English assigment follows the Danish version

Exam question paper for:

Written examination: 20. December 2025

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

Duration: 4 hours

Aids allowed: All printed aids plus pocket calculator model TI30XS or TI30XB

The final answers should be handed in by filling out a separate "Answer Sheet".

This exam consists of 30 questions of the "multiple choice" type, which are divided between 16 exercises.
Only hand in the "Answer Sheet" and not the entire question paper.

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 off your own result to the number of decimals given in the answer options before you choose your answer.*

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

```
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

You draw many random samples, each of size $n = 150$, from a population that is skewed to the right (meaning the distribution has a tail extending far to the right). The distribution has mean 40 and a standard deviation 10.

Question I.1 (1)

According to the Central Limit Theorem (CLT), which of the following statements about the distribution of the sample means is correct?

1 ☐ The distribution of the sample means will be highly skewed to the right.

2 ☐ The mean of the sample means will be much greater than 40 because of the skewness.

3 ☐ The distribution of the sample means will be approximately normally distributed and centered around 40.

4 ☐ The standard deviation of the sample means will be 10, same as the population.

5 ☐ The distribution of the sample means will become uniform as the sample size increases.

6 ☐ Don't know / No answer

Continue on page 4

Exercise II

The discrete random variable X has the following distribution:

x	1	3	5	6
$f(x)$	0.1	0.4	0.3	0.2

where $f(x) = P(X = x)$ is the probability density function (pdf).

Question II.1 (2)

What is the mean $\mathbf{E}[X]$ and variance $\mathbf{V}[X]$?

1 ☐ $\mathbf{E}[X] = 3.75$ and $\mathbf{V}[X] = 1.92$

2 ☐ $\mathbf{E}[X] = 4.00$ and $\mathbf{V}[X] = 1.92$

3 ☐ $\mathbf{E}[X] = 4.40$ and $\mathbf{V}[X] = 2.40$

4 ☐ $\mathbf{E}[X] = 3.75$ and $\mathbf{V}[X] = 2.40$

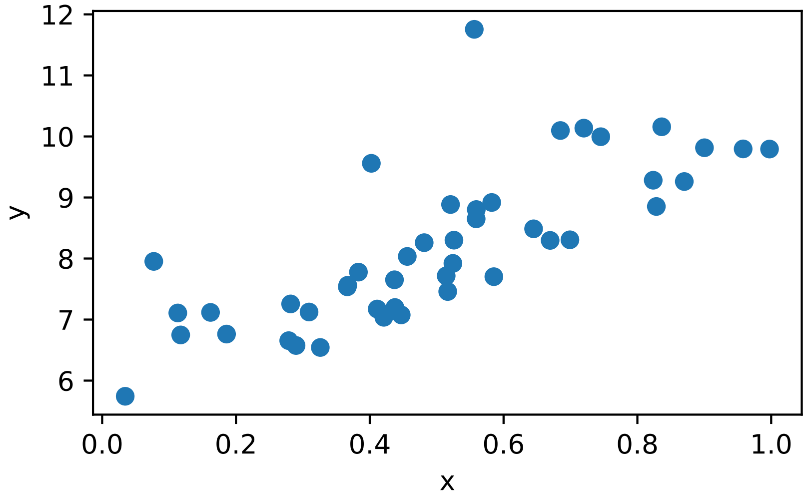
5 ☐ $\mathbf{E}[X] = 4.00$ and $\mathbf{V}[X] = 2.40$

6 ☐ Don't know / No answer

Continue on page 5

Exercise III

Some data has been obtained for which we simply call the observed values "y" and "x". The data is visualized in the scatter plot below.



The following information about the data is provided:

$$\bar{x} = 0.5024$$

$$\bar{y} = 8.1964$$

$$Sxx = \sum_{i}^{n} (x_i - \bar{x})^2 = 2.5365$$

$$Sxy = \sum_{i}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 10.2153$$

Where n is the number of observations in the data.

The data was stored in Python in a DataFrame called "dat", containing the columns "x" and "y". A simple linear regression model was fitted to the data using the following command in Python:

```
fit = smf.ols(formula = 'y ~ x', data = dat).fit()
```

The resulting regression table is printed below (although certain values have been substituted by the letters A, B, C and D):

```
print(fit.summary(slim=True))

                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.583
Model:                            OLS   Adj. R-squared:                  0.573
No. Observations:                  45   F-statistic:                     60.12
Covariance Type:            nonrobust   Prob (F-statistic):           1.06e-09
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.1731      0.289     21.388      0.000       5.591       6.755
x                   A          B      7.754      0.000           C           D
==============================================================================
```

Question III.1 (3)

Consider the statistical model that was fitted to the data and the values represented by the letters A and B (inserted in the regression table above).

Which of the following statements is correct?

1 ☐ $A = \hat{\beta}_1$ is the estimate of the parameter β_1 in the statistical model:
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $\varepsilon_i \sim N(0, \sigma^2)$.
