | | | | | | | | |

The exam paper contains 24 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 company produces cylindrical water barrels with a total volume capacity of 100 liters. Given their manufacturing process they have determined uncertainties with respect to barrel radius and height. They are now interested in determining the uncertainty with respect to the total volume capacity.

The formula for a cylindrical volume is $V = H \cdot \pi \cdot R^2$, with $H$ and $R$ being height and radius, respectively. $H$ and $R$ are assumed to be independent random variables. The following information is provided.

$\mu_H = 0.6$ m, $\sigma_H = 5 \cdot 10^{-3}$ m

$\mu_R = 0.23$ m, $\sigma_R = 3 \cdot 10^{-3}$ m

$\mu_V = 0.1$ m$^3$, $\sigma_V = ?$

**Question I.1 (1)**

Which of the following formulas can be used to approximate the uncertainty of the total volume capacity, which is the standard deviation of the total volume ($\sigma_V$)? **Hint**: Use the command `pi` for your R calculations in order to obtain the correct precision of your result!

1 ☐ $\sigma_V = 0.166^2 \cdot \sigma_R^2 + 0.867^2 \cdot \sigma_H^2$

2 ☐ $\sigma_V = 0.166^2 \cdot \sigma_H^2 + 0.867^2 \cdot \sigma_R^2$

3 ☐ $\sigma_V = \sqrt{0.166^2 \cdot \sigma_H^2 + 0.867^2 \cdot \sigma_R^2}$

4 ☐ $\sigma_V = \sqrt{0.166^2 \cdot \sigma_R^2 + 0.867^2 \cdot \sigma_H^2}$

5 ☐ $\sigma_V = \sqrt{0.166^2 \cdot \sigma_R^2 + 0.434^2 \cdot \sigma_H^2}$

**Question I.2 (2)**

Which of the following code simulates the production of 1000 barrels and computes their respective volume capacities? It is assumed that the data are normally distributed.

1 ☐ `rnorm(1000, 0.6, 0.03) * pi * rnorm(1000, 0.23, 0.05)^2`

2 ☐ `pnorm(1000, 0.6, 0.003) * pi * pnorm(1000, 0.23, 0.005)^2`

3 ☐ `rnorm(1000, 0.6, 0.005) * pi * rnorm(1000, 0.23, 0.003)^2`

4 ☐ `qnorm(1000, 0.6, 0.005) * pi * qnorm(1000, 0.23, 0.003)^2`

5 ☐ `qnorm(500, 0.6, 0.003) + pi * qnorm(500, 0.23, 0.005)^2`

Continue on page 4

**Exercise II**

Military radar and missile detection systems are designed to warn a country of an enemy attack. A reliability question is whether a detection system will be able to identify an attack and issue a warning. Assume that a particular detection system has a 0.90 probability of detecting a missile attack.

**Question II.1 (3)**

If two detection systems are installed in the same area and operate independently, what is the probability that both of the systems will detect the attack?

1 ☐ 0.9

2 ☐ 0.81

3 ☐ 0.92

4 ☐ 0.99

5 ☐ 0.999

Continue on page 5

**Exercise III**

The exam score of students are assumed to be normally distributed. In a 100-marks final exam, the mean score is 55 and standard deviation is 19. The top-scoring 15% students will be given the grade 12 according to the Danish 7-point scale.

**Question III.1 (4)**

What is the minimum score required to achieve grade 12 in the exam?

1 ☐ 70.99

2 ☐ 74.69

3 ☐ 76.76

4 ☐ 78.80

5 ☐ 82.80

Continue on page 6

**Exercise IV**

A survey showed that a majority of the respondents plan on doing their holiday shopping online because they don't want to spend money on gas driving from store to store. Suppose we have a group of 10 shoppers; 7 prefer to do their holiday shopping online and 3 prefer to do their holiday shopping in stores. A random sample of 3 of these 10 shoppers is selected for a more in-depth study of how the economy has impacted their shopping behavior.

