This exam paper is available in both Danish and English. The English version appears after the Danish version.

Exam question paper for:

Written examination: 26. June 2026

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

Duration: 4 hours

Aids allowed: All printed materials and a pocket calculator of type TI30XS or TI30XB

Weighting: The questions are weighted equally (see details below)

Final answers must be submitted by completing the separate "Answer Sheet".

This examination consists of 30 multiple-choice questions distributed across 10 exercises.
Only the "Answer Sheet"must be submitted; do not hand in the exam paper itself.

Multiple choice questions: *The exam is 30 multiple choice questions, with possible answers: 1, 2, 3, 4, 5, and 6 (where 6 always refers to the answer "Don't know/No answer"). A correct answer gives 5 points, a wrong answer gives −1 point, and "Don't know/No answer" (6) gives 0 points. It is not allowed to select more than one answer option for a single question. If this happens anyway, 0 points are given for the question.*

The use of Python code in this exam: *This examination includes Python code. Note that the following libraries and abbreviations are used:*

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import scipy.stats as stats
import statsmodels.api as sm
import statsmodels.formula.api as smf
import statsmodels.stats.power as smp
import statsmodels.stats.proportion as smprop
```

Continue on page 2

Exercise I

A student is taking a multiple choice exam with 30 questions. For each question there are five options where one and only one is correct. The student is wondering what the probability of answering at least 15 correctly if he answers every question completely at random.

Question I.1 (1)

Which of the following Python commands calculate the desired probability?

1 ☐ `1 - stats.hypergeom.cdf(14, 30, 6, 15)`

2 ☐ `1 - stats.hypergeom.cdf(15, 30, 5, 15)`

3* ☐ `1 - stats.binom.cdf(14, 30, 1/5)`

4 ☐ `1 - stats.hypergeom.pmf(15, 30, 6, 14)`

5 ☐ `1 - stats.binom.pmf(15, 30, 1/5)`

6 ☐ Don't know / No answer

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

Answering at random imply that the number of correct answer follow a Binomial with success probability $p = 1/5$, and $n = 30$. Further the probability in question is

$$P(X \geq 15) = 1 - P(X \leq 14) \tag{1}$$

where $X \sim Binom(1/6, 30)$, which can be calculated by

```
1-stats.binom.cdf(14,30,1/5)

np.float64(0.00023122561166777356)
```

```
------------------------------------ FACIT-END ------------------------------------
```

In the exam students are awarded 5 points for a correct answer and -1 point for an incorrect answer.

Question I.2 (2)

Let the random variable Y be the total number of points achieved in the exam, still assuming that the student answer completely at random, what is the standard deviation of Y?

1 ☐ $\sqrt{30\left(\frac{4}{5}+5^2\cdot\frac{1}{5}\right)}=\sqrt{174}$

2* ☐ $\sqrt{30\left(1.2^2\cdot\frac{4}{5}+4.8^2\cdot\frac{1}{5}\right)}=\sqrt{172.8}$

3 ☐ $30\sqrt{\frac{4}{5}+5^2\cdot\frac{1}{5}}=30\sqrt{5.8}$

4 ☐ $30\sqrt{\frac{4}{5^2}}=30\sqrt{0.16}$

5 ☐ $\sqrt{30\cdot\frac{4}{5^2}}=\sqrt{4.8}$

6 ☐ Don't know / No answer

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

The X_i be random variable denoting the point for one question. The expected value X_i is

$$E[X_i] = -1\cdot\frac{4}{5}+5\cdot\frac{1}{5}=\frac{1}{5} \tag{2}$$

and the variance of X_i is

$$V[X_i] = \left(-1-\frac{1}{5}\right)^2\frac{4}{5}+\left(5-\frac{1}{5}\right)^2\frac{1}{5}=1.2^2\frac{4}{5}+4.8^2\frac{1}{5} \tag{3}$$

and hence the variance of the total number of points (Y) is

$$V[Y] = 30\left(1.2^2\frac{4}{5}+4.8^2\frac{1}{5}\right) \tag{4}$$

taking the square root we get answer 2.

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

Continue on page 4

Exercise II

A researcher is planning to study the effect of a new drug that reduces the number of recovery days needed after a certain procedure. In a previous pilot study, it was found that patients who were given the drug needed, on average, 19 days to recover from the procedure. This result seems promising, as patients who are not given the drug need, on average, 23 days to recover. The estimated standard deviation of the number of recovery days is 12 days, and it is assumed that this standard deviation applies to both patients who are given the drug and those who are not.

Since the previous pilot study was based on only a small number of patients, the researcher plans to conduct a new study with a larger sample. The researcher will measure the number of recovery days in a sample of n patients and estimate the average number of recovery days, and it is assumed that the data are normally distributed. The researcher will then test whether this average is significantly different from 23 days (the null hypothesis). A significance level of $\alpha = 0.05$ will be used, and the researcher wants the study to have a statistical power of 70% ($1 - \beta$=0.70) under an alternative hypothesis where the mean and variance equal the values observed in the pilot study.

To answer the following question, we provide the following quantiles from a standard normal distribution:

$z_{0.50}$	$z_{0.60}$	$z_{0.70}$	$z_{0.80}$	$z_{0.90}$	$z_{0.95}$	$z_{0.975}$	$z_{0.99}$	$z_{0.995}$
0.00	0.25	0.52	0.84	1.28	1.64	1.96	2.33	2.58

Question II.1 (3)

How large should the sample size n be?

1 ☐ 5

2 ☐ 42

3 ☐ 44

4* ☐ 56

5 ☐ 71

6 ☐ Don't know / No answer

`----------------------------------- FACIT-BEGIN -----------------------------------`

Method 3.65

$$n = \left(\frac{\sigma\left(z_{1-\beta}+z_{1-\alpha/2}\right)}{\mu_0-\mu_1}\right)^2 = \left(\frac{12(0.52+1.96)}{4}\right)^2 = 55.4$$

We must round up, so the answer is 56.

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

Continue on page 6

Question II.2 (4)

Which of the following statements is FALSE?

1 ☐ The risk of making a Type I error in the new study is 5%

2* ☐ The risk of making a Type II error in the new study is 70%

3 ☐ The significance level α is the same as the risk of making a Type I error

4 ☐ The statistical power is the same as *the probability of correctly rejecting the null hypothesis if the null hypothesis is false.*

5 ☐ If the true effect of the drug is smaller than hypothesized, then the statistical power of the study will also be smaller.

6 ☐ Don't know / No answer

`----------------------------------- FACIT-BEGIN -----------------------------------`

The risk of making a Type II error in the new study is 30% ($\beta = 0.30$), hence option 2 is false. Lets just go through the other options:

1: This is the definition of the significance level and hence it is correct

3: Again the definition of significance level (just phrased slightly different)

4: This is the definition of power, hence it is also true

5: The larger the true effect the larger the power, and of course the other way around, hance this is also correct

and hence the correct answer is option 2.

`------------------------------------ FACIT-END ------------------------------------`

Continue on page 7

Exercise III

In a study, a university wants to investigate whether the level of Generative AI declaration in student reports is associated with the type of academic project.

After introducing a compulsory library course on responsible use of Generative AI, students are required to explicitly state in the methods chapter whether they used Generative AI tools (such as ChatGPT, Claude, or Copilot) as part of their project or for writing their reports.

The level of AI declaration in the report is classified as: (1) Explicit declaration (clearly documented use), (2) Partial declaration (mentioned "what" but not "how"), and (3) No declaration.

The project type is classified as: (1) Bachelor project, (2) Master thesis, and (3) Other project report.

In a random sample of 240 student reports, the results are summarized in the table below:

	AI Declaration Level:			
Project Type:	Explicit	Partial	None	Total
Bachelor project	40	20	20	80
Master thesis	50	30	20	100
Other report	10	20	30	60
Total	100	70	70	240

The researcher wants to test the hypothesis of independence between Generative AI declarations and project type. The desired significance level α has been stored in Python in a variable called `alpha`.

Question III.1 (5)

Which of the following Python code lines correctly calculates the critical value for the relevant hypothesis test?

1 ☐
```
critical_value = stats.chi2.ppf(1 - alpha/2, df = 239)
```

2* ☐
```
critical_value = stats.chi2.ppf(1 - alpha, df=4)
```

3 ☐
```
critical_value = stats.f.ppf(1 - alpha, dfn=2, dfd=4)
```

4 ☐
```
critical_value = stats.f.cdf(1 - alpha, dfn=2, dfd=6)
```

5 ☐
```
critical_value = stats.t.cdf(1 - alpha/2, df=8)
```

6 ☐ Don't know / No answer

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

This problem involves two categorical variables with three categories each:

- Project type (Bachelor, Master, Other)
- AI declaration level (Explicit, Partial, None)

Thus the data forms a 3×3 contingency table.

We therefore apply a Chi-square test of independence.

Degrees of freedom:

$$df = (r-1)(c-1) = (3-1)(3-1) = 4$$

The hypothesis test compares the observed test statistic with the critical value:

$$\chi^2_{1-\alpha,df}$$

In Python, the critical value is obtained using the inverse cumulative distribution function (percent point function) of the chi-square distribution:

