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

---

# Exam paper:

*Written examination*: 28.05.2026

*Course name and number*: **02402 Statistics (Polytechnical Foundation)**

*Duration*: 4 hours

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

---

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

This examination consists of 30 multiple-choice questions distributed across 11 exercises.
**Only the "Answer Sheet" must be submitted; do not hand in the exam paper itself.**

**Multiple choice questions:** *Note that in each question, one and only one of the answer options is correct. Furthermore, not all suggested answer options are necessarily meaningful. When performing calculations, always round your result to the number of decimal places used in the answer options before selecting your answer.*

**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 3

**Exercise I**

Let $X$ and $Y$ be independent random variables. The expected values are $\mathbb{E}[X] = 5$ and $\mathbb{E}[Y] = -3$, while the variances are $\mathbb{V}[X] = 9$ and $\mathbb{V}[Y] = 16$.

**Question I.1 (1)**

What is the covariance between $X$ and $Y$?

1 ☐ -25

2 ☐ -5

3 ☐ 0

4 ☐ 5

5 ☐ 25

6 ☐ Don't know / No answer

**Question I.2 (2)**

Let $Z = -4X + 3Y$. What is the expectation of $Z$?

1 ☐ -29

2 ☐ -27

3 ☐ 0

4 ☐ 27

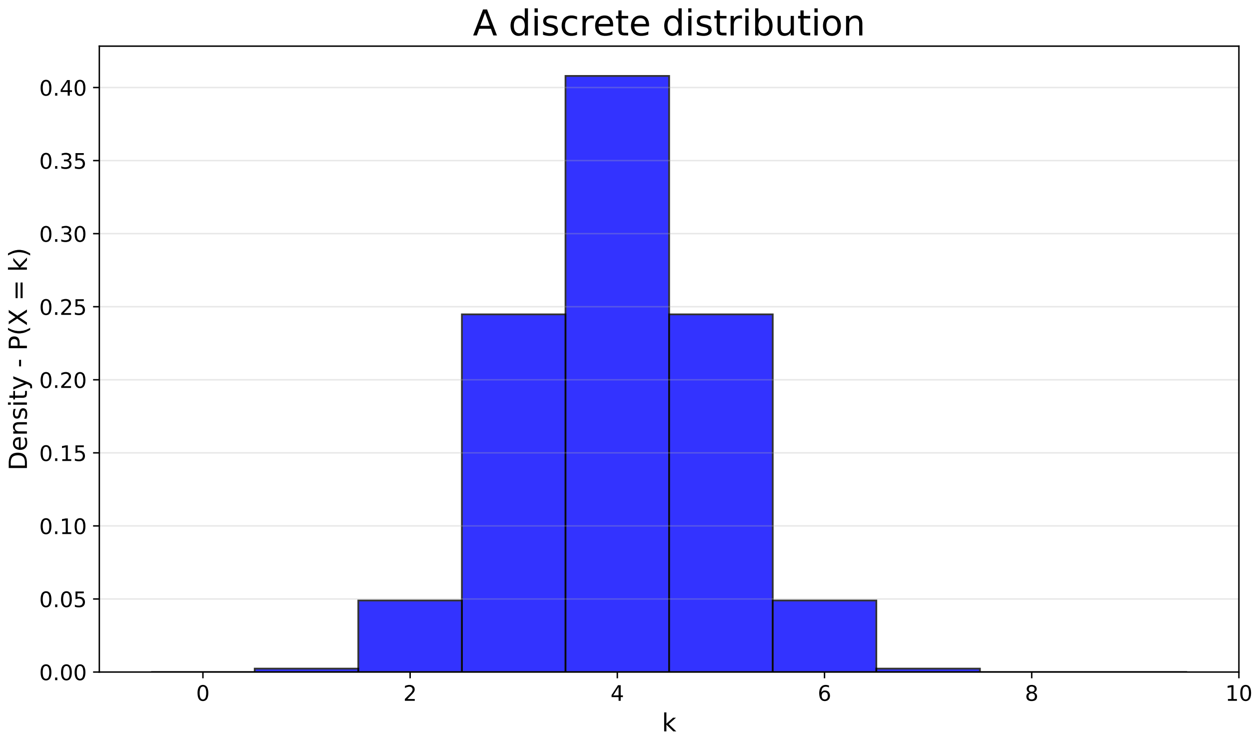
5 ☐ 29

6 ☐ Don't know / No answer

Continue on page 4

## Exercise II

Consider the following discrete distribution on the non-negative integers, represented by its probability density function (pdf), and let $X$ follow this distribution.




### Question II.1 (3)

Which of the following options correctly declares the distribution?

1 ☐ It is a binomial distribution with $n = 10$ and $p = 0.5$.

2 ☐ It is a binomial distribution with $n = 10$ and $p = 0.9$.

3 ☐ It is a hypergeometric distribution with $n = 7$, $a = 8$, and $N = 14$.

4 ☐ It is a Poisson distribution with $\lambda = 4$.

5 ☐ It is a Poisson distribution with $\lambda = 10$.

6 ☐ Don't know / No answer

Continue on page 5

**Question II.2 (4)**

What is $\mathbb{P}(X < 5)$?

1 ☐ Approximately 0.25

2 ☐ Approximately 0.30

3 ☐ Approximately 0.40

4 ☐ Approximately 0.70

5 ☐ Approximately 0.95

6 ☐ Don't know / No answer

**Question II.3 (5)**

What is the variance of $X$?

1 ☐ Approximately 0.2

2 ☐ Approximately 0.7

3 ☐ Approximately 0.9

4 ☐ Approximately 4.2

5 ☐ Approximately 17.4

6 ☐ Don't know / No answer

Continue on page 6