**Question IV.1 (5)**

What is the probability that the majority (either 2 or 3) prefer shopping online?

1 ☐ 0.0006

2 ☐ 0.0136

3 ☐ 0.8167

4 ☐ 0.9012

5 ☐ 0.9984

Continue on page 7

**Exercise V**

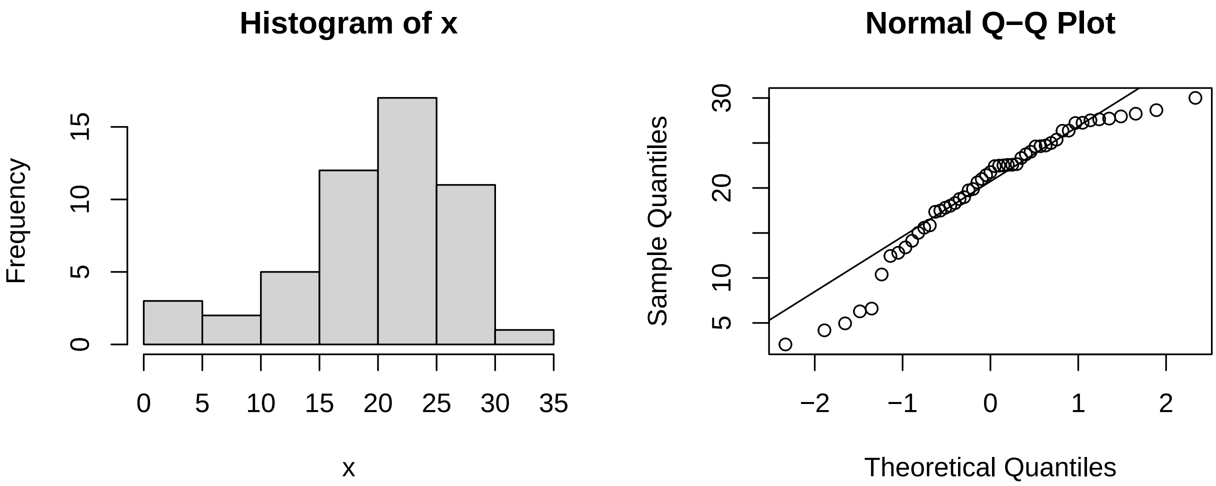
Power generation with solar cells is essential for the transition to a sustainable energy system. In a test, two similar sized systems were placed next to each other such that they were exposed to exactly the same conditions. The power generation from each system was measured over a period of 50 days. Let $x_i$ and $y_i$ denote the observed energy generation during Day $i$ in kWh in the period for System 1 and 2, respectively. Hence, $n = n_x = n_y = 50$.

Let $\delta_i = x_i - y_i$ denote the daily difference, and $\bar{\delta} = \frac{1}{50}\sum_{i=1}^{50}\delta_i$ and $s_{\delta}^2 = \frac{1}{49}\sum_{i=1}^{50}(\delta_i - \bar{\delta})^2$.

**Question V.1 (6)**

When calculating a confidence interval for the mean value using the $t$-distribution the assumption of normal distribution of the population can be considered.

For the observed values from System 1 the assumption is checked with the following histogram and the normal QQ-plot:




What is the appropriate conclusion with regards to the validity of the confidence interval based on the two plots and consideration of the sample size?

1 ☐ No evident deviation from the normal distribution is found, hence the normal distribution assumption is fulfilled and thus the confidence interval is valid. The sample size doesn't influence this conclusion.

2 ☐ A deviation from the normal distribution is found, hence the normal distribution assumption is not fulfilled and thus the confidence interval is not valid. The sample size doesn't influence this conclusion.

3 ☐ A deviation from the normal distribution is observed, hence the normal distribution assumption is not fulfilled, and since the sample size is above 30 and according to the

central limit theorem the normal assumption is an issue and thus the confidence interval is not valid.