$B = \hat{\sigma}_{\hat{\beta}_1}$ is the estimated standard error of $\hat{\beta}_1$.

2 ☐ $A = \hat{\beta}_0$ is the estimate of the parameter β_0 in the statistical model:
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $\varepsilon_i \sim N(0, \sigma^2)$.
$B = \hat{\sigma}_{\hat{\beta}_0}$ is the estimated standard error of $\hat{\beta}_0$.

3 ☐ $A = \hat{\beta}_1$ is the estimate of the parameter β_1 in the statistical model:
$y_i = \beta_1 x_i + \varepsilon_i$, with $\varepsilon_i \sim N(0, \sigma^2)$.
$B = \hat{\sigma}_{\hat{\beta}_1}$ is the estimated standard error of $\hat{\beta}_1$.

4 ☐ $A = \hat{\beta}_0$ is the estimate of the parameter β_0 and $B = \hat{\beta}_1$ is the estimate of the parameter β_1 in the statistical model:
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $\varepsilon_i \sim N(0, \sigma^2)$.

5 ☐ $A = \hat{\beta}_1$ is the estimate of the parameter β_1 and $B = \hat{\beta}_0$ is the estimate of the parameter β_0 in the statistical model:
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $\varepsilon_i \sim N(0, \sigma^2)$.

6 ☐ Don't know / No answer

Question III.2 (4)

What is the value of A?

1 ☐ $A = 0.5024$

2 ☐ $A = 10.2153$

3 ☐ $A = 5.10765$

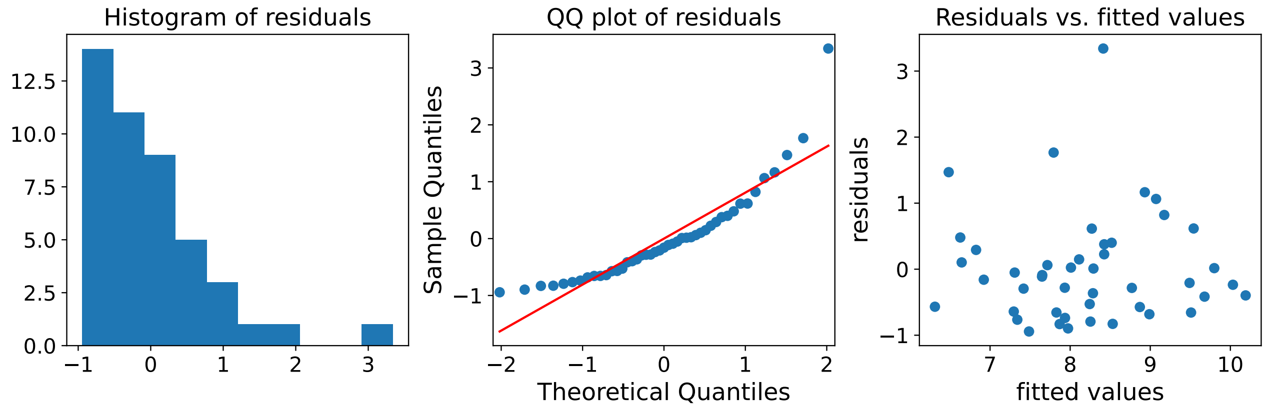
4 ☐ $A = 45/10 = 4.5$

5 ☐ $A = 4.0273$

6 ☐ Don't know / No answer

Question III.3 (5)

In order to check the model assumptions the following plots were produced:



Which of the following statements is correct?

1 ☐ The histogram and qq-plot of the residuals indicate that the assumption about normality ($\varepsilon_i \sim N(0, \sigma^2)$) is violated - i.e. the residuals do not seems to follow a normal distribution.

2 ☐ The qq-plot of the residuals indicate that the assumption about independence is violated - i.e. the residuals do not seems to be independent.

3 ☐ The plots do not indicate a violation of the model assumptions.

4 ☐ The scatter plot of the residuals vs. fitted values indicate that the assumption about independence is violated - i.e. the residuals do not seems to be independent.

5 ☐ The qq-plot of the residuals indicate that the residuals follow a normal distribution with zero mean: $\varepsilon_i \sim N(0, \sigma^2)$.

6 ☐ Don't know / No answer

Continue on page 8

Exercise IV

A group of researchers study cell activity in four different species of mice. They collect samples with sample sizes: $n_1 = 50$, $n_2 = 150$, $n_3 = 150$, $n_4 = 50$ for species 1, 2, 3 and 4 respectively. The researchers fit a mathematical model of the form:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \quad i \in \{1, 2, 3, 4\},$$

where the errors ε_{ij} are assumed to be independent.

The researchers compute that $SS(Tr) = 6.0479$ (here "Tr" relates to the different species) and $SST = 163.234$, yielding a p-value of 0.0018 when testing the null hypothesis:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$

Question IV.1 (6)

Which of the following statements give the correct value of the test statistic in the F-test, and also states a correct conclusion?

