*Written examination*: 24 May 2024

*Course name and number*: **02402 Statistics**

*Duration:* 4 hours

*Aids and facilities allowed:* All aids - No internet access

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

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

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

The exam paper contains 26 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**

27% of a certain (infinite) population owns a car. A group of 15 individuals is randomly selected.

**Question I.1 (1)**

What is the probability that exactly 4 out of the 15 individuals own a car?

1 ☐ 0.7720

2 ☐ 0.6192

3 ☐ 0.4965

4 ☐ 0.3871

5 ☐ 0.2276

**Exercise II**

In a manufacturing process, the time between failures of a certain machine follows an exponential distribution, and the mean time between failures (MTBF) is 200 hours.

**Question II.1 (2)**

What is the probability that the machine will fail within 100 hours of the last failure?

1 ☐ 0.6065

2 ☐ 0.1353

3 ☐ 0.8647

4 ☐ 0.6509

5 ☐ 0.3935

**Exercise III**

You are tracking the number of emails received per hour, and on average, you receive 5 emails per hour.

**Question III.1 (3)**

Assuming that the emails arrive according to a Poisson process, what is the probability of receiving exactly 3 emails in the next hour?

1 ☐ 0.0404

2 ☐ 0.1008

3 ☐ 0.1404

4 ☐ 0.1668

5 ☐ 0.1718

**Exercise IV**

Consider a fair standard deck of 52 playing cards with 4 aces. You randomly draw 5 cards from the deck without replacement.

**Question IV.1 (4)**

What is the probability of getting exactly 2 aces?

1 ☐ 0.0399

2 ☐ 0.0132

3 ☐ 0.0347

4 ☐ 0.1102

5 ☐ 0.2113

Exercise V

Consider a discrete random variable $X$ with the distribution specified in the below table:

| $x$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $P(X = x)$ | 0.2 | 0.3 | 0.4 | 0.1 |

**Question V.1 (5)**

What is the variance of $X$?

1 ☐ 4.80

2 ☐ 4.42

3 ☐ 3.36

4 ☐ 3.08

5 ☐ 2.78

Continue on page 5

**Exercise VI**

The building industry has to reduce its $CO_2$ emissions from materials production and construction. A sample of $CO_2$ emissions from a particular part of the construction process was taken from 20 different construction sites. The observations can be assumed independent, and the unit is kg/m$^2$, where the square meter refers to the floor area of the constructed building.

The sample can be read into R by:

```
x <- c(875, 776, 915, 806, 1030, 1197, 768, 456,
       1171, 873, 777,1108, 1031, 1009, 772)
```

## Question VI.1 (6)

What is the value of the quantile in the $t$-distribution used for calculating the 90% confidence interval for the mean using this sample?

1 ☐ 1.341

2 ☐ 1.761

3 ☐ 2.131

4 ☐ 2.145

5 ☐ 2.977

## Question VI.2 (7)

The industry wants to know if the mean emission is different from 1000 kg/m$^2$, and therefore it tests the null hypothesis $H_0 : \mu = 1000$ against the alternative hypothesis $H_1 : \mu \neq 1000$.

What is the $p$-value for the usual test of the null hypothesis?

1 ☐ $4.053 \cdot 10^{-11}$

2 ☐ 0.05

3 ☐ 0.07556

4 ☐ 0.1

5 ☐ 0.92444

**Question VI.3 (8)**

The hypothesis test in the previous question is based on two assumptions. Which of the following statements regarding the assumptions is correct?