4 ☐ A deviation from the normal distribution is observed, hence the normal distribution assumption is not fulfilled, however since the sample size is above 30 according to the central limit theorem the normal assumption is not an issue and thus the confidence interval is valid.

5 ☐ None of the above are appropriate conclusions.

## Question V.2 (7)

Which test statistic is correct to use for testing the hypothesis

$$H_0 : \mu_X - \mu_Y = 0$$

i.e. difference in mean energy generated by the two systems?

1 ☐ $t_{\text{obs}} = \frac{\bar{x}-\bar{y}}{\sqrt{s_x/n_x+s_y/n_y}}$

2 ☐ $t_{\text{obs}} = \frac{\bar{x}-\bar{y}}{\sqrt{s_x^2/n_x+s_y^2/n_y}}$

3 ☐ $t_{\text{obs}} = \frac{\bar{\delta}}{\sqrt{s_x^2/n_x+s_y^2/n_y}}$

4 ☐ $t_{\text{obs}} = \frac{\bar{x}-\bar{y}}{\sqrt{s_\delta/n}}$

5 ☐ $t_{\text{obs}} = \frac{\bar{\delta}}{\sqrt{s_\delta^2/n}}$

Continue on page 9

**Exercise VI**

An office worker was given the advice to exercise more and at least take 10000 steps per day. Using a step counter the office worker counted the number of steps per day. After 10 days he entered the number from each day rounded to nearest 100 into R by:

```
x <- c(8500, 10300, 6800, 10600, 4900, 6200, 10800, 5700, 5100, 9000)
```

The office worker now wanted to use the data to conclude if he did enough exercise according to the advice.

### Question VI.1 (8)

What is the 99% confidence interval for the mean value of the number of steps walked per day calculated with the collected sample?

1 ☐ $[1297, 14283]$

2 ☐ $[4274, 11306]$

3 ☐ $[5400, 10180]$

4 ☐ $[6126, 9454]$

5 ☐ $[6442, 9138]$

### Question VI.2 (9)

What is conclusion of the test of the hypothesis

$$H_0 : \mu = 10000$$

with the collected data on a 5% significance level (both the conclusion and argument must be correct)?

1 ☐ The null hypothesis is rejected, since the $p$-value is <u>below</u> 5%.

2 ☐ The null hypothesis is rejected, since the $p$-value is <u>above</u> 5%.

3 ☐ The null hypothesis is <u>not</u> rejected, since the $p$-value is <u>below</u> 5%.

4 ☐ The null hypothesis is <u>not</u> rejected, since the $p$-value is <u>above</u> 5%.

5 ☐ None of the above conclusions and arguments are correct.

**Question VI.3 (10)**

In an experiment, the office worker wants to compare two periods where steps per day are observed. Each period has the length of 14 days and he wants to test at a significance level 1% and have the power of the test at 80%. He wants to be able to detect the population mean difference of 1000 steps. What is the standard deviation of the population in this case (rounded to down to nearest whole number)?

1 ☐ $\sigma = 235$

2 ☐ $\sigma = 470$

3 ☐ $\sigma = 724$

4 ☐ $\sigma = 909$

5 ☐ $\sigma = 1061$

Continue on page 11

## Exercise VII

We observed three variables and carried out a multiple linear regression

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

```
summary(lm(y~x1+x2))

##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.880 -2.584  1.109  1.828  6.598
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9653     1.3833  -0.698  0.49673
## x1            1.2933     0.3481   3.715  0.00231 **
## x2           11.8911     0.6950  17.109 8.82e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.76 on 14 degrees of freedom
## Multiple R-squared:  0.956,Adjusted R-squared:  0.9497
## F-statistic: 152.1 on 2 and 14 DF,  p-value: 3.196e-10
```

### Question VII.1 (11)

Which of the following statements is correct regarding the output line "`x1`" in the lm result?