1 ☐ The value of the test statistic is $F = 5.079$. The null hypothesis cannot be rejected at significance level $\alpha = 0.01$.

2 ☐ The value of the test statistic is $F = 5.079$. The null hypothesis cannot be rejected at significance level $\alpha = 0.001$.

3 ☐ The value of the test statistic is $F = 2.96$. The null hypothesis is rejected at significance level $\alpha = 0.01$.

4 ☐ The value of the test statistic is $F = 2.96$. The null hypothesis is rejected at significance level $\alpha = 0.001$.

5 ☐ The value of the test statistic is $F = 0.99$. The null hypothesis is rejected at significance level $\alpha = 0.05$.

6 ☐ Don't know / No answer

Question IV.2 (7)

The projection matrix associated with the model is

$$\boldsymbol{H} = \begin{bmatrix} \frac{1}{n_1}\boldsymbol{E}_{n_1,n_1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \frac{1}{n_2}\boldsymbol{E}_{n_2,n_2} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \frac{1}{n_3}\boldsymbol{E}_{n_3,n_3} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \frac{1}{n_4}\boldsymbol{E}_{n_4,n_4} \end{bmatrix},$$

where $\boldsymbol{E}_{n_i,n_i}$ is a $n_i \times n_i$ matrix of ones. Under the null-hypothesis $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, then the appropriate projection matrix is

$$\boldsymbol{H}_0 = \frac{1}{n}\boldsymbol{E}_{n,n},$$

where $n = n_1 + n_2 + n_3 + n_4$.

Given that the null-hypothesis is true, what is the the distribution of

$$Q = \frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H}_0)\boldsymbol{Y},$$

where σ^2 is the true (but unknown) variance of ε_i?

1 ☐ $Q \sim F(4, n-3)$.

2 ☐ $Q \sim \chi^2(3)$.

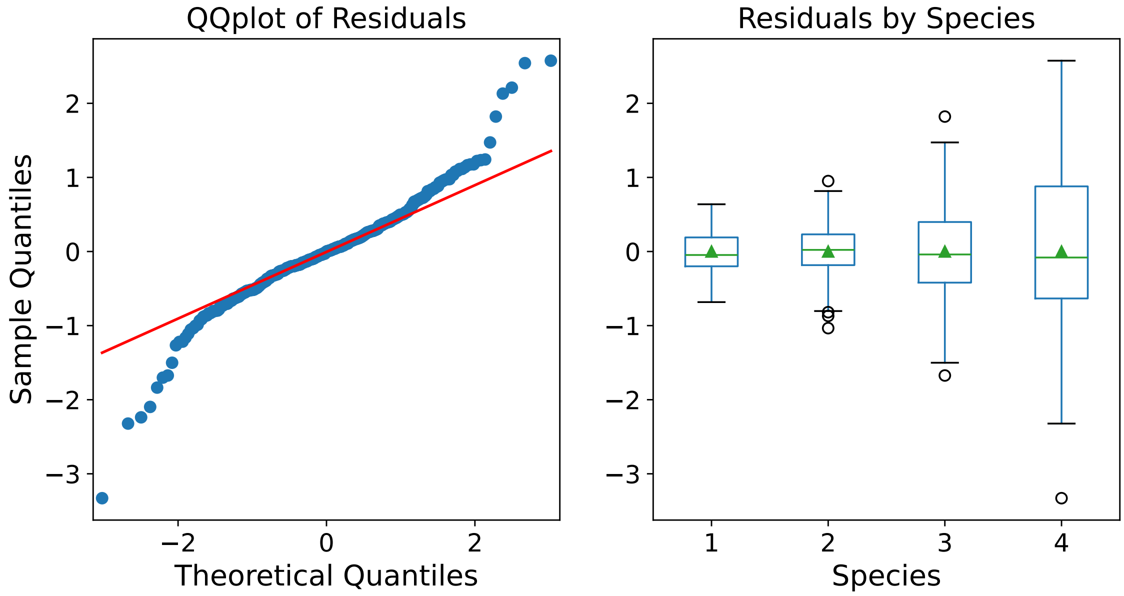
3 ☐ $Q \sim \chi^2(n-4)$.

4 ☐ $Q \sim F(4, n)$.

5 ☐ $Q \sim F(3, n-4)$.

Question IV.3 (8)

The researchers proceed to perform model validation (i.e., model checking). They produce the diagnostic plots presented below:



Which of the following statements is correct (all arguments must be true)?

1 ☐ The QQ-plot indicates a violation of the assumption about normally distributed parameters. The box plots indicate an incorrect estimation of the model parameters.

2 ☐ The QQ-plot indicates a violation of the assumption about normally distributed residuals. The box plots indicate a violation of the assumption of equal residual variance across species.

3 ☐ The QQ-plot indicates a violation of the assumption about normally distributed residuals. The box plots indicate a violation of the assumption that $\mu_{\varepsilon_{ij}} = 0$.

4 ☐ The QQ-plot indicates a violation of the assumption about equal sample size within each group. The box plots indicate a disproportionate number of outliers in the data.