1 ☐ The assumption of independence is satisfied because the sample size is small, and the assumption of normality is satisfied according to the Central Limit Theorem

2 ☐ The assumption of independence must be verified by checking that the observations were taken randomly from the population, and the assumption of normality is satisfied according to the Central Limit Theorem

3 ☐ The assumption of independence must be verified by checking that the observations were taken randomly from the population, and the assumption of normality can be assessed with an analysis of variance

4 ☐ The assumption of independence must be verified by checking that the observations were taken randomly from the population, and the assumption of normality can be assessed with a QQ-plot

5 ☐ The assumption of independence is satisfied because the sample size is small, and the assumption of normality can be assessed with a QQ-plot

Continue on page 7

**Exercise VII**

A researcher obtains a data set with 4000 observations divided into four groups each containing 1000 observations. The researcher formulates a one-way ANOVA model for the data:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad i \in \{1, ..., 4\}, \quad j \in \{1, ..., 1000\},$$

where the errors are assumed to be independent. The researcher then considers the usual null hypothesis:

$$\mathcal{H}_0 : \quad \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0,$$

and produces the below ANOVA table associated with the model, where some numbers have been replaced by letters:

| Source | DF | SS | MS | Test statistic | $p$-value |
|---|---|---|---|---|---|
| Group | A | B | C | 1491.24 | $< 0.001$ |
| Residual | D | 16084.03 | E | | |
| Total | 3999 | 34090.85 | | | |

## Question VII.1 (9)

Which of the following options correctly estimates the standard deviation of the error?

1 ☐ $\hat{\sigma} = \sqrt{B}$

2 ☐ $\hat{\sigma} = B$

3 ☐ $\hat{\sigma} = \sqrt{C/A}$

4 ☐ $\hat{\sigma} = \sqrt{E}$

5 ☐ $\hat{\sigma} = \sqrt{E/D}$

**Question VII.2 (10)**

What can be seen from the ANOVA table?

1 ☐ The model assumptions are violated

2 ☐ The average within group variability, MSE, is larger than the average between group variability, MS(group)

3 ☐ The average between group variability, MS(group), is larger than the average within group variability, MSE

4 ☐ None of the group means are significantly different from zero on a 5% significance level

5 ☐ All the group means are significantly different from zero on a 5% significance level

Assume that the researcher performs model validation and finds no indications that the model assumptions are violated.

**Question VII.3 (11)**

Which of the following plots could represent the data (assuming the data has not been transformed during the analysis)?