```
stats.chi2.ppf(1 - alpha, df=4)
```

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

Question III.2 (6)

Continuing with the contingency table on Generative AI declaration in student reports (Bachelor projects, Master theses, and other reports), a researcher now wants to perform a test of independence between generative AI declaration and student report type, using Python.

The observed counts from the study are stored in Python as the following array:

```
observed = np.array([[40, 20, 20],
                     [50, 30, 20],
                     [10, 20, 30]])
```

Which of the following Python code lines correctly performs the test for this contingency table?

1 ☐ `stats.chi2.cdf(observed)`

2 ☐ `stats.chi2.ppf(observed)`

3* ☐ `stats.chi2_contingency(observed)`

4 ☐ `stats.ttest_ind(observed)`

5 ☐ `stats.chi2.pdf(observed)`

6 ☐ Don't know / No answer

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

The correct Python function for performing a χ^2-test of independence on a contingency table is:

`stats.chi2_contingency()`

This function takes the contingency table (observed frequencies) as input and returns:

- the Chi-square test statistic
- the p-value
- the degrees of freedom
- the expected frequencies

Example usage:

```
chi2, p, dof, expected = stats.chi2_contingency(observed)
```

This test evaluates whether the two categorical variables (here: project type and AI declaration level) are statistically independent.

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

Continue on page 10

Exercise IV

Researchers record test scores of three large language models (GPT, Claude, and Gemini) in three domains (Reasoning, Coding, and Mathematics). The scores of the models and across the domains are given below:

Point scores	GPT model	Claude model	Gemini model	Average
Reasoning	92.4	84.3	91.9	89.5
Coding	80.0	82.0	76.2	79.4
Mathematics	98.8	95.3	98.6	97.6
Average	90.4	87.2	88.9	88.8

The researchers fit a two-way ANOVA model

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

with $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, to the data under the usual assumptions. For the next two questions, we simply assume that the model assumptions are satisfied (although this is probably not the case).

Question IV.1 (7)

What is the estimated effect of the GPT model according to the model?

1 ☐ 90.40

2 ☐ 30.1

3 ☐ 10.0

4 ☐ 9.5

5* ☐ 1.6

6 ☐ Don't know / No answer

```
---------------------------------- FACIT-BEGIN ----------------------------------
```

Equations (8-38) through (8-40) provide parameter estimates for the two-way ANOVA model. The equations show that the treatment and block effects are estimated by taking the treatment/block average and subtracting the overall average. Thus, in the notation of the book:

$$\hat{\beta}_1 = 90.4 - 88.8 = 1.6.$$