5 ☐ The diagnostic plots do not indicate any violation of model assumptions.

6 ☐ Don't know / No answer

Continue on page 12

Exercise V

A researcher claims that the average daily screen time for educational use by university students is 6 hours. To test this claim, a random sample of $n = 20$ students was selected, giving: $\bar{x} = 5.4$ hours and $s = 1.2$ hours. Assume screen time is approximately normally distributed. A hypothesis test for the null hypothesis $H_0 : \mu = 6$ is conducted and the computed p-value is 0.03.

Question V.1 (9)

Which of the following statements is correct?

1 ☐ Using a significance level, $\alpha = 0.05$, the null hypothesis $H_0 : \mu = 6$ is rejected. The researcher concludes that the mean screen time for educational use is not 6 hours.

2 ☐ Using a significance level, $\alpha = 0.05$, the alternative hypothesis $H_A : \mu = 0$ is accepted. The researcher concludes that the mean screen time for educational use is significantly less than 6 hours.

3 ☐ Using a significance level, $\alpha = 0.01$, the null hypothesis $H_0 : \mu = 6$ is rejected. The researcher concludes that the mean screen time for educational use is not 6 hours.

4 ☐ Using a significance level, $\alpha = 0.01$, the alternative hypothesis $H_A : \mu = 0$ is accepted. The researcher concludes that the mean screen time for educational use is significantly less than 6 hours.

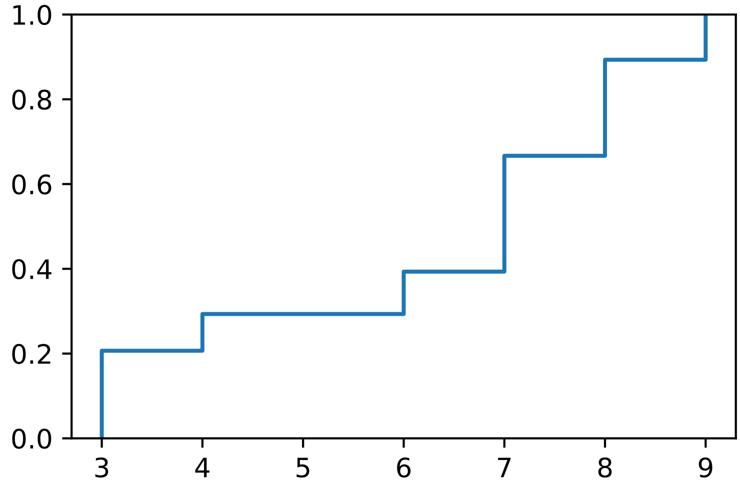
5 ☐ There is not enough information to make a decision about H_0.

6 ☐ Don't know / No answer

Continue on page 13

Exercise VI

150 observations of a discrete stochastic variable are simulated in Python and the resulting values are visualized in the following empirical cumulative distribution (ecdf) plot:



Question VI.1 (10)

Which of the following Python codes could generate the simulated observations?

1 ☐ `np.random.choice(a=[3,4,5,6,7,8,9], size=150, p=[1/7,1/7,1/7,1/7,1/7,1/7,1/7])`

2 ☐ `stats.uniform.rvs(loc=3, scale=6, size=150)`

3 ☐ `stats.uniform.rvs(loc=3, scale=9, size=100)`

4 ☐ `np.random.choice(a=[3,4,6,7,8,9], size=150, p=[2/10,1/10,1/10,3/10,2/10,1/10])`

5 ☐ `stats.norm.rvs(loc=6, scale=2, size=150)`

6 ☐ Don't know / No answer

Continue on page 14

Exercise VII

A researcher is studying whether students' active participation in lectures is related to higher course satisfaction. In a small pilot study with $n = 15$ students, it was found that about 60% of students, who frequently participated in lectures, rated their overall course satisfaction as "high". Since the pilot study is relatively small, the estimate of 60% is related to a high degree of uncertainty. To plan a larger study for the next semester, the researcher wants to estimate the true proportion of "high course satisfaction", among students who frequently participate in lectures. The researcher would like to give a 95% confidence interval for this proportion, with the *Margin of Error* (ME) being only 0.05.

To answer the following questions you may need the quantile from the standard normal distribution: $z_{0.975} = 1.96$.

Question VII.1 (11)

What minimum sample size n is required for the follow-up study?

1 ☐ $n = 78$

2 ☐ $n = 185$

3 ☐ $n = 240$

4 ☐ $n = 369$

5 ☐ $n = 412$

6 ☐ Don't know / No answer

Question VII.2 (12)

If no prior estimate of p were available, what is then the required sample size?