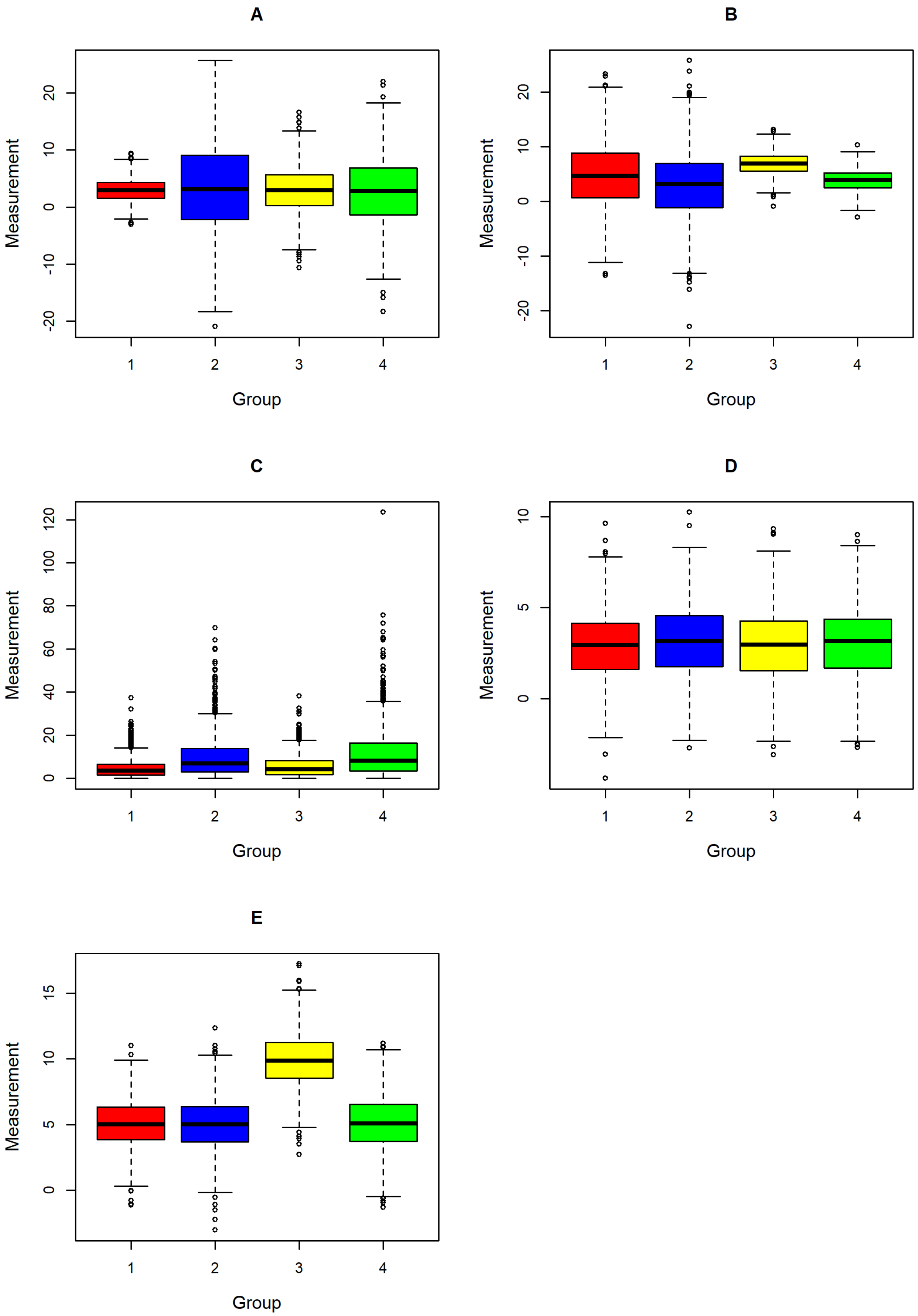
1 ☐ A

2 ☐ B

3 ☐ C

4 ☐ D

5 ☐ E


A
B
C
D
E
Measurement
Group
1
2
3
4
-20
-10
0
10
20
40
60
80
100
120
5
15

**Exercise VIII**

A food company considers replacing its original cake with a new and more sustainable one. Therefore, the company asked 100 consumers to taste both cakes and indicate their preference in a completely randomized trial. The results of the trial are given below:

| | |
|---|---|
| Consumers preferring the original cake: | 41 |
| Consumers preferring the new cake: | 59 |

Let $p$ denote the (population) proportion of consumers that prefers the original cake.

**Question VIII.1 (12)**

Which procedure can be used to determine if there is statistically significant evidence that consumers prefer one cake over the other and what is the conclusion using $\alpha = 0.05$?

1 ☐ Since more consumers preferred the new cake in the trial, there is statistically significant evidence that consumers prefer the new cake

2 ☐ Test the null hypothesis: $H_0 : p = 0.41$ resulting in a $p$-value of 0.025% leading to a conclusion that there is statistically significant evidence that consumers prefer one cake over the other

3 ☐ Test the null hypothesis: $H_0 : p = 0.59$ resulting in a $p$-value of 0.025% leading to a conclusion that there is statistically significant evidence that consumers prefer one cake over the other

4 ☐ Test the null hypothesis: $H_0 : p \neq 0.5$ resulting in a $p$-value of 92.814% leading to a conclusion that there is no statistically significant evidence that consumers prefer one cake over the other

5 ☐ Test the null hypothesis: $H_0 : p = 0.5$ resulting in a $p$-value of 7.186% leading to a conclusion that there is no statistically significant evidence that consumers prefer one cake over the other

**Question VIII.2 (13)**

The food company conducted similar trials in two different countries. In both countries, the trials included 150 consumers. Let $\hat{p}_{\mathrm{A}} = 0.52$ and $\hat{p}_{\mathrm{B}} = 0.64$ denote the proportions of consumers that preferred the original cake in country A and country B, respectively, and assume that the trials in the two countries were independent.

Which expression correctly calculates the test statistic for the usual hypothesis test for the difference of proportions between the two countries?