**Exercise III**

In a study of sentencing practices, researchers collected data on prison sentences (measured in years) imposed by different judges for comparable criminal offenses. Let $Y_{ij}$ denote sentence $j$ given by judge $i$. The data are modeled as $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, where $\mu$ is the overall average sentence across all judges, $\alpha_i$ is the difference between the average sentence of judge $i$ and the overall average, and $\varepsilon_{ij}$ is a random error term. The error terms are assumed to be independent and identically distributed with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, and all sentences are treated as independent observations.

**Question III.1 (6)**

Which statement is correct based on the model described above?

1 ☐ The model allows the error terms to follow any continuous distribution with mean zero.

2 ☐ The model allows judges to have different mean sentences and different sentence variances.

3 ☐ The model allows judges to have different mean sentences but not different sentence variances.

4 ☐ The model allows judges to have different sentence variances but not different mean sentences.

5 ☐ The model does not allow judges to have different sentence variances nor different mean sentences.

6 ☐ Don't know / No answer

**Question III.2 (7)**

The researchers expect the judges' average sentences to differ substantially, but each judge to give very consistent sentences (each judge tends to give very similar sentences across cases). Assuming the model assumptions are satisfied, which findings would support the researchers' expectations?

1 ☐ A small value of SS(judges) and a small value of SSE.

2 ☐ A small value of SS(judges) but a large value of SSE.

3 ☐ A large value of SS(judges) but a small value of SSE.

4 ☐ A large value of SS(judges) and a large value of SSE.

5 ☐ No findings can support the researchers' expectations if the model assumptions are satisfied.

6 ☐ Don't know / No answer

The model is fitted to the data included in the study, which are shown in the table below:

| Judge A | Judge B | Judge C | Judge D | Judge E |
|---|---|---|---|---|
| 4.3 | 7.8 | 10.1 | 4.7 | 13.2 |
| 14.1 | 8.4 | 10.5 | 10.3 | 17.9 |
| 6.9 | 10.2 | 10.4 | 8.1 | 12.4 |
| 20.5 | 11.3 | 9.9 | 7.6 | 12.5 |
| 2.2 | 9.3 | 10.1 | | 10.1 |
| 10.8 | | 11.0 | | 19.8 |
| | | 10.2 | | 26.5 |
| | | 12.6 | | 11.1 |

**Question III.3 (8)**

Which statement about the model assumptions is correct based on the data? (Hint: Try to visualise the data as boxplots.)