1 ☐ $n = 138$

2 ☐ $n = 185$

3 ☐ $n = 240$

4 ☐ $n = 385$

5 ☐ $n = 420$

6 ☐ Don't know / No answer

Question VII.3 (13)

Disregarding the results from the two previous questions, the researcher decides to conduct a study using a random sample of 200 students. Among the 200 students 50 were classified as "Active participants" and 150 were classified as "Not active participants". The researcher wants to compare the proportion of "high course satisfaction" between the two groups and obtains the following survey results:

Active Participants: 30 out of 50 reported "high course satisfaction"
Not active Participants: 75 out of 150 reported "high course satisfaction"

The researcher conducts a two sample proportions hypothesis test, comparing the proportion of students reporting "high course satisfaction" within the two groups. The researcher uses a significance level of $\alpha = 0.05$.

Which of the following conclusions is correct (both the calculation and the conclusion must be correct)?

1 ☐ The calculated $z_{obs} = 1.23 < 1.96$, so we fail to reject the null hypothesis of equal proportions within the two groups ($H_0 : p_1 = p_2$)

2 ☐ The calculated $z_{obs} = 2.01 > 1.96$, so we reject the null hypothesis of equal proportions within the two groups ($H_0 : p_1 = p_2$) and conclude that there is a significant difference between the two groups.

3 ☐ The calculated $z_{obs} = 1.19 < 1.96$, so we fail to reject the null hypothesis of equal proportions within the two groups ($H_0 : p_1 = p_2$)

4 ☐ The calculated $z_{obs} = 1.23 < 1.96$, so we reject the null hypothesis of equal proportions within the two groups ($H_0 : p_1 = p_2$) and conclude that there is a significant difference between the two groups.

5 ☐ The calculated $z_{obs} = 2.01 > 1.96$, so we fail to reject the null hypothesis of equal proportions within the two groups ($H_0 : p_1 = p_2$).

6 ☐ Don't know / No answer

Continue on page 16

Question VII.4 (14)

In another study the researcher wants to investigate if students' lecture participation level is associated with their exam results. This time the student participation level is classified as "High", "Moderate", or "Low" and the exam results are classified as "High", "Medium", or "Low". In a random sample of 200 students the participation level and exam results are recorded and presented in the table below:

	Exam Result:			
Lecture Participation:	High	Medium	Low	Total
High	30	15	5	50
Moderate	25	30	15	70
Low	10	20	50	80
Total	65	65	70	200

The researcher wants to test whether the distribution of exam results is independent from lecture participation level.

The desired significance level α has been stored in Python in a variable called "`alpha`". Which of the following Python statements correctly calculates the critical value for the relevant hypothesis test?

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

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

Continue on page 17

Exercise VIII

Consider a multiple linear regression model written in the matrix formulation:

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

for which the design matrix $\mathbf{X}$ is as follows (here we only show the top rows of the matrix):

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & (x_1)^3 \\ 1 & x_2 & (x_2)^3 \\ 1 & x_3 & (x_3)^3 \\ 1 & x_4 & (x_4)^3 \\ 1 & x_5 & (x_5)^3 \\ 1 & x_6 & (x_6)^3 \\ 1 & x_7 & (x_7)^3 \\ 1 & x_8 & (x_8)^3 \\ 1 & x_9 & (x_9)^3 \\ 1 & x_{10} & (x_{10})^3 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

Question VIII.1 (15)

How can the same regression model be written (in non-matrix formulation)?

1 ☐ $y_i = \beta_0 + \beta_a a_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

2 ☐ $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

3 ☐ $y_i = \beta_0 + \beta_1 x_i + 2 \cdot \beta_2 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

4 ☐ $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^3 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

5 ☐ $y_i = \beta_0 + \beta_1 (x_i + x_i^3), \quad \varepsilon_i \sim N(0, \sigma^2)$

6 ☐ Don't know / No answer

Question VIII.2 (16)

The model parameters are estimated using the formula (in matrix formulation):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Which statement is correct:

1 ☐ The vector $\hat{\boldsymbol{\beta}}$ has four elements $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}^2)$ and the values of these parameters are chosen such that the sum of squared residuals $(\sum_{i=1}^{n} \varepsilon_i^2)$ is minimized.

2 ☐ The vector $\hat{\boldsymbol{\beta}}$ has three elements $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2)$ and the values of these parameters are chosen such that the total sum of squares $(SST = \sum_{i=1}^{n}(y_i - \bar{y})^2)$ is minimized.

3 ☐ The vector $\hat{\boldsymbol{\beta}}$ has three elements $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2)$ and the values of these parameters are chosen such that the sum of squared residuals $(\sum_{i=1}^{n} \varepsilon_i^2)$ is minimized.

4 ☐ The vector $\hat{\boldsymbol{\beta}}$ has three elements $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2)$ and the values of these parameters are chosen such that the variance of x-values (s_x^2) is maximized.

5 ☐ The vector $\hat{\boldsymbol{\beta}}$ has two elements $(\hat{\beta}_0,\ \hat{\beta}_1)$ and the values of these parameters are chosen such that $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is minimized.