1 ☐ The Std. Error expresses the uncertainty on the estimate of the regression slope $\beta_2$.

2 ☐ The Std. Error expresses the uncertainty on the expected value of an observation, where $x_1 = 1$ and $x_2 = 0$.

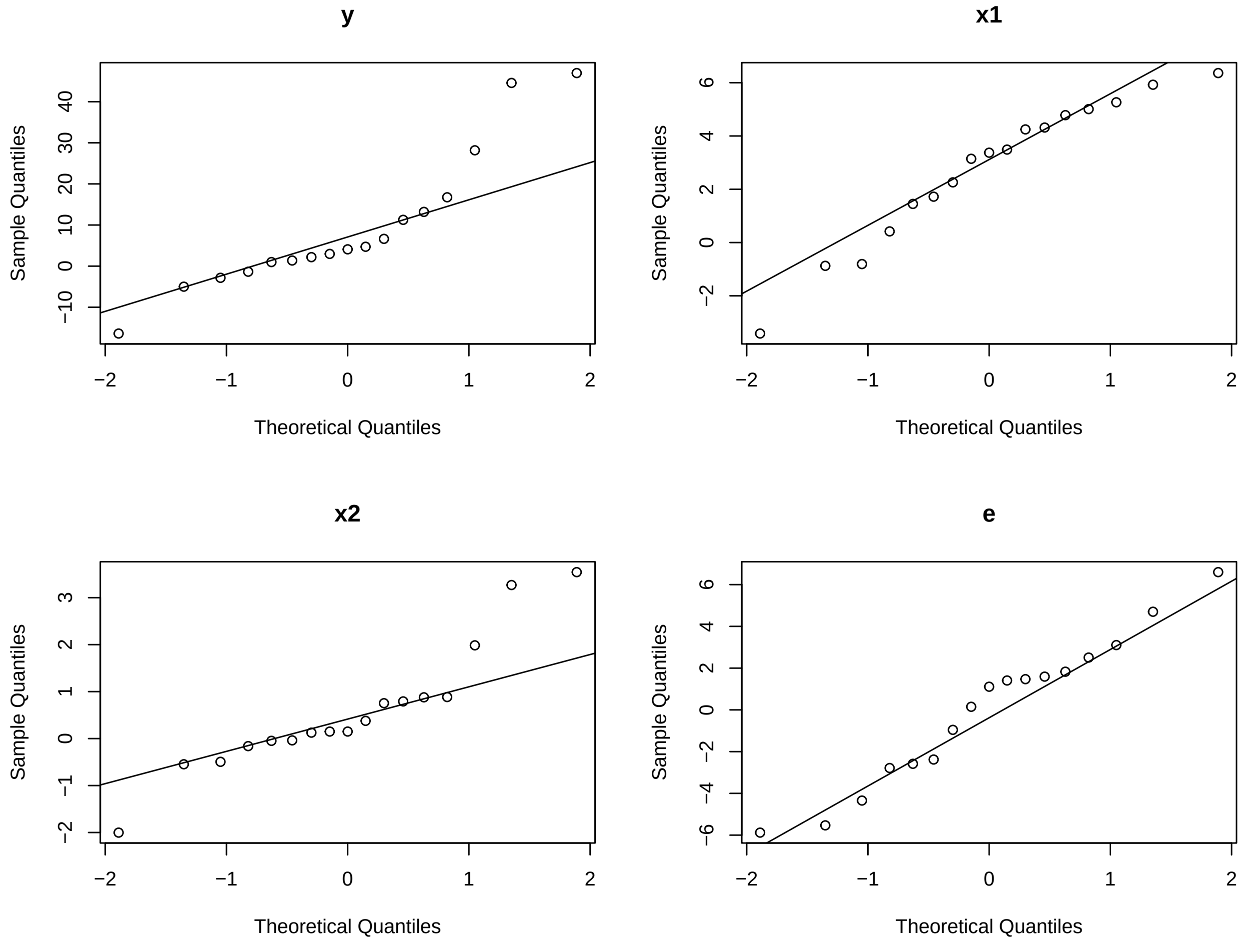
3 ☐ The $t$-value is a measure of model validity. A small $t$-value indicates a valid model.