1 ☐ $\frac{0.52-0.64}{\sqrt{\frac{2\cdot 0.58\cdot 0.42}{150}}}$

2 ☐ $\frac{0.52-0.64}{\sqrt{\frac{2\cdot 0.64\cdot 0.36}{150}}}$

3 ☐ $\frac{0.52-0.64}{\sqrt{\frac{2\cdot 0.64\cdot 0.52}{150}}}$

4 ☐ $\frac{0.58}{\sqrt{150\cdot 0.58\cdot 0.42}}$

5 ☐ $\frac{0.58}{\sqrt{150\cdot 0.52\cdot 0.64}}$

Continue on page 12

**Exercise IX**

Most statistical simulation techniques fall into two categories: parametric and non-parametric bootstrapping.

**Question IX.1 (14)**

Which statement correctly distinguishes between parametric and non-parametric bootstrapping?

1 ☐ Parametric bootstrapping relies on distributional assumptions, while non-parametric bootstrapping makes no assumptions about the underlying distribution

2 ☐ Parametric bootstrapping is computationally faster than non-parametric bootstrapping

3 ☐ Non-parametric bootstrapping requires a larger sample size compared to parametric bootstrapping

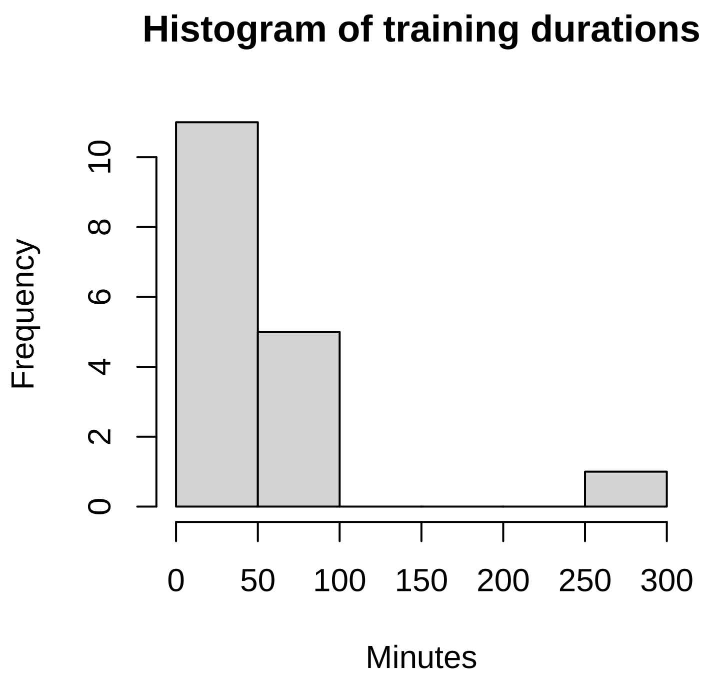
4 ☐ Parametric bootstrapping is only applicable for large sample sizes, while non-parametric bootstrapping is also suitable for small sample sizes

5 ☐ Parametric bootstrapping is used for categorical data, while non-parametric bootstrapping is suitable for continuous data

**Exercise X**

A research team has developed a new training system for trainees. To assess the system, the team conducts an experiment recording training durations (in minutes) of trainees using the training system. The team assumes that the training duration follows an exponential distribution and lets $Y \sim \exp(\lambda)$ denote a random training duration.

A sample of training durations has been collected for $n = 17$ trainees. A histogram of the sample is shown below:




The observed training durations have been saved in the vector `y`, and the below R code is run:

```
set.seed(54873)
# Number of simulations
k <- 10000

# Number of observations in the sample
ny <- length(y)

# Simulate k bootstrap samples each with ny observations
simysamples <- replicate(k, rexp(ny, rate = 1/mean(y)))

# Calculate the mean in each of the k bootstrap samples
simmu <- apply(simysamples, 2, mean)

# Find the two relevant quantiles of the k simulated means
quantile(simmu, c(0.05, 0.95),type=2)

##     5%    95%
## 32.912 74.364
```

**Question X.1 (15)**

The team wants to test the null hypothesis $\mathcal{H}_0 : \mu_Y = 50$ against a two-sided alternative hypothesis.