6 ☐ Don't know / No answer

Question VIII.3 (17)

Let n denote the number of observations in the data set, and further let $\hat{\sigma}^2$ denote the usual unbiased (or central) variance estimator for the regression model. The following number is now calculated

$$Q = \mathbf{y}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}$$

What is the value of Q?

1 ☐ $(n-1)\hat{\sigma}^2$

2 ☐ $(n-3)\hat{\sigma}^2$

3 ☐ $(n-2)\hat{\sigma}^2$

4 ☐ $\hat{\sigma}^2$

5 ☐ $n\hat{\sigma}^2$

6 ☐ Don't know / No answer

Question VIII.4 (18)

Now another model is considered

$$\mathbf{Y} = \mathbf{X}_2\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$$

The matrix $\boldsymbol{X}_2$ is constructed such that

$$\boldsymbol{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}_2(\mathbf{X}_2^T\mathbf{X}_2)^{-1}\mathbf{X}_2^T = \boldsymbol{H}_2$$

Which of the following statements is (in general) true?

1 ☐ The parameter estimates ($\hat{\boldsymbol{\beta}}$) are equal in the two case, but the fitted values ($\hat{\boldsymbol{Y}}$) might differ.

2 ☐ The parameter estimates ($\hat{\boldsymbol{\beta}}$) and the fittede values ($\hat{\boldsymbol{Y}}$) are equal in the two cases.

3 ☐ The fitted values ($\hat{\boldsymbol{Y}}$) are the same for the two models, but the variance estimate ($\hat{\sigma}^2$) might differ.

4 ☐ The fitted values ($\hat{\boldsymbol{Y}}$) are the same for the two models, but the parameters estimated ($\hat{\boldsymbol{\beta}}$) might differ.

5 ☐ The desing matrices are equal ($\boldsymbol{X} = \boldsymbol{X}_2$) and hence all values (parameter estimates, fitted values, and variance) agree for the two models.

6 ☐ Don't know / No answer

Continue on page 20

Exercise IX

A researcher collects a random sample of size $n = 10$ from a normally distributed population. The sample standard deviation is $s = 4.2$.

To answer this question you may need the following quantiles from the χ^2-distribution with ν degrees of freedom:

$\chi^2_{0.025}(\nu = 9) = 2.70, \quad \chi^2_{0.975}(\nu = 9) = 19.02$

Question IX.1 (19)

Find the 95% confidence interval for the population standard deviation σ.

1 ☐ [2.93, 7.25]

2 ☐ [2.89, 7.67]

3 ☐ [3.10, 7.38]

4 ☐ [2.85, 7.80]

5 ☐ [3.40, 8.20]

6 ☐ Don't know / No answer

Continue on page 21

Exercise X

A fitness coach wants to test whether a new 4-week training program has a significant effect on resting heart rate. She measures the resting heart rates (in beats per minute) of 8 participants before and after the program.

	Resting heart rate:							
Before program:	78	85	90	76	88	82	79	84
After program:	74	80	85	72	83	78	75	80

Assume the data follow normal distributions and the test is conducted at significance level $\alpha = 0.05$.

Question X.1 (20)

Assume the data has been read into Python in Variables named `Before` and `After`.
Which Python command should be used to correctly test whether the program significantly changed the mean resting heart rate?

1 ☐ `stats.ttest_1samp(After, popmean=75)`

2 ☐ `stats.ttest_ind(Before, After, equal_var=False)`

3 ☐ `stats.ttest_1samp(Before - After, popmean=Before.mean()-After.mean())`

4 ☐ `stats.ttest_1samp(Before - After, popmean=0)`

5 ☐ `stats.ttest_ind(Before, After, equal_var=True)`

6 ☐ Don't know / No answer

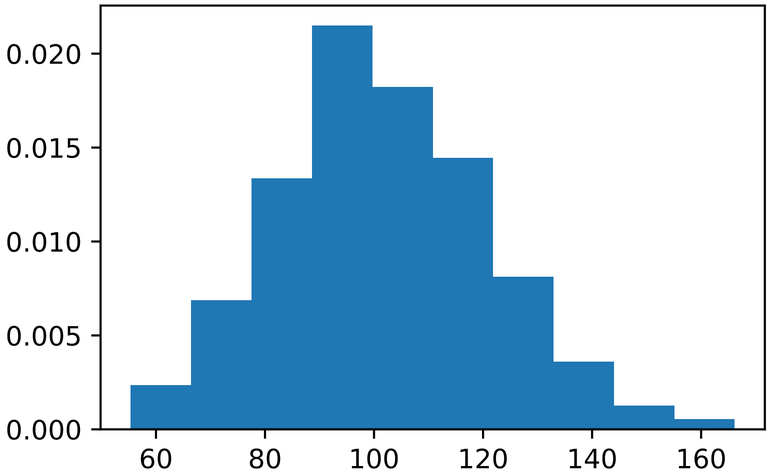
Continue on page 22

Exercise XI

A simulation has been carried out with the following Python code:

```
A = stats.uniform.rvs(size=500, loc=0, scale = 5)
B = stats.norm.rvs(size=500, loc=10, scale = 1)
C = A + B**2

plt.hist(C, density=True)
plt.show()
```



Question XI.1 (21)

Which distributions do the stochastic variables A and B follow?