4 ☐ The $t$-value can be used to assess if there is a significant association between $x_1$ and $y$.

5 ☐ Neither the Std. Error nor the $t$-value are related to the uncertainty of the model.

### Question VII.2 (12)

To assess the normality assumption of the model, four QQ-plots using $y_i$, $x_{1,i}$, $x_{2,i}$ and the residuals $e_i$ were generated:




Which of these plots should be used to assess the assumption?

1 ☐ The QQ plot using $y_i$

2 ☐ The QQ plot using $x_{1,i}$

3 ☐ The QQ plot using $x_{2,i}$

4 ☐ The QQ plot using $e_i$

5 ☐ None of the above.

Continue on page 13

**Exercise VIII**

To monitor the long-term effects of environmental policies at a place of interest, the environmental agency measured the amount of dissolved oxygen in water (DO) in mg/L, which is an indicator of water quality.

The data was read into R:

```
year <- 1990:2015
DO <- c(1.52, 2.88, 1.60, 2.24, 2.45, 1.84, 2.03, 2.33, 2.81,
        2.46, 2.36, 2.23, 2.81, 2.70, 2.63, 2.00, 2.40, 2.45,
        2.48, 2.51, 2.55, 2.77, 2.70, 2.23, 2.88, 3.09)
```

and a linear regression was carried out. Note that some output has been masked with an `X`:

```
summary(lm(DO ~ year))

## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -52.916308  17.494219  -3.025  X         X
## year          0.027634   0.008736   3.163  X         X
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3341 on 24 degrees of freedom
```

### Question VIII.1 (13)

What is the estimated total increase in DO during a period of five years?

1 ☐ -0.028 mg/L

2 ☐ 0.028 mg/L

3 ☐ 0.14 mg/L

4 ☐ 2.8%

5 ☐ 17.5 mg/L

### Question VIII.2 (14)

The environmental agency wants to test the null hypothesis $H_0 : \beta_1 = 0$ using a significance level of 1%, where $\beta_1$ is the parameter representing the slope in the linear regression. Which of the following conclusions is correct (both argument and conclusion must be correct)?

1 ☐ The critical values for the test are $\pm 2.06$. Since $t_{\text{obs}} > 2.06$, the hypothesis is rejected.

2 ☐ The critical values for the test are $\pm 2.80$. Since $t_{\text{obs}} > 2.80$, the hypothesis is rejected.

3 ☐ The critical values for the test are $\pm 2.06$. Since $t_{\text{obs}} > 2.06$, the hypothesis is accepted

4 ☐ The critical values for the test are $\pm 2.80$. Since $\hat{\beta}_1$ is within this interval, the hypothesis is accepted.

5 ☐ The $p$-value for the test is $p = 0.004$. Since $p < 0.01$, the hypothesis is accepted.

## Question VIII.3 (15)

Suppose the DO value in 2022 was 2.21 mg/L. Which of the following statements is correct (both argument and conclusion must be correct)?

1 ☐ The 95% prediction interval is [2.17, 3.75]. The observation fits reasonably well with the model.

2 ☐ The 95% confidence interval is [2.38, 3.34]. The observation fits reasonably well with the model.

3 ☐ The 95% confidence interval is [2.38, 3.34]. The observation does <u>not</u> fit well with the model.

4 ☐ The 95% prediction interval is [2.17, 3.75]. The observation does <u>not</u> fit well with the model.

5 ☐ None of the above statements are correct.

Continue on page 15

**Exercise IX**

An interruption of Internet service occurred for the customers living in a city. When customers called the Internet service provider's office, a recorded message told them that the company was aware of the service outage and that it was anticipated that service would be restored in two hours. Assume that two hours is the mean time to do the repair and that the repair time has an exponential probability distribution.