```
----------------------------------- FACIT-END -----------------------------------
```

Continue on page 11

Question IV.2 (8)

What is the domain sum of squares, SS(domain), according to the model?

1 ☐ 71520

2* ☐ 499

3 ☐ 166

4 ☐ 15

5 ☐ This cannot be determined without further information.

6 ☐ Don't know / No answer

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

According to theorem 8.20:

$$SS(domain) = l \cdot \sum_{i=1}^{k} \hat{\alpha}_i^2 = 3 \cdot \sum_{i=1}^{3} \hat{\alpha}_i^2 = 3\left((89.5-88.8)^2 + (79.4-88.8)^2 + (97.6-88.8)^2\right) = 498.87.$$

Therefore, the correct answer is that SS(domain) equals 499.

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

Continue on page 12

Exercise V

Let Y_1 and Y_2 be independent and normally distributed random variables with mean $\mu = 0$ and variance $\sigma^2 = 3^2$. Further, let $Y = Y_1 - Y_2$ be the difference between Y_1 and Y_2.

Question V.1 (9)

Let z_α be the α-quantile of the standard normal distribution. Which of the following intervals covers 90% of all outcomes of Y?

1 ☐ $[z_{0.025} 2 \cdot 3^2; z_{0.975} 2 \cdot 3^2]$

2 ☐ $[z_{0.05} 6; z_{0.95} 6]$

3* ☐ $[z_{0.05} \sqrt{2} \cdot 3; z_{0.95} \sqrt{2} \cdot 3]$

4 ☐ $[z_{0.05} \frac{3}{\sqrt{2}}; z_{0.95} \frac{3}{\sqrt{2}}]$

5 ☐ $[z_{0.025} 3; z_{0.975} 3]$

6 ☐ Don't know / No answer

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

First find the mean and variance of Y,

$$E[Y] = 0; \quad V[Y] = 3^2 + 3^2 = 2 \cdot 3^2 \tag{5}$$

and hence we have

$$Z = \frac{Y}{\sqrt{2} \cdot 3} \sim N(0,1) \tag{6}$$

and

$$P\left(z_{0.05} < Z < z_{0.95}\right) = P\left(z_{0.05} < \frac{Y}{\sqrt{2} \cdot 3} < z_{0.95}\right) = P\left(z_{0.05}\sqrt{2} \cdot 3 < Y < z_{0.95}\sqrt{2} \cdot 3\right) \tag{7}$$

this is answer no. 3.

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

Now consider the random variable $R = \sqrt{Y_1^2 + Y_2^2}$ (you may think of R as the distance from a center, e.g. a shooter is trying to hit a target and Y_1 and Y_2 represent the deviation in the horizontal and vertical axis).

Question V.2 (10)

What is the expectation and variance of R^2?

1 ☐ $E(R^2) = 9$, and $V(R^2) = 4$

2 ☐ $E(R^2) = 18$, and $V(R^2) = 2 \cdot 3^3$

3 ☐ $E(R^2) = 18$, and $V(R^2) = \sqrt{2} \cdot 3^2$

4* ☐ $E(R^2) = 18$, and $V(R^2) = 3^4 \cdot 4$

5 ☐ $E(R^2) = 9$, and $V(R^2) = \sqrt{2} \cdot 3^2$

6 ☐ Don't know / No answer

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

We can write R^2 as

$$R^2 = Y_1^2 + Y_2^2 = 3^2 \left(\frac{Y_1^2}{3^2} + \frac{Y_2^2}{3^2} \right) \tag{8}$$

and since $\frac{Y_i^2}{3^2} \sim \chi^2(1)$ and hence

$$R^2 = 3^2 X^2; \quad X^2 \sim \chi^2(2) \tag{9}$$

and therefore

$$E(R^2) = 3^2 E(X^2) = 3^2 \cdot 2 = 18 \tag{10}$$
$$V(R^2) = 3^4 V(X^2) = 3^4 \cdot 2 \cdot 2 = 3^4 \cdot 4 \tag{11}$$

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

Continue on page 14

Now, let μ_{R^2} and $\sigma^2_{R^2}$ denote the mean and variance of R^2.

Question V.3 (11)

According to the approximate non-linear error propagation rule, what is the variance of $R = \sqrt{R^2}$?

1* ☐ $V(R) \approx \frac{\sigma^2_{R^2}}{4\mu_{R^2}}$

2 ☐ $V(R) \approx \frac{\sigma^2_{R^2}}{4(\mu_{R^2})^2}$

3 ☐ $V(R) \approx \frac{\sigma^2_{R^2}}{2\mu_{R^2}}$

4 ☐ $V(R) \approx \frac{\sigma^2_{R^2}}{2(\mu_{R^2})^2}$

5 ☐ $V(R) \approx \frac{\sigma^2_{R^2}}{(\mu_{R^2})^2}$

6 ☐ Don't know / No answer

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

$R = \sqrt{R^2}$ and hence $\frac{\partial R}{\partial R^2} = \frac{\partial \sqrt{R^2}}{\partial R^2} = \frac{1}{2\sqrt{R^2}}$, now inserting in the error propagation we get

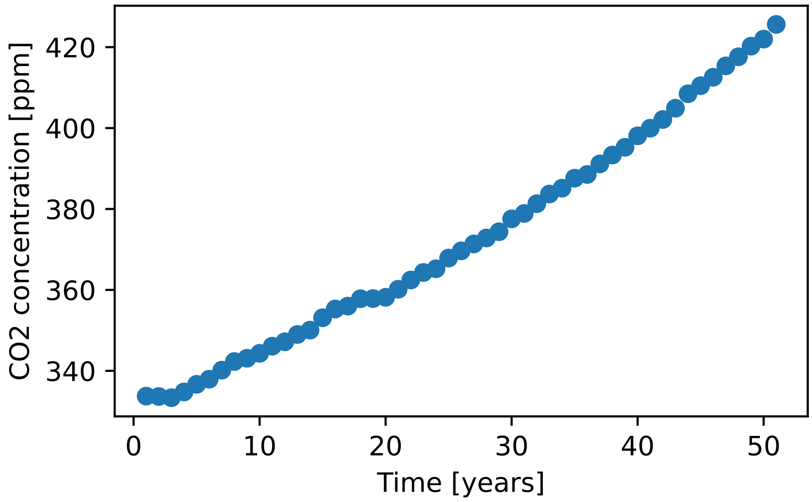
$$V(R) \approx \left(\frac{1}{2\sqrt{\mu_{R^2}}} \right)^2 \sigma^2_{R^2} = \frac{\sigma^2_{R^2}}{4\mu_{R^2}} \tag{12}$$

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

Continue on page 15

Exercise VI

The plot below shows average yearly CO_2-concentrations as a function of time since 1973 (i.e. the last observation is 2024) at a specific measurement station.



Question VI.1 (12)

Let r denote the empirical correlation coefficient between "Time" and CO_2-concentration. Which of the following statements about r could be correct?

1 ☐ $r \approx 0.8$

2 ☐ $r \approx 0$

3* ☐ $r \approx 0.99$

4 ☐ $r \approx -0.5$

5 ☐ $r \approx 0.5$

6 ☐ Don't know / No answer

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

The points fall almost on a straigth line with a positive slope. So imidiately answer 2 and 4 are out. Comparing to the figure in chapter 1.4 it is also clear that the correlation is stronger than 0.8 (actually it is clear that it is above 0.95), and hence the corrrect answer is 3.

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

In order to explore the development in CO_2-concentration as a function of time, the following analysis have been performed in Python (`value` denotes the CO_2-concentration, and `co2` is a data-frame containing the data shown above). Some values have been replaced by letters.

Continue on page 17

```
fit1 = smf.ols(formula = "value ~ time", data = co2).fit()
print(fit1.summary(slim=True))
                            OLS Regression Results
==============================================================================
Dep. Variable:                  value   R-squared:                       R2
Model:                            OLS   Adj. R-squared:                  R2a
No. Observations:                  51   F-statistic:                     4128.
Covariance Type:            nonrobust   Prob (F-statistic):           5.68e-49
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    325.14       0.86           t1       p1         l1            u1
time           1.84       0.03           t2       p2         l2            u2
==============================================================================