What is the result of the hypothesis test based on the collected sample and the calculations above (both conclusion and argument must be correct)?

1 ☐ Since 50 is contained in the 90% parametric bootstrap confidence interval, the null hypothesis cannot be rejected at $\alpha = 10\%$

2 ☐ Since 50 is contained in the 90% parametric bootstrap confidence interval, the null hypothesis is rejected at $\alpha = 10\%$

3 ☐ Since 50 is contained in the 95% parametric bootstrap confidence interval, the null hypothesis cannot be rejected at $\alpha = 5\%$

4 ☐ Since 50 is contained in the 90% non-parametric bootstrap confidence interval, the null hypothesis cannot be rejected at $\alpha = 10\%$

5 ☐ Since 50 is contained in the 95% non-parametric bootstrap confidence interval, the null hypothesis cannot be rejected at $\alpha = 5\%$

**Question X.2 (16)**

What is the sample mean of the collected (original) sample?

1 ☐ $\bar{y} = 74.364 - 32.912 = 41.452$

2 ☐ $\bar{y} = 74.364 \cdot 0.05 + 32.912 \cdot 0.95 = 34.985$

3 ☐ $\bar{y} = 74.364 \cdot 0.95 - 32.912 \cdot 0.05 = 69.000$

4 ☐ $\bar{y} = (74.364 + 32.912)/2 = 53.638$

5 ☐ The sample mean cannot be determined from the given information

**Exercise XI**

The electricity consumption (in KWh) for electronic vehicles belonging to residents outside greater Copenhagen, who charge their vehicles at public charging stations, was monitored on a weekly basis during the winter of 2023. The observations were stored in the vector `x` and the following non-parametric bootstrapping procedure was run:

```
set.seed(696969)
# Observations
x <- c(230,205,203,212,292,260,317,321,235)

# Number of bootstrap samples
k <- 100000

# Number of observations in each bootstrap sample
n <- length(x)

# Generate k bootstrap samples
simsamples <- replicate(k, sample(x, replace = TRUE))

# Calculate the median of each generated sample
simmedians <- apply(simsamples, 2, median)

# 90% confidence interval for the median
quantile(simmedians, c(0.05, 0.95),type=2)

##  5% 95%
## 212 292

# 99% confidence interval for the median
quantile(simmedians, c(0.005, 0.995),type=2)

##  0.5% 99.5%
##   205   317
```

**Question XI.1 (17)**

Which sample could constitute a non-parametric bootstrap sample found in `simsamples`?

1 ☐ $(250, 203, 292, 260, 268)$

2 ☐ $(199, 203, 292, 260, 268)$

3 ☐ $(230, 205, 203, 212, 292, 260, 317, 321, 235)$

4 ☐ $(199, 205, 203, 212, 292, 260, 317, 321, 235)$

5 ☐ $(212, 292)$

**Question XI.2 (18)**

Which statement is not correct? Each statement is about a null hypothesis for the median level of the electricity consumption $q_{0.5}$ and the conclusion is drawn based on the results of the R code. (Note again: there are 4 true and 1 false statements. You must find the false statement.)

1 ☐ The null hypothesis $H_0 : q_{0.5} = 200$ is rejected on a 1% significance level

2 ☐ The null hypothesis $H_0 : q_{0.5} = 200$ is rejected on a 5% significance level

3 ☐ The null hypothesis $H_0 : q_{0.5} = 200$ is rejected on a 10% significance level

4 ☐ The null hypothesis $H_0 : q_{0.5} = 300$ is rejected on a 1% significance level

5 ☐ The null hypothesis $H_0 : q_{0.5} = 300$ is rejected on a 10% significance level