**Question IX.1 (16)**

What is the probability that the repair will take between one hour and two hours?

1 ☐ 0.3935

2 ☐ 0.5521

3 ☐ 0.0821

4 ☐ 0.3934

5 ☐ 0.2387

Continue on page 16

**Exercise X**

This exercise contains three questions related to simulation and bootstrapping.

## Question X.1 (17)

Heights of 15 students were measured in cm and read into R.

Perform parametric bootstrapping applying 1000 simulations while assuming a normal distribution of the student heights.

Use the following R code. Copy it to R and fill in the missing code at the `?` symbols:

```
heights <- c(162, 172, 178, 154, 173, 174, 166, 166,
             166, 164, 167, 163, 165, 170, 177)
set.seed(1234)
k <- 1000
sim_samples <- replicate(k, ?)
sim_stats <- apply(sim_samples, 2, ?)
quantile(sim_stats, c(?, ?))
```

Remember to run the `set.seed(1234)` when you calculate the result.

Which of the following is the obtained 95% confidence interval for the median student height?

1 ☐ $[164.29, 171.64]$

2 ☐ $[164.01, 171.61]$

3 ☐ $[164.69, 171.15]$

4 ☐ $[164.73, 171.14]$

5 ☐ $[163.66, 170.81]$

## Question X.2 (18)

We wish to generate 10 random numbers coming from a log normal distribution with $\alpha = 0$ and $\beta = 1$. Which of the following commands is correct?

1 ☐ `runif(10)`

2 ☐
```
pnorm(runif(10))
```

3 ☐
```
plnorm(runif(10))
```

4 ☐
```
qnorm(runif(10))
```

5 ☐
```
qlnorm(runif(10))
```

## Question X.3 (19)

Which of the following is part of the code required for performing non-parametric bootstrapping based on a sample $x$?

1 ☐
```
sim_means = apply(x, 2, mean)
quantile(sim_means, c(0.025, 0.975))
```

2 ☐
```
sample(rnorm(length(x), mean = mean(x), sd = sd(x)),
       size = 1000, replace = FALSE)
```

3 ☐
```
replicate(10000, sample(x, replace = FALSE))
```

4 ☐
```
replicate(10000, sample(x, replace = TRUE))
```

5 ☐
```
sample(rnorm(length(x), mean = mean(x), sd = sd(x)),
       size = 1000, replace = TRUE)
```

Continue on page 18

**Exercise XI**

Aircraft paint can be applied to metal surfaces by two methods, namely dipping and spraying. A process engineering group is interested in learning whether four paints differ in their adhesion properties when being applied by the two methods.

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 adhesion force of the $i$'th paint when being applied using the $j$'th method. $\alpha_i$ and $\beta_j$ represent effect sizes corresponding to applied paint and method, respectively.

The result of fitting the model is given in the ANOVA table below:

```
## Analysis of Variance Table
##
## Response: adhesion
##           Df  Sum Sq Mean Sq F value Pr(>F)
## paint      3 0.93308 0.31103  1.0876 0.4733
## method     ? 1.26474       ?       ?      ?
## Residuals  3 0.85790 0.28597
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Question XI.1 (20)**

How many observations were included in the study?

1 ☐ We do not have sufficient information to determine this.

2 ☐ 4

3 ☐ 6

4 ☐ 8

5 ☐ 16

**Question XI.2 (21)**

Another engineering team conducted a similar study as described above and obtained the following results using 4 different paints:

```
## Analysis of Variance Table
##
## Response: adhesion
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## paint      3 2.53730 0.84577 43.5483 0.005671 **
## method     1 0.00099 0.00099  0.0511 0.835659
## Residuals  3 0.05826 0.01942
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The engineers would like to perform pairwise post-hoc comparisons of all paints and wish to obtain a 5% family wise risk of making a type I error. Which of the following Bonferroni corrections should be applied?