print(np.sqrt(fit1.scale))
3.00
```

Question VI.2 (13)

What is the approximate values of `l2` and `u2`?

1 ☐ `l2`≈ 1.42, `u2`≈ 2.26

2* ☐ `l2`≈ 1.78, `u2`≈ 1.90

3 ☐ `l2`≈ 1.00, `u2`≈ 2.68

4 ☐ `l2`≈ 1.83, `u2`≈ 1.85

5 ☐ `l2`≈ 0.16, `u2`≈ 3.52

6 ☐ Don't know / No answer

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

The numbers `l2` and `u2` makes the 95% confidence interval for the parameter, i.e. it is defined as

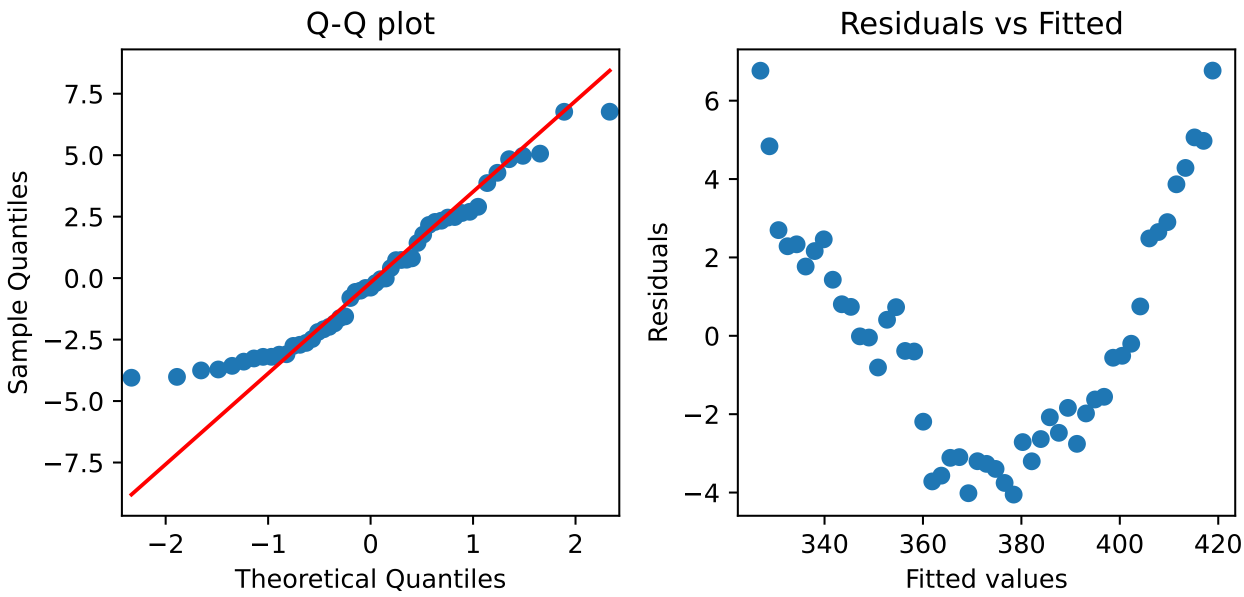
$$[\texttt{l2}, \texttt{u2}] = \hat{\beta}_1 \pm t_{0.975}(49) se_{\beta_1} \tag{13}$$

the 0.975 quantile in the t-distribution is around 2 (it is 1.96 for the normal and slightly higher for the specific t-distribution). Hence

$$[\texttt{l2}, \texttt{u2}] \approx 1.84 \pm 2 \cdot 0.03 = [1.78, 1.90]. \tag{14}$$

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

In order to validate the model assumptions the diagnostic plots (based on the model presented above) below have been created.



Question VI.3 (14)

Based on the diagnostic plots, which of the following statements is correct (all parts of the answer must be correct)?

1 ☐ The independence assumption is clearly violated (Q-Q plot), and the variance of the residuals depends on the level (Residuals vs. Fitted).

2 ☐ A data transformation should be considered (Q-Q plot), and "Residual vs. Fitted" suggest a log-transformation.

3 ☐ The "Q-Q plot" suggests that the normality assumption is fulfilled (most dots are on the straight line), further it seems that the correlation between fitted values and residuals is close to zero, and hence all assumptions seems to be fulfilled.

4 ☐ The "Q-Q plot" suggest that systematic effects are missing in the model, and further "Residuals vs. fitted" suggest that the residual variance is a function of fitted values.

5* ☐ The normality assumption seems to be violated (Q-Q plot), and there are clearly systematic effects which are not included in the model (Residual vs. Fitted).

6 ☐ Don't know / No answer

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

The QQ-plot assess the normality assumption (which is violated in this case), and Residuals vs. Fitted show clear systematic effects (that should be modeled) in the residuals, this is answer 5. We berifly go through the other answwering options:

1) The independent assumption is not checked by the QQ-plot

2) Resiudal vs. fitted does not suggest a log-transformation.

3) The assumption are definitely not fulfilled

4) The QQ-plot does not suggest systematic effects.

```
------------------------------------ FACIT-END ------------------------------------
```

Continue on page 20

After careful consideration and model development, the following model is proposed (with input variables time= $\{1, 2, ..., 51\}$, time2= $\{1, 2^2, ..., 51^2\}$, and time3= $\{1, 2^3, ..., 51^3\}$).

```
fit2 = smf.ols(formula = "value ~ time + time2 + time3", data = co2).fit()
print(fit2.summary(slim=True))
                            OLS Regression Results
==============================================================================
Dep. Variable:                  value   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.999
No. Observations:                  51   F-statistic:                 2.109e+04
Covariance Type:            nonrobust   Prob (F-statistic):           1.60e-73
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    330.6110      0.466    708.890      0.000     329.673     331.549
time           1.3488      0.077     17.535      0.000       1.194       1.504
time2          0.0018      0.003      0.517      0.608      -0.005       0.009
time3          0.0002   4.32e-05      3.796      0.000    7.72e-05       0.000
==============================================================================