Continue on page 17

**Exercise XII**

In financial engineering, the capital market line relates the return of optimal portfolios with their volatility. The model is given as:

$$R_i = \alpha + \beta v_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where the errors are assumed to be independent. In the model, $\alpha$ is referred to as the risk-free return and $\beta$ as the optimal Sharpe ratio. Moreover, $R_i$ and $v_i$ denote the return and the volatility of portfolio $i$, respectively. Consider the sample in the below table:

| Portfolio | 1 | 2 | 3 | 4 | 5 | Mean |
|---|---|---|---|---|---|---|
| Volatility | 2.4% | 3.9% | 6.1% | 10.2% | 22.0% | 8.92% |
| Return | 5.0% | 8.3% | 15.5% | 23.8% | 45.2% | 19.56% |

The model is fitted to the sample data in the table using the method of least squares.

**Question XII.1 (19)**

What are the estimates of the model parameters?

1 ☐ $\hat{\alpha} = 1.4712, \quad \hat{\beta} = 2.0279, \quad \hat{\sigma} = 1.7507$

2 ☐ $\hat{\alpha} = 0.0147, \quad \hat{\beta} = 2.0279, \quad \hat{\sigma} = 0.0175$

3 ☐ $\hat{\alpha} = 2.0279, \quad \hat{\beta} = 0.0147, \quad \hat{\sigma} = 0.0175$

4 ☐ $\hat{\alpha} = 0.0147, \quad \hat{\beta} = 2.0279, \quad \hat{\sigma} = \sqrt{0.0175}$

5 ☐ $\hat{\alpha} = -0.0064, \quad \hat{\beta} = 0.4887, \quad \hat{\sigma} = 0.0085$

The principle of 'no free lunches' in financial engineering states that you cannot expect a return without taking risk. Inspired by this principle, we test the null hypothesis that the risk-free return ($\alpha$) is zero against a two-sided alternative hypothesis.

**Question XII.2 (20)**

Assuming the model assumptions are satisfied, which statement about the usual test of the null hypothesis is correct using a 5% significance level?

1 ☐ The $p$-value is 32.82%, and the null hypothesis is therefore rejected

2 ☐ The $p$-value is 32.82%, and the null hypothesis is therefore accepted

3 ☐ The $p$-value is 0.04%, and the null hypothesis is therefore rejected

4 ☐ The $p$-value is 0.04%, and the null hypothesis is therefore accepted

5 ☐ None of the above statements are correct

**Question XII.3 (21)**

For which of the following new portfolios is the model estimate of the expected return associated with the least uncertainty, as indicated by the narrowest 95% confidence interval?

1 ☐ Portfolio A with a volatility of 2%

2 ☐ Portfolio B with a volatility of 4%

3 ☐ Portfolio C with a volatility of 6%

4 ☐ Portfolio D with a volatility of 8%

5 ☐ Portfolio E with a volatility of 10%

Continue on page 19

**Exercise XIII**

A robot has been playing a game for two consecutive years and each year its scores have been recorded. Researchers are interested in determining whether there has been a significant change in the robot's scores from the first year to the second.

The first year the robot played $n_1$ games with a mean score of $\bar{x}_1$ and a (sample) standard deviation of $s_1$. The second year it played $n_2$ games and had a mean score of $\bar{x}_2$ and a (sample) standard deviation of $s_2$.

**Question XIII.1 (22)**

The researchers want to test whether the mean scores of the two years are significantly different. They assume the observations are independent and normally distributed, and they ensure that the samples include enough games to satisfy the assumptions of the appropriate (usual) test.

Which formula correctly calculates the observed test statistic used in the usual test?

1 ☐ $t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(s_1^2 - s_2^2)/n_1 - n_2}}$

2 ☐ $t_{\text{obs}} = \frac{(\bar{x}_1 + \bar{x}_2)}{\sqrt{(s_1^2 - s_2^2)/n_1 - n_2}}$

3 ☐ $t_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$

4 ☐ $t_{\text{obs}} = \frac{\bar{x}_1}{\sqrt{s_1^2/n_1}} - \frac{\bar{x}_2}{\sqrt{s_2^2/n_2}}$

5 ☐ $t_{\text{obs}} = \frac{\bar{x}_1}{\sqrt{s_1^2/n_1}} + \frac{\bar{x}_2}{\sqrt{s_2^2/n_2}}$

**Question XIII.2 (23)**

The researchers want to schedule some games between two robots. Since the games can vary substantially in scores, the researchers are only interested in the score difference in each game. Consequently, the researchers want to apply a one sample $t$-test.