1 ☐ $\alpha_{\text{Bonferroni}} = 0.025$

2 ☐ $\alpha_{\text{Bonferroni}} = \frac{0.05}{4}$

3 ☐ $\alpha_{\text{Bonferroni}} = \frac{0.05}{3}$

4 ☐ $\alpha_{\text{Bonferroni}} = \frac{0.05}{6}$

5 ☐ $\alpha_{\text{Bonferroni}} = 0.05$

Continue on page 20

Exercise XII

The following data represents weights of food (in kilograms) consumed per day by adult deer collected at different months of the year:

```
consumption <- c(5.4, 4.5, 4.6, 4.4, 4.9, 3.3, 4.1, 4.6, 4.8, 4.6,
                 5.2, 4.7, 4.4, 4.8, 4.8, 5.2, 4.9, 4.8, 5.6, 5.5)
month <- c(rep("Feb",5), rep("May",5), rep("Aug", 5), rep("Nov", 5))
```

## Question XII.1 (22)

What is the total sum of squares (SST) of the consumption?

1 ☐ 0.1843

2 ☐ 0.2535

3 ☐ 2.1215

4 ☐ 2.9480

5 ☐ 5.0695

## Question XII.2 (23)

Read in the data in R and test the null hypothesis of equal food consumption across all months! Apply a significance level $\alpha = 0.05$. Which statement is correct?

1 ☐ We analyse the data using multiple linear regression. We accept the null hypothesis of equal food consumption because $p = 0.3 > \alpha$

2 ☐ We analyse the data using one-way Anova. We accept the null hypothesis of equal food consumption because $p = 0.3 > \alpha$

3 ☐ We analyse the data using multiple linear regression. We reject the null hypothesis of equal food consumption because $p = 0.3 > \alpha$

4 ☐ We analyse the data using one-way Anova. We accept the null hypothesis of equal food consumption because $p = 0.03 < \alpha$

5 ☐ We analyse the data using one-way Anova. We reject the null hypothesis of equal food consumption because $p = 0.03 < \alpha$

Continue on page 21

**Exercise XIII**

Phone calls at Regional Airways arrive at rate of 48 calls per hour at the reservation desk for Regional 8Airways following a Poisson distribution.

**Question XIII.1 (24)**

Let the random variable $X$ denote the number of calls. What is the variance of $X$?

1 ☐ 2304

2 ☐ 48

3 ☐ 9.6

4 ☐ 4

5 ☐ 2

**Question XIII.2 (25)**

Suppose no calls are currently on hold. If the agent takes 5 minutes to complete the current call, what is the probability that at two or above callers will be waiting?

1 ☐ 0.9817

2 ☐ 0.9084

3 ☐ 0.7619

4 ☐ 0.5665

5 ☐ 0.7760

Continue on page 22

**Exercise XIV**

Each year Statistics Denmark carries out surveys in the Danish population. One of them is the Survey on Living Conditions (SILC), which include questions about financial vulnerability. A question posed in the questionnaire is: "How easy is it for your household to make ends meet financially?"

The following answers were observed for unemployed persons the last five years:

| | 2017 | 2018 | 2019 | 2020 | 2021 | Sum |
|---|---|---|---|---|---|---|
| Very easy | 9 | 8 | 10 | 11 | 10 | 48 |
| Easy | 15 | 14 | 20 | 18 | 15 | 82 |
| Neither easy nor hard | 26 | 24 | 26 | 27 | 27 | 130 |
| Difficult | 23 | 25 | 19 | 20 | 24 | 111 |
| Very hard | 28 | 30 | 26 | 24 | 24 | 132 |
| Sum | 101 | 101 | 101 | 100 | 100 | 503 |