print(np.sqrt(fit2.scale))
0.77
```

Question VI.4 (15)

Based on the Python summary above (i.e., regarding the inputs as independent) and using significance level $\alpha = 0.05$, which of the following conclusions is correct (all parts of the answer must be correct)?

1 ☐ Since $1.6 \cdot 10^{-73} << 0.05$, all parameters should be kept in the model.

2 ☐ Since both 0.0018 and 0.0002 are both less than 0.05, we may exclude time2 and time3 from the model.

3* ☐ Since $0.608 > 0.05$, time2 may be omitted.

4 ☐ Since 0.466 and 0.077 are both larger than 0.05, the Intercept and time may be omitted in the model.

5 ☐ Since the confidence interval for time3 is $[7.72 \cdot 10^{-5}; 0.000]$, time3 should be excluded from the model.

6 ☐ Don't know / No answer

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

We go through the options:

1) the p-value $1.6 \cdot 10^{-73}$, is not related to the hypothesis that none of the parameters can be omitted.

2) Parameter estimates should not be compared to the significance level.

3) The p-value relate to the parameter and since it is greater than 0.05 it imply that `time2` may be omitted from the model (hence it is correct).

4) Again parameters should not be compared to the significance level

5) The p-value for `time3` suggest that it should be kept in the model, the 0.000 in the upper bound is simply due to rounding in the summary.

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

The purpose of the model is to make predictions, and the following confidence interval for a prediction 5 years ahead (i.e., at `time`=56), has been proposed

$$\begin{aligned}\hat{y} \pm ME =& 330.6 + 1.35 \cdot 56 + 0.0018 \cdot 56^2 + 0.0002 \cdot 56^3 \pm \\ & t_{0.975}(47) \cdot \sqrt{0.466^2 + 0.077^2 \cdot 56^2 + 0.003^2 \cdot 56^4 + \left(4.32 \cdot 10^{-5}\right)^2 \cdot 56^6}\end{aligned}$$

where $t_{0.975}(47)$ is the 0.975 quantile of the t-distribution with 47 degrees of freedom.

Question VI.5 (16)

Which of the following comments on the proposed confidence interval is correct?

1 ☐ The degrees of freedom for the t-distribution is wrong. It should be 51.

2 ☐ The calculation is correct if it can be verified that the inputs (`time`, `time2`, and `time3`) are approximately normal.

3* ☐ The calculation under the square root assumes an orthogonal design, which is incorrect in the presented case.

4 ☐ As we are only looking 5 years ahead, the 56 should be replaced by 5 everywhere.

5 ☐ As the empirical correlations between the inputs (`time`, `time2`, and `time3`) is approximately 0, the calculation will also be approximately correct.

6 ☐ Don't know / No answer

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

In general the confidence interval is given by

$$\boldsymbol{x}_{New}^T\hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2}(n-p)\sqrt{\hat{\sigma}^2\boldsymbol{x}_{New}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{New}} \tag{15}$$

In our case $\boldsymbol{x}_{New}^T = [1, 56, 56^2, 56^3]$, and also $se_\beta = \sqrt{\texttt{diag}(\boldsymbol{X}^T\boldsymbol{X})^{-1}}$. What is written under the square root is

$$se_{\beta_0}^2 \cdot 1^2 + se_{\beta_1}^2 \cdot 56^2 + se_{\beta_2}^2 \cdot 56^4 + se_{\beta_3}^2 \cdot 56^6 \tag{16}$$

Hence it assumes that $\boldsymbol{X}^T\boldsymbol{X}$ is a diagonal matrix (which is wrong in the presented case), this is answer option 3.

For the other options we have

1) The degrees of freedom for the t-distribution is correct.

2) There is no requirement on the distribution of the input.

4) Time should be `56`

5) The empirical correlation be the input is not 0. And even if it was correct it would be be sufficient for the calculation to be correct, what is need is the inner product between columns of the design matrix to be zero, and e.g. the inner product between column 1 and 2 is $\sum_{i=1}^{5} 1i > 0$.

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

It is now decided to test the 3rd order model

$$Y_i = \beta_0 + \beta_1\texttt{time}_i + \beta_2\texttt{time}_i^2 + \beta_3\texttt{time}_i^3 + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2),$$

against the null hypothesis that the data can be described by the first order model

$$Y_i = \beta_0 + \beta_1\texttt{time}_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2).$$

In both models it is assumed that the ϵ_i's are iid..

Continue on page 23

Question VI.6 (17)

What is the usual test statistic for comparing (i.e. testing the hypothesis $\beta_2 = \beta_3 = 0$) the two models (some of the answers uses numbers from both the presented Python summaries)?

1 ☐ $\frac{21090-4128}{2} = 8481$

2 ☐ $\frac{21090-4128}{3^2} = 1885$

3 ☐ $\frac{3^2-0.77^2}{0.77^2} = 14.2$

4* ☐ $\frac{(3^2 \cdot 49 - 0.77^2 \cdot 47)/2}{0.77^2} = 347.8$

5 ☐ $\frac{(3^2 \cdot 49 - 0.77^2 \cdot 47)/2}{0.77^2/47} = 16375$

6 ☐ Don't know / No answer

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

Denote the projection matrices resulting from `fit1` and `fit2` by $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ respectively. The statistic we are looking for is

$$\frac{(\boldsymbol{y}^T(\boldsymbol{H}_2 - \boldsymbol{H}_1)\boldsymbol{y})/2}{\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_2)\boldsymbol{y}/47} \tag{17}$$

the number in the denomiator is given directly as

$$\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_2)\boldsymbol{y}/47 = 0.77^2 \tag{18}$$

or

$$\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_2)\boldsymbol{y} = 0.77^2 \cdot 47. \tag{19}$$

In the same way we have

$$\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{y} = 3^2 \cdot 49. \tag{20}$$

We can now rewrite (part of) the denomiator as

$$\begin{aligned}
\boldsymbol{y}^T(\boldsymbol{H}_2 - \boldsymbol{H}_1)\boldsymbol{y} =& \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{I} + \boldsymbol{H}_2 - \boldsymbol{H}_1)\boldsymbol{y} && (21)\\
=& \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_1 - (\boldsymbol{I} - \boldsymbol{H}_2))\boldsymbol{y} && (22)\\
=& \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{y} - \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_2)\boldsymbol{y} && (23)\\
=& 3^2 \cdot 49 + 0.77^2 \cdot 47. && (24)
\end{aligned}$$

Collection we get

$$\frac{(\boldsymbol{y}^T(\boldsymbol{H}_2 - \boldsymbol{H}_1)\boldsymbol{y})/2}{\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_2)\boldsymbol{y}/47} = \frac{(3^2 \cdot 49 + 0.77^2 \cdot 47)/2}{0.77^2}. \tag{25}$$

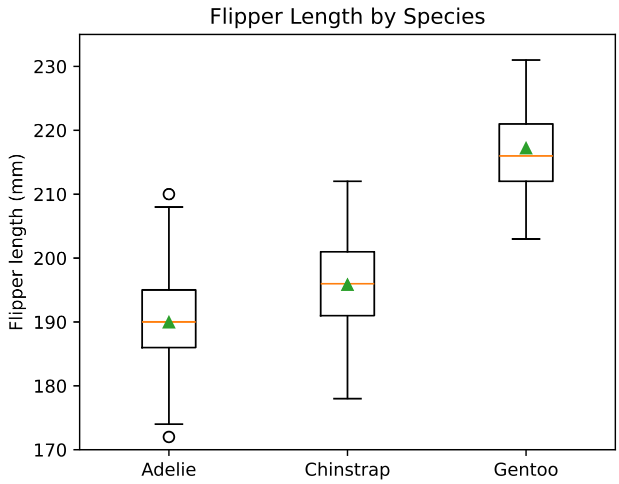