1 ☐ A follows a normal distribution with mean $\mu_A = 0$ and standard deviation $\sigma_A = 5$.
B follows a normal distribution with mean $\mu_B = 10$ and standard deviation $\sigma_B = 1$.

2 ☐ A follows a uniform distribution with mean $\mu_A = 0$ and standard deviation $\sigma_A = 5$.
B follows a normal distribution with mean $\mu_B = 10$ and standard deviation $\sigma_B = 1$.

3 ☐ A is uniformly distributed between $\alpha_A = 0$ and $\beta_A = 5$.
B follows a normal distribution with mean $\mu_B = 10$ and standard deviation $\sigma_B = 1$.

4 ☐ A is uniformly distributed between $\alpha_A = -5$ and $\beta_A = 5$.
B follows a normal distribution with mean $\mu_B = 10$ and standard deviation $\sigma_B = 1$.

5 ☐ The distributions of A and B cannot be determined from the Python code.

6 ☐ Don't know / No answer

Question XI.2 (22)

From the simulation we see that $C = A + B^2$.

Say we perform a new simulation for which $\mu_A = 2.5$, $\sigma_A = 5/\sqrt{12}$, $\mu_B = 10$ and $\sigma_B = 1$.

Use error propagation to estimate σ_C. Which of the following is correct?

1 ☐ $\sigma_C = \sqrt{5/\sqrt{12} + 20}$

2 ☐ $\sigma_C = 5/\sqrt{12} + 200$

3 ☐ $\sigma_C = \sqrt{25/12 + 400}$

4 ☐ $\sigma_C = \sqrt{25 + 400}$

5 ☐ $\sigma_C = \sqrt{25/12}$

6 ☐ Don't know / No answer

Continue on page 24

Exercise XII

A researcher is planning an experiment to estimate the mean reduction in blood pressure when given a specific drug. From a pilot study, the estimated standard deviation in blood pressure is $\sigma = 8$ mmHg. The researcher wants to detect a mean *effect size* (difference in average blood pressure) of $\mu_0 - \mu_1 = 5$ mmHg with a significance level of $\alpha = 0.05$ and a power of $1-\beta = 0.80$. The researcher plans to conduct the experiment on a sample of n test persons, where each test person has their blood pressure measured both with and without the drug.

To solve this exercise you may need the following quantiles from a standard normal distribution:

$z_{1-\beta} = z_{0.80} = 0.84$ and $z_{1-\alpha/2} = z_{0.975} = 1.96$.

Question XII.1 (23)

What is the required sample size needed for the experiment, given the information above?

1 ☐ $n = 40$

2 ☐ $n = 20$

3 ☐ $n = 41$

4 ☐ $n = 5$

5 ☐ $n = 21$

6 ☐ Don't know / No answer

Continue on page 25

Exercise XIII

To gain insight into the adoption of generative AI tools (such as ChatGPT or Copilot) for learning support, a university research team surveyed a pilot group of $n = 50$ students. Among these, $x = 30$ students report using such AI tools regularly to assist with study tasks such as writing, coding, or revising concepts.

To solve this exercise you may need one of the following quantiles:

From a standard normal distribution: $z_{0.95} = 1.64$ and $z_{0.975} = 1.96$.

From a t-distribution (with ν degrees of freedom): $t_{0.95}(\nu = 49) = 1.68$ and $t_{0.975}(\nu = 49) = 2.01$.

Question XIII.1 (24)

Which of the following is the correct 95%-confidence interval (given that one uses the method described in the book) for the true proportion (p) of students who regularly use generative AI tools for learning, and what are the assumptions behind the calculation (both confidence interval and argument must be true)?

1 ☐ The 95%-confidence interval for the true proportion is $[0.46; 0.74]$. The calculation is based on the assumption that the sample size (n) is large enough for the sample proportion ($\hat{p}$) to be approximately normally distributed.

2 ☐ The 95%-confidence interval for the true proportion is $[0.46; 0.74]$. The calculation is based on the assumption that the sample size (n) follows a Binomial distribution.

3 ☐ The 95%-confidence interval for the true proportion is $[0.48,\ 0.72]$. The calculation is based on the assumption that the sample size (n) is large enough for the Central Limit Theorem to be valid.

4 ☐ The 95%-confidence interval for the true proportion is $[0.48,\ 0.72]$. The calculation is based on the assumption that the sample size (n) is large enough for the t-distribution to be well approximated by a standard normal distribution.

5 ☐ The 95%-confidence interval for the true proportion is $[0.48,\ 0.72]$. The calculation is based on the assumption that the sample size (n) follows a Binomial distribution.