1 ☐ The model assumptions are satisfied.

2 ☐ The model assumptions are not satisfied because the data is not normally distributed.

3 ☐ The model assumptions are not satisfied because the study includes fewer than five sentences from one of the judges.

4 ☐ The model assumptions are not satisfied because the judges exhibit considerably different sentence variances.

5 ☐ The model assumptions are not satisfied because the judges have contributed different numbers of sentences.

6 ☐ Don't know / No answer

Continue on page 8

## Exercise IV

Consider a simple linear regression model

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

under the usual independence assumptions.

### Question IV.1 (9)

Which one of the following statements is false?

1 ☐ The model has three parameters.

2 ☐ The random variables $\varepsilon_i$ are called errors.

3 ☐ The normality assumption is checked with a normal QQ-plot of the residuals.

4 ☐ The 95% confidence interval for the slope can be wider than the 95% confidence interval for the intercept.

5 ☐ The 95% confidence interval for the mean response at a given $x_0$ can be wider than the 95% prediction interval for an individual response at $x_0$.

6 ☐ Don't know / No answer

### Question IV.2 (10)

Consider a case where the model has been fitted to a dataset with $n = 4$ observations:

| ID ($i$) | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Observation ($y_i$) | 4 | 6 | 8 | 10 |
| Model prediction ($\hat{y}_i$) | 4 | 5 | 10 | 9 |
| Residual ($e_i$) | 0 | 1 | -2 | 1 |

What is the residual sum of squares (RSS) for the model?

1 ☐ The residual sum of squares is -2.

2 ☐ The residual sum of squares is 0.

3 ☐ The residual sum of squares is 1.

4 ☐ The residual sum of squares is 4.

5 ☐ The residual sum of squares is 6.

6 ☐ Don't know / No answer

Continue on page 9

**Exercise V**

Consider a multiple curvilinear regression model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, and the first row of the design matrix $\boldsymbol{X}$ is given by

$$[\cos(x_{11}) \quad \cos(x_{21}) \quad x_{11} + x_{21} \quad x_{11} \cdot x_{21}].$$

**Question V.1 (11)**

What is the model equation for an individual observation?

1 ☐ $Y_i = \beta_0 + \beta_1 \cos(x_{1i}) + \beta_2 \cos(x_{2i}) + \beta_3 x_{1i} + \beta_4 x_{2i} + \beta_5 x_{1i} \cdot \beta_6 x_{2i} + \varepsilon_i$

2 ☐ $Y_i = \beta_0 + \beta_1 \cos(x_{1i}) + \beta_2 \cos(x_{2i}) + \beta_3 (x_{1i} + x_{2i}) + \beta_4 (x_{1i} \cdot x_{2i}) + \varepsilon_i$

3 ☐ $Y_i = \beta_0 + \beta_1 \cos(x_{1i}) + \beta_2 \cos(x_{2i}) + \beta_3 x_{1i} + \beta_4 x_{2i} + \beta_5 (x_{1i} \cdot x_{2i}) + \varepsilon_i$

4 ☐ $Y_i = \beta_1 \cos(x_{1i}) + \beta_2 \cos(x_{2i}) + \beta_3 x_{1i} + \beta_4 x_{2i} + \beta_5 (x_{1i} \cdot x_{2i})$

5 ☐ $Y_i = \beta_1 \cos(x_{1i}) + \beta_2 \cos(x_{2i}) + \beta_3 (x_{1i} + x_{2i}) + \beta_4 (x_{1i} \cdot x_{2i}) + \varepsilon_i$

6 ☐ Don't know / No answer

**Question V.2 (12)**

How is the vector of residuals, $\boldsymbol{r}$, calculated?