`------------------------------------ FACIT-END ------------------------------------`

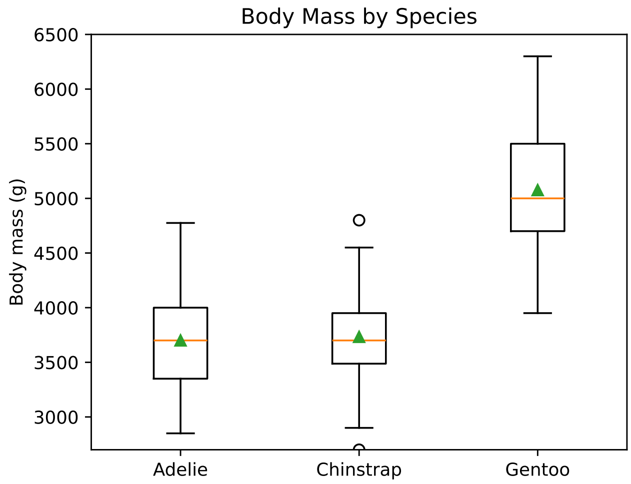
Continue on page 24

Exercise VII

We consider data on penguins from three different species: `Adelie`, `Chinstrap`, and `Gentoo`. The data are stored in a Pandas DataFrame named `penguins` and include the variables: `species`, `flipper_length_mm`, and `body_mass_g`.

The data are visualized in the plots below:





Summary statistics for each species are:

		Flipper length (mm)		Body mass (g)	
Species	n	Mean	Std. dev.	Mean	Std. dev.
Adelie	151	189.95	6.54	3700.7	458.6
Chinstrap	68	195.82	7.13	3733.1	384.3
Gentoo	123	217.19	6.48	5076.0	504.1

Assume that both flipper length and body mass within each species is approximately normally distributed.

To answer the following questions, you may need one or more of the following quantiles ($t_{1-\alpha,\nu}$ is the $1-\alpha$ quantile in a t-distribution with ν degrees of freedom):

Degrees of freedom	$t_{0.95,\nu}$	$t_{0.975,\nu}$	$t_{0.995,\nu}$
$\nu = 67$	1.668	1.996	2.651
$\nu = 120$	1.658	1.980	2.617
$\nu = 122$	1.657	1.980	2.617
$\nu = 150$	1.655	1.976	2.609
$\nu = 200$	1.653	1.972	2.601
$\nu = 250$	1.651	1.969	2.596
$\nu = 300$	1.650	1.968	2.592

Question VII.1 (18)

What is the usual 95% confidence interval for the mean flipper length of Chinstrap penguins?

1 ☐ [195.7; 195.9]

2 ☐ [176.9; 214.7]

3 ☐ [183.9; 207.7]

4* ☐ [194.1; 197.5]

5 ☐ [181.5; 210.1]

6 ☐ Don't know / No answer

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

Sample statistics for Chinstrap penguins:

$$\bar{x} = 195.82, \quad s = 7.13, \quad n = 68$$

We use the t-quantile $t_{0.975,\nu=67} = 1.996$.

The upper and lower limits of the 95% confidence interval are calculated as:

$195.82 \pm 1.996 \cdot 7.13/\sqrt{68}$

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

An older source reports that the mean flipper length of Chinstrap penguins is $\mu_0 = 190$ mm. Use the sample data from the current study to test

$$H_0 : \mu = 190 \quad \text{vs} \quad H_1 : \mu \neq 190.$$

Question VII.2 (19)

At significance level $\alpha = 0.05$, what are the correct test statistic and conclusion of the test?

1 ☐ The test statistic becomes $t \approx 0.816$, and the conclusion is that the observed mean flipper length in the Chinstrap penguins is NOT significantly different from 190 mm (since 0.816 is smaller than 1.996).

2 ☐ The test statistic becomes $t \approx 0.816$, and the conclusion is that the observed mean flipper length in the Chinstrap penguins IS significantly different from 190 mm (since 0.816 is larger than 0.05).

3* ☐ The test statistic becomes $t \approx 6.73$, and the conclusion is that the observed mean flipper length in the Chinstrap penguins IS significantly different from 190 mm (since 6.73 is larger than 1.996).

4 ☐ The test statistic becomes $t \approx 6.73$, and the conclusion is that the observed mean flipper length in the Chinstrap penguins is NOT significantly different from 190 mm (since 6.73% is larger than 5%).

5 ☐ The test statistic becomes $t \approx 5.45$, and the conclusion is that the observed mean flipper length in the Chinstrap penguins is NOT significantly different from 190 mm (since 5.45% is larger than 5%).

6 ☐ Don't know / No answer

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

Sample statistics for Chinstrap penguins:

$$\bar{x} = 195.82, \quad s = 7.13, \quad n = 68$$

Test statistic:

$$t = (195.82 - 190)/(7.13/\sqrt{68}) = 6.73$$

The test statistic should be compared to the critical value 1.996.

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

We wish to investigate whether there is a significant difference in mean body mass between **Adelie** and **Chinstrap** penguins, which corresponds to testing

$$H_0 : \mu_A = \mu_C \qquad \text{vs} \qquad H_1 : \mu_A \neq \mu_C.$$

We will use a significance level of $\alpha = 0.05$ and apply a two-sample t-test that assumes equal variances in the two groups. For this purpose, you may use the pooled two-sample estimate of the variance (estimated from the Adelie and Chinstrap samples):

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_C - 1)s_C^2}{n_A + n_C - 2} = 190977 = 437^2.$$

Question VII.3 (20)

Which of the following statements is TRUE?

1 ☐ The estimated difference $\hat{\mu}_A - \hat{\mu}_C$ is -32.4 grams, which is a significant difference, since it is smaller than 0.05% of the average body mass of both Adelie and Chinstrap penguins.

2 ☐ The relevant test statistic is approximately $t = -0.07$, and the final conclusion is that there is a significant difference in body mass between the Adelie and the Chinstrap species, since $|-0.07| = 0.07$ is greater than 0.05.

3* ☐ The relevant test statistic is approximately $t = -0.5$, and the final conclusion is that we do NOT find a statistically significant difference in mean body mass between the Adelie and the Chinstrap species.

4 ☐ The relevant test statistic is approximately $t = -0.5$, but given the information in the exercise, we are not able to determine if the difference is significant or not.

5 ☐ The relevant test statistic is approximately $t = -0.07$, and the final conclusion is that we do NOT find a significant difference in body mass between the Adelie and the Chinstrap species, since $|-0.07| = 0.07$ is greater than 0.05.

6 ☐ Don't know / No answer

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

For **Adelie**:

$$n_A = 151, \quad s_A = 458.6, \quad \bar{x}_A = 3700.7$$

For **Chinstrap**:

$$n_C = 68, \quad s_C = 384.3, \quad \bar{x}_C = 3733.1$$

And we already have:

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_C - 1)s_C^2}{n_A + n_C - 2} = 190977 = 437^2$$

$$s_p = 437$$

Therefore:

$$t_{pooled} = \frac{\bar{x}_A - \bar{x}_C}{\sqrt{s_p^2\left(\frac{1}{n_A} + \frac{1}{n_C}\right)}} = \frac{3700.7 - 3733.1}{\sqrt{437^2\left(\frac{1}{68} + \frac{1}{151}\right)}} \approx -0.5$$

Also, the degrees of freedom are:

$$n_A + n_C - 2 = 217$$

From the quantiles given in the assignment, we do not see the exact quantile we need (which is $t_{0.975,\nu=217}$), but we can see that the critical value will be between $1.980(t_{0.975,\nu=120})$ and 1.996 $(t_{0.975,\nu=67})$. Therefore, the value $|-0.5| = 0.5$ is so small that we will not be able to reject the null hypothesis.

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

Continue on page 30

Question VII.4 (21)

Which statement is TRUE regarding the pooled two-sample t-test?

1* ☐ The method assumes that the two populations have equal variances.