The researchers want to detect if there is a mean score difference of at least 10 points, and they want the test to have a power of 80% using a significance level of 5%.

Assuming a standard deviation in score differences of 50, how many games should at least be scheduled?

1 ☐ 17 or 19 depending on whether you apply the normal approximation

2 ☐ 32 or 34 depending on whether you apply the normal approximation

3 ☐ 155 or 157 depending on whether you apply the normal approximation

4 ☐ 197 or 199 depending on whether you apply the normal approximation

5 ☐ 392 or 394 depending on whether you apply the normal approximation

Continue on page 21

Exercise XIV

Engineers at CERN run experiments under various physical conditions. A particular experiment is concerned with determining the kinetic energy of a special quantum particle at three different temperatures and five different magnetic field strengths. Since the experiments are very expensive, the engineers only run the experiment once under each combination of temperature and magnetic field strength. The engineers therefore formulate the model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad i \in \{1, ..., 3\}, \quad j \in \{1, ..., 5\},$$

where the errors are assumed to be independent.

**Question XIV.1 (24)**

The engineers test the usual null hypothesis of no difference between the magnetic field strength effects using a significance level of 1%. What is the critical value in the usual test of the null hypothesis?

1 ☐ 5.99

2 ☐ 7.01

3 ☐ 7.34

4 ☐ 8.65

5 ☐ 8.81

Before conducting the experiment, the engineers planned on comparing the kinetic energy of the quantum particle at all the three different temperatures. Therefore, they calculate pairwise confidence intervals for all the possible differences between the temperature effects. To control the family wise type I error rate, the engineers decide to use Bonferroni correction.

**Question XIV.2 (25)**

If they choose an overall significance level of 1%, what corrected significance level should be used in the confidence intervals?