1 ☐ $\boldsymbol{r} = \boldsymbol{\epsilon}$

2 ☐ $\boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

3 ☐ $\boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$

4 ☐ $\boldsymbol{r} = \widehat{\boldsymbol{Y}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

5 ☐ $\boldsymbol{r} = \boldsymbol{Y} - (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})$

6 ☐ Don't know / No answer

Continue on page 10

Model selection yields that the most appropriate model is

$$Y_i = \beta_1 \cos(x_{1i}) + \beta_2(x_{1i} \cdot x_{2i}) + \varepsilon_i,$$

where the errors, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, are assumed to be independent. The model is fitted to a dataset with 102 observations using ordinary least squares estimation, which produces the following parameter estimates:

$$\hat{\beta}_1 = 24, \quad \hat{\beta}_2 = 15, \quad \hat{\sigma} = 4.$$

**Question V.3 (13)**

What is the residual sum of squares, $\text{RSS}(\hat{\beta}_1, \hat{\beta}_2)$, for the fitted model?

1 ☐ 1600

2 ☐ 1584

3 ☐ 801

4 ☐ 400

5 ☐ 396

6 ☐ Don't know / No answer

Continue on page 11

**Exercise VI**

Each week, the Danish state lottery draws 7 numbers at random without replacement from the numbers 1 to 36 (both included). A lottery ticket consists of 7 distinct numbers, and to win the jackpot, all 7 numbers on your ticket must match the drawn numbers. (Note: The order of the numbers on a ticket does not matter.)

**Question VI.1 (14)**

What is the chance (probability) of winning the jackpot with a single ticket in the lottery?

1 ☐ 1 in 78.364.164.096

2 ☐ 1 in 42.072.307.200

3 ☐ 1 in 2.176.782.336

4 ☐ 1 in 8.347.680

5 ☐ 1 in 5.040

6 ☐ Don't know / No answer

**Exercise VII**

Consider two independent random variables $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 \sim \mathcal{N}(0, 1)$. A theoretical model posits that $Y = f(X_1, X_2) = X_1^2 + X_2^2$.

**Question VII.1 (15)**

According to the non-linear error propagation rule, what is the variance of $Y$?

1 ☐ The variance of $Y$ is 0.

2 ☐ The variance of $Y$ is 2.

3 ☐ The variance of $Y$ is 4.

4 ☐ The variance of $Y$ is 8.

5 ☐ The variance of $Y$ is 16.

6 ☐ Don't know / No answer

Continue on page 12

**Exercise VIII**

A university investigates the effects of generative AI platforms on academic performance.

**Question VIII.1 (16)**

In a large class, a sample of 15 students who used the recommended platform had an average exam score of 82 points with a standard deviation of 6.2 points. Assuming normality of the exam scores, what is the 95% confidence interval for the mean exam score of students using the recommended platform? (You can use that $t_{0.975}(14) = 2.145$.)

1 ☐ $(78.6, 85.4)$

2 ☐ $(79.7, 84.3)$

3 ☐ $(80.0, 84.0)$

4 ☐ $(81.0, 83.0)$

5 ☐ $(77.5, 86.5)$

6 ☐ Don't know / No answer

**Question VIII.2 (17)**

In another large class, a sample of 12 students who used an alternative platform had a mean score of 79 points (SD was 5 points) on the exam. The university applies a two-sided $t$-test to compare the mean score of the students using the alternative platform with the historical class mean score of 75 points. The test yields a $p$-value of 0.018 and a 95% confidence interval of $(75.82, 82.18)$. What is the appropriate conclusion of the test at the 5% significance level?

1 ☐ The mean score of students using the alternative platform is significantly lower than the historical class mean score.

2 ☐ The mean score of students using the alternative platform is not significantly higher than the historical class mean score.

3 ☐ The mean score of students using the alternative platform is not significantly different from the