2 ☐ The method requires equal sample sizes in the two groups.

3 ☐ The method is inappropriate because the sample sizes are larger than 30.

4 ☐ The method allows the two populations to have different variances and adjusts the degrees of freedom accordingly.

5 ☐ The method tests whether the two sample variances are equal.

6 ☐ Don't know / No answer

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

The pooled method assumes the samples come from population with equal variances.

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

The model can be written as a general linear model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}),$$

where $\boldsymbol{Y}$ is the collection of all resposes (first those from Adelie, then those from Chinstrap). The design matrix is now chosen as

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1}_{151} & \frac{68}{151+68}\mathbf{1}_{151} \\ \mathbf{1}_{68} & -\frac{151}{151+68}\mathbf{1}_{68} \end{bmatrix},$$

where $\mathbf{1}_n$ is a column vector of ones with n-elements, and $\boldsymbol{\beta} = [\beta_1, \beta_2]$.

Question VII.5 (22)

Let $\hat{\sigma}^2$ denote estimate of σ^2. What are the standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ (denoted se_{β_1} and se_{β_2})?

1* ☐ $se_{\beta_1} = \frac{\hat{\sigma}}{\sqrt{68+151}}$ and $se_{\beta_2} = \hat{\sigma}\sqrt{\frac{68+151}{68\cdot 151}}$

2 ☐ $se_{\beta_1} = \frac{\hat{\sigma}}{\sqrt{151}}$ and $se_{\beta_2} = \frac{\hat{\sigma}}{\sqrt{68}}$

3 ☐ $se_{\beta_1} = \hat{\sigma}\sqrt{\frac{68}{151}}$ and $se_{\beta_2} = \hat{\sigma}\sqrt{\frac{151}{68}}$

4 ☐ $se_{\beta_1} = \hat{\sigma}\sqrt{\frac{68}{68+151}}$ and $se_{\beta_2} = \hat{\sigma}\sqrt{\frac{151}{151+68}}$

5 ☐ $se_{\beta_1} = \hat{\sigma}\sqrt{\frac{151}{68+151}}$ and $se_{\beta_2} = \hat{\sigma}\sqrt{\frac{68}{151+68}}$

6 ☐ Don't know / No answer

`---------------------------------- FACIT-BEGIN ----------------------------------`

We need to find the diagonal elements of

$$\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{26}$$

and we start by the matrix product

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} \mathbf{1}_{151}^T & \mathbf{1}_{68}^T \\ \frac{68}{151+68}\mathbf{1}_{151}^T & -\frac{151}{151+68}\mathbf{1}_{68}^T \end{bmatrix} \begin{bmatrix} \mathbf{1}_{151} & \frac{68}{151+68}\mathbf{1}_{151} \\ \mathbf{1}_{68} & -\frac{151}{151+68}\mathbf{1}_{68} \end{bmatrix} \tag{27}$$

$$= \begin{bmatrix} 151+68 & \frac{68}{151+68}\cdot 151 - \frac{151}{151+68}68 \\ \frac{68}{151+68}\cdot 151 - \frac{151}{151+68}\cdot 68 & \left(\frac{68}{151+68}\right)^2\cdot 151 + \left(\frac{151}{151+68}\right)^2\cdot 68 \end{bmatrix} \tag{28}$$

$$= \begin{bmatrix} 151+68 & 0 \\ 0 & \frac{68\cdot 151}{(151+68)^2}(151+68) \end{bmatrix} \tag{29}$$

$$= \begin{bmatrix} 151+68 & 0 \\ 0 & \frac{68\cdot 151}{151+68} \end{bmatrix} \tag{30}$$

taking the square root of the doagonal elements of the inverse we get answer 1.

`---------------------------------- FACIT-END ----------------------------------`

Continue on page 32

Exercise VIII

A transportation researcher is studying whether the adoption rate of electric cars (EVs) increased after a government introduced incentives such as free municipal parking and expanded charging station infrastructure.

In a small pilot survey conducted before the incentives, it was found that about 15% of households owned an electric car. Since the pilot study was relatively small ($n = 20$ households), the estimate of 15% has considerable uncertainty.

To plan a larger follow-up study, the researcher wishes to estimate the true proportion of households owning an electric car after the incentives. The researcher wishes to construct a 95% confidence interval with a *Margin of Error* (ME) of only 6%.

Question VIII.1 (23)

What minimum sample size n is required for the follow-up study, assuming the true proportion is equal the proportion found in the pilot study?

1 ☐ $n = 120$

2* ☐ $n = 137$

3 ☐ $n = 210$

4 ☐ $n = 260$

5 ☐ $n = 310$

6 ☐ Don't know / No answer

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

The margin of error is

$$ME = z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Solving for n:

$$n = p(1-p)\left(\frac{z_{1-\alpha/2}}{ME}\right)^2$$

Insert the known values:

$$n = 0.15(1 - 0.15)\left(\frac{1.96}{0.06}\right)^2 = 0.1275(32.67)^2 = 0.1275 \times 1067.11 = 136.1$$

Thus, the researcher should plan for a sample size of 137 households.

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

The researcher later conducts a survey of 300 households. With the result

- Before the incentives: 36 out of 150 households owned an electric car.
- After the incentives: 51 out of 150 households owned an electric car.

The researcher wants to test whether the proportion of EV ownership changed after the incentives using a significance level $\alpha = 0.05$.

Continue on page 34

Question VIII.2 (24)

Which of the following conclusions is correct?

1* ☐ $z_{obs} = 1.91 < 1.96$, so we fail to reject H_0 and CANNOT conclude that EV adoption increased significantly.

2 ☐ $z_{obs} = 1.45 < 1.96$, so we fail to reject H_0 and conclude that EV adoption did not increase significantly.

3 ☐ $z_{obs} = 0.87 < 1.96$, so we fail to reject H_0 and conclude that EV adoption did not increase significantly.

4 ☐ $z_{obs} = 1.91 < 1.96$, so we fail to reject H_0 and conclude that EV adoption increased significantly.

5 ☐ $z_{obs} = 3.02 > 1.96$, so we reject H_0 and conclude that EV adoption increased significantly.

6 ☐ Don't know / No answer

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

We estimate the proportions:

$$\hat{p}_1 = \frac{36}{150} = 0.24$$

$$\hat{p}_2 = \frac{51}{150} = 0.34$$

Under $H_0 : p_1 = p_2 = p$ the pooled estimate is

$$\hat{p} = \frac{36 + 51}{300} = \frac{87}{300} = 0.29$$

The test statistic is

$$z_{obs} = \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{150} + \frac{1}{150}\right)}} = \frac{0.34 - 0.24}{\sqrt{0.29(0.71)(0.0133)}} \frac{0.10}{0.0463} = \frac{0.10}{0.0524} = 1.91$$

Since

$$|z_{obs}| = 1.91 < 1.96$$

we **fail to reject the null hypothesis** at the significance level $\alpha = 0.05$.

Therefore, there is not sufficient statistical evidence to conclude that the proportion of households owning an electric car changed after the incentives.

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

Continue on page 36

Exercise IX