The following answers were observed for employed persons the last five years:

| | 2017 | 2018 | 2019 | 2020 | 2021 | Sum |
|---|---|---|---|---|---|---|
| Very easy | 19 | 20 | 23 | 20 | 24 | 106 |
| Easy | 31 | 29 | 30 | 31 | 32 | 153 |
| Neither easy nor hard | 30 | 31 | 29 | 30 | 28 | 148 |
| Difficult | 15 | 13 | 12 | 15 | 12 | 67 |
| Very hard | 6 | 7 | 6 | 5 | 5 | 29 |
| Sum | 101 | 100 | 100 | 101 | 101 | 503 |

**Question XIV.1 (26)**

Considering only the answers from unemployed persons in 2021 it was observed that 24 out of 100 answered "Very hard". Which of the following answers contains the best calculation of the 95% confidence interval for this proportion?

1 ☐ $0.2708333 \pm 2.58\sqrt{\frac{0.1975}{96}}$

2 ☐ $0.24 \pm 1.96\sqrt{\frac{0.1824}{100}}$

3 ☐ $0.2708333 \pm 1.96\sqrt{\frac{0.1975}{96}}$

4 ☐ $0.24 \pm 2.58\sqrt{\frac{0.1824}{100}}$

5 ☐ $0.24 \pm 1.64\sqrt{\frac{0.1824}{100}}$

## Question XIV.2 (27)

Considering only the answers from employed persons in 2021 it was observed that 5 out of 101 answered “Very hard”. Which of the following answers contains the best calculation of the 95% confidence interval for this proportion?

1 ☐ $0.0666667 \pm 1.96\sqrt{\frac{0.0622222}{105}}$

2 ☐ $0.049505 \pm 1.96\sqrt{\frac{0.0470542}{101}}$

3 ☐ $0.0666667 \pm 2.58\sqrt{\frac{0.0622222}{105}}$

4 ☐ $0.049505 \pm 2.58\sqrt{\frac{0.0470542}{101}}$

5 ☐ $0.049505 \pm 1.64\sqrt{\frac{0.0470542}{101}}$

## Question XIV.3 (28)

It is of interest to conclude if the distribution of the answers from the employed and unemployed groups is different. Use the answers from 2021 of the two groups to test if there is a significant difference. What is the conclusion on a 5% significance level (both conclusion and argument must be correct)?

1 ☐ There is a significant difference, since the $p$-value is 0.0345.

2 ☐ There is no significant difference, since the $p$-value is 0.9655.

3 ☐ There is no significant difference, since the $p$-value is 0.9997.

4 ☐ There is no significant difference, since the $p$-value is 0.01215.

5 ☐ There is a significant difference, since the $p$-value is 0.00001047.

## Question XIV.4 (29)

It is of interest to examine if the distribution of answers has changed over time. Considering the answers from unemployed persons. Under the usual null hypothesis what is the expected number of answers in the category “Easy” in 2019?

1 ☐ $e_{23} = 3.2604$

2 ☐ $e_{23} = 4.0159$

3 ☐ $e_{23} = 13.3677$

4 ☐ $e_{23} = 16.465$

5 ☐ $e_{23} = 20$

## Question XIV.5 (30)

It is still of interest to examine if the distribution of answers has changed over time. Again, considering the answers from the unemployed persons using the usual test. Which of the following is the correct `R` call for calculation of the $p$-value and conclusion on a 5% significance level?

1 ☐ `2*(1 - pt(4.5303, df=16)) = 0.0003414`, hence a significant change over time.

2 ☐ `pnorm(4.5303) = 0.9999971`, hence no significant change over time.

3 ☐ `1 - pchisq(4.5303, df=16) = 0.9976`, hence no significant change over time.

4 ☐ `pf(4.5303, df1=4, df2=4) = 0.9137`, hence no significant change over time.

5 ☐ `1 - pf(4.5303, df1=4, df2=4) = 0.08627`, hence no significant change over time.

The exam is finished. Enjoy the summer!