1 ☐ 0.333%

2 ☐ 0.500%

3 ☐ 0.667%

4 ☐ 1.000%

5 ☐ 3.000%

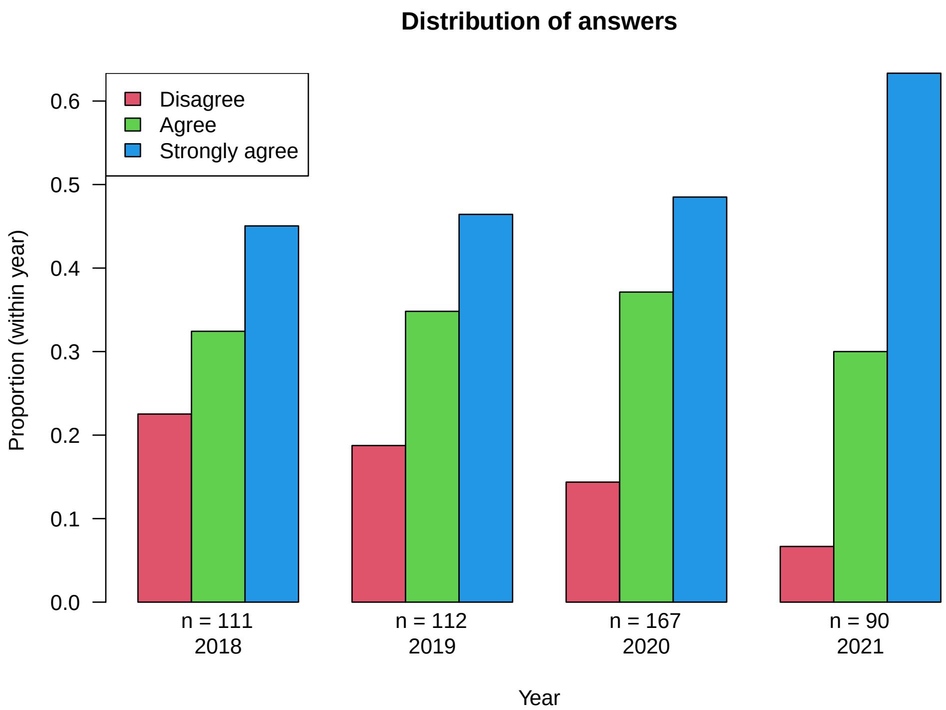
Continue on page 22

Exercise XV

The annual student evaluation of an introductory statistics course included the statement ”I spend excessive time on my computer.” A table displaying the frequencies of answers for four consecutive years is given below:

| | 2018 | 2019 | 2020 | 2021 |
|---|---|---|---|---|
| Disagree | 25 | 21 | 24 | 6 |
| Agree | 36 | 39 | 62 | 27 |
| Strongly agree | 50 | 52 | 81 | 57 |

To visually represent the responses, the proportions of each answer were plotted across the years:




Consider the null hypothesis: ”The distribution of answers is the same for each year.”

**Question XV.1 (26)**

The usual test statistic follows a $\chi^2$-distribution under the null hypothesis. What is the critical value of the test statistic on a significance level $\alpha = 0.05$?

1 ☐ 10.64

2 ☐ 12.59

3 ☐ 14.45

4 ☐ 16.81

5 ☐ 18.55

**Question XV.2 (27)**

The usual test of the null hypothesis leads to what $p$-value and conclusion on a 5% significance level?

1 ☐ Since the $p$-value is 0.2463, none of the years have a significantly different distribution of answers than the other years

2 ☐ Since the $p$-value is 0.06288, at least one year has a significantly different distribution of answers than the other years

3 ☐ Since the $p$-value is 0.06288, none of the years have a significantly different distribution of answers than the other years

4 ☐ Since the $p$-value is 0.03144, at least one year has a significantly different distribution of answers than the other years

5 ☐ Since the $p$-value is 0.03144, none of the years have a significantly different distribution of answers than the other years

**Question XV.3 (28)**

The usual test of the null hypothesis relies on an approximation of the $\chi^2$-distribution, which is valid when the expected count is at least five in all cells. The cell with the lowest expected count (value) is "Disagree" in 2021.

Is the approximation of the $\chi^2$-distribution valid in this case (both conclusion and argument must be correct)?

1 ☐ No, since the lowest expected cell count is 2.25

2 ☐ Yes, since the lowest expected cell count is 6

3 ☐ Yes, since the lowest expected cell count is 14.25

4 ☐ Yes, since the lowest expected cell count is 17.58

5 ☐ Yes, since the lowest expected cell count is 45

Continue on page 25

**Exercise XVI**

Engineers at a large Danish company fit a multiple linear regression model

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

where the errors are assumed to be independent. The model is fitted to a data set with 33 observations and the following results are obtained from the usual R-output:

| Parameter | Estimate | Std. error | $t$-value | $p$-value |
|---|---|---|---|---|
| $\beta_0$ | 1.250 | 0.763 | 1.650 | 0.1094 |
| $\beta_1$ | 4.045 | 0.366 | 11.052 | $< 0.0001$ |
| $\beta_2$ | 6.932 | 1.764 | 9.390 | $< 0.0001$ |

The engineers first test the null hypothesis:

$$\mathcal{H}_{01}: \quad \beta_1 = 4.5$$

against a two-sided alternative hypothesis.

**Question XVI.1 (29)**

What is the value of the usual test statistic under the null hypothesis?

1 ☐ -7.141

2 ☐ -1.243

3 ☐ 1.243

4 ☐ 1.924

5 ☐ 11.052

**Question XVI.2 (30)**

The usual test of the null hypothesis $\mathcal{H}_{01}$ produces what $p$-value?

1 ☐ 11.17%

2 ☐ 11.19%

3 ☐ 22.28%

4 ☐ 22.34%

5 ☐ 22.38%

Continue on page 26

The exam is finished. Enjoy the summer!