A researcher assesses the toxicity of various drugs against a cardiovascular disease using a one-way ANOVA model, which yields the following entries in the ANOVA table:

Source	DF	SS
Drug	3	33.66
Residual	36	220.20
Total	39	253.86

Question IX.1 (25)

Given the model assumptions are satisfied, what is the observed F-statistic for the null hypothesis that the mean toxicity (mean effect) is the same for all the drugs against the usual alternative hypothesis?

1 ☐ $F_{\text{obs}} = 11.22$

2 ☐ $F_{\text{obs}} = 6.12$

3* ☐ $F_{\text{obs}} = 1.83$

4 ☐ $F_{\text{obs}} = 0.55$

5 ☐ $F_{\text{obs}} = 0.15$

6 ☐ Don't know / No answer

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

Theorem 8.6 states that the F-statistic is

$$F_{\text{obs}} = \frac{SS(Tr)/(k-1)}{SSE/(n-k)} = \frac{33.66/3}{220.20/36} = 1.83.$$

The values in the denominators are the degrees of freedom for the treatments and the residuals, and these are found in the ANOVA table under DF.

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

Question IX.2 (26)

The observed F-statistic should be compared to a quantile from which distribution when testing the null hypothesis that the mean toxicity (mean effect) is the same for all the drugs against the usual alternative hypothesis?

1* ☐ An F-distribution with $\nu_1 = 3$ and $\nu_2 = 36$ degrees of freedom.

2 ☐ An F-distribution with $\nu_1 = 4$ and $\nu_2 = 39$ degrees of freedom.

3 ☐ An F-distribution with $\nu_1 = 36$ and $\nu_2 = 39$ degrees of freedom.

4 ☐ A t-distribution with $\nu = 36$ degrees of freedom.

5 ☐ A t-distribution with $\nu = 39$ degrees of freedom.

6 ☐ Don't know / No answer

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

Theorem 8.6 states that the F-statistic follows an F-distribution with $\nu_1 = k - 1 = 3$ and $\nu_2 = n - k = 36$. These values are the degrees of freedom for the treatments and the residuals, and they are found in the ANOVA table under DF.

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

Question IX.3 (27)

What is the estimated standard deviation of the residuals in the one-way ANOVA model?

1 ☐ 220.20

2 ☐ 14.84

3 ☐ 6.12

4 ☐ 3.35

5* ☐ 2.47

6 ☐ Don't know / No answer

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

In ANOVA models, the residual variance is estimated by MSE, cf. page 314. Hence,

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-k} = \frac{220.20}{36}.$$

Consequently, the estimated residual standard deviation is given as

$$\hat{\sigma} = \sqrt{\frac{220.20}{36}} = 2.47.$$

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

Question IX.4 (28)

The Bonferroni correction is used to correct the significance level (α) when multiple tests are conducted. Which of the following statements about the effect on the risk of Type I and Type II errors is correct for the Bonferroni correction?

1 ☐ The Bonferroni correction increases the Type I risk but decreases the Type II risk of the individual hypothesis tests.

2* ☐ The Bonferroni correction decreases the Type I risk but increases the Type II risk of the individual hypothesis tests.

3 ☐ The Bonferroni correction increases the Type I risk and decreases the power of the individual hypothesis tests.

4 ☐ The Bonferroni correction decreases the Type I risk and increases the power of the individual hypothesis tests.

5 ☐ The Bonferroni correction increases the Type I risk of the individual hypothesis tests but decreases the family-wise Type I risk.

6 ☐ Don't know / No answer

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

The idea of applying a Bonferroni correction is to reduce the family-wise Type I error rate by reducing the Type I risk in each of the individual tests. However, for a fixed experimental design, decreasing the Type I risk implies increasing the Type II risk i.e., decreasing the power.

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

Continue on page 39

Exercise X

Consider the standard normal random variable $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}_2)$ and the matrices

$$\boldsymbol{A}_1 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}; \quad \boldsymbol{A}_2 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}; \quad \boldsymbol{A}_3 = \begin{bmatrix} \sqrt{2} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

and define the random variables $\boldsymbol{Y}_i = \boldsymbol{A}_i \boldsymbol{Z}_i$.

Question X.1 (29)

Which of the following statements, on the variance-covariance matrices ($V[\boldsymbol{Y}_i]$), is true?

1 ☐ $V[\boldsymbol{Y}_i] \neq V[\boldsymbol{Y}_j]$ for $i \neq j$.

2* ☐ $V[\boldsymbol{Y}_1] = V[\boldsymbol{Y}_2] = V[\boldsymbol{Y}_3]$.

3 ☐ $V[\boldsymbol{Y}_1] = V[\boldsymbol{Y}_2]$, but $V[\boldsymbol{Y}_1] \neq V[\boldsymbol{Y}_3]$

4 ☐ $V[\boldsymbol{Y}_1] = V[\boldsymbol{Y}_3]$, but $V(\boldsymbol{Y}_1] \neq V[\boldsymbol{Y}_2]$

5 ☐ $V[\boldsymbol{Y}_2] = V[\boldsymbol{Y}_3]$, but $V[\boldsymbol{Y}_1] \neq V[\boldsymbol{Y}_2]$

6 ☐ Don't know / No answer

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

The variance-covariance matrix is given by $V[\boldsymbol{Y}_i] = V[\boldsymbol{A}_i \boldsymbol{Z}] = \boldsymbol{A}_i V[\boldsymbol{Z}] \boldsymbol{A}_i^T = \boldsymbol{A}_i \boldsymbol{A}_i^T$. Hence we get

$$V[\boldsymbol{Y}_1] = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \tag{31}$$

$$V[\boldsymbol{Y}_2] = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \tag{32}$$

$$V[\boldsymbol{Y}_2] = \begin{bmatrix} \sqrt{2} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{2} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \tag{33}$$

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

Now consider $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}_4)$ and the orthogonal projection matrix

$$\boldsymbol{H} = \frac{1}{4} \begin{bmatrix} 3 & 1 & 1 & 1 \\ 1 & 3 & -1 & -1 \\ 1 & -1 & 3 & -1 \\ 1 & -1 & -1 & 3 \end{bmatrix}.$$

Question X.2 (30)

What is $E[\boldsymbol{Z}^T \boldsymbol{H} \boldsymbol{Z}]$?

1 ☐ 1

2 ☐ 2

3 ☐ 4

4 ☐ 6

5* ☐ 3

6 ☐ Don't know / No answer

`------------------------------------ FACIT-BEGIN ------------------------------------`

Since $\boldsymbol{H}$ is an orthogonal projection matrix then $\boldsymbol{Z}^T \boldsymbol{H} \boldsymbol{Z} \sim \chi^2(trace(\boldsymbol{H}))$, and since $trace(\boldsymbol{H}) = \frac{1}{4} \cdot 4 \cdot 3 = 3$, it imply that $\boldsymbol{Z}^T \boldsymbol{H} \boldsymbol{Z} \sim \chi^2(3)$, and hence $E(\boldsymbol{Z}^T \boldsymbol{H} \boldsymbol{Z}) = 3$.

`------------------------------------- FACIT-END -------------------------------------`

The exam is finished.