6 ☐ Don't know / No answer

Continue on page 26

Exercise XIV

An insurance company receives claims according to a Poisson process with an intensity of two claims per month. This entails that claims arrive independently and that the waiting time between two consecutive claims follows an exponential distribution with a mean of half a month. Consequently, the number of claims the insurance company receives during a given month follows a Poisson distribution with a mean of two claims per month.

Question XIV.1 (25)

What is the probability that the insurance company receives fewer than two claims during a given month?

1 ☐ 13.5%

2 ☐ 27.1%

3 ☐ 30.3%

4 ☐ 40.6%

5 ☐ 50.5%

6 ☐ Don't know / No answer

Question XIV.2 (26)

Which expression correctly calculates the probability that the waiting time between two consecutive claims exceeds one month?

1 ☐ $1 - \exp(-\frac{1}{2})$

2 ☐ $\int_0^1 2\exp(-2x)dx$

3 ☐ $\int_1^\infty 2\exp(-2x)dx$

4 ☐ $\int_1^\infty \frac{1}{2}\exp(-\frac{1}{2}x)dx$

5 ☐ $\int_0^1 \frac{1}{2}\exp(-\frac{1}{2}x)dx$

6 ☐ Don't know / No answer

Continue on page 27

Exercise XV

Consider the two-way ANOVA model with four different "Treatments" and five different "Blocks":

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \quad i \in \{1, ..., 4\}, \quad j \in \{1, ...5\},$$

where the errors ε_{ij} are assumed to be independent, i refers to the four different treatments and j refers to the five different blocks.

A data set has been obtained, and the following averages have been calculated:

Overall	31.5	Block 1	31.0
Treatment 1	35.1	Block 2	32.0
Treatment 2	29.4	Block 3	32.3
Treatment 3	28.7	Block 4	30.3
Treatment 4	32.8	Block 5	31.9

Furthermore, the treatment sum of squares, $SS(Tr)$, has been calculated to be 134.5 (here "Tr" refers to the treatments).

Moreover, let $\hat{\sigma}$ denote the model estimate of σ.

Question XV.1 (27)

What are the covariances $\text{Cov}(Y_{12}, Y_{13})$ and $\text{Cov}(Y_{22}, Y_{42})$ according to the model?

1 ☐ $\text{Cov}(Y_{12}, Y_{13}) = 0$ and $\text{Cov}(Y_{22}, Y_{42}) = 0$

2 ☐ $\text{Cov}(Y_{12}, Y_{13}) = 0$ and $\text{Cov}(Y_{22}, Y_{42}) = \beta_2$

3 ☐ $\text{Cov}(Y_{12}, Y_{13}) = \alpha_1$ and $\text{Cov}(Y_{22}, Y_{42}) = 0$

4 ☐ $\text{Cov}(Y_{12}, Y_{13}) = \alpha_1$ and $\text{Cov}(Y_{22}, Y_{42}) = \beta_2$

5 ☐ $\text{Cov}(Y_{12}, Y_{13}) = \mu + \alpha_1$ and $\text{Cov}(Y_{22}, Y_{42}) = \mu + \beta_2$

6 ☐ Don't know / No answer

Question XV.2 (28)

Testing the null hypothesis $H_0 : \alpha_i = 0$ for $i = 1, ..., 4$ leads to which value of the F-statistic?

1 ☐ $F = \frac{44.83}{\hat{\sigma}^2}$

2 ☐ $F = \frac{44.83}{\hat{\sigma}}$

3 ☐ $F = \frac{134.5}{\hat{\sigma}^2}$

4 ☐ $F = \frac{538}{\hat{\sigma}}$

5 ☐ $F = \frac{538}{\hat{\sigma}^2}$

6 ☐ Don't know / No answer

Continue on page 29

Exercise XVI

Let X be a standard normal distributed random variable, and $Y_1 \sim \chi^2(5)$ and $Y_2 \sim \chi^2(5)$. It is assumed that X, Y_1 and Y_2 are independent.

Question XVI.1 (29)

What is the mean of

$$Y = \frac{Y_1}{X^2 + Y_2} \quad ?$$

1 ☐ $E[Y] = \frac{5}{4}$

2 ☐ $E[Y] = \frac{6}{4}$

3 ☐ $E[Y] = \frac{6}{4}$

4 ☐ $E[Y] = \frac{5}{6}$

5 ☐ $E[Y] = \frac{5}{3}$

Question XVI.2 (30)

What is a, if the following must hold

$$P\left(\frac{X}{\sqrt{Y_1}} < a\right) = 0.95 \quad ?$$

1 ☐ $a = t_{0.95}(5)\sqrt{4}$

2 ☐ $a = \frac{t_{0.95}(5)}{\sqrt{5}}$

3 ☐ $a = F_{0.95}(1,5)\sqrt{5}$

4 ☐ $a = t_{0.95}(4)\sqrt{5}$

5 ☐ $a = \frac{F_{0.95}(1,4)}{\sqrt{4}}$

The exam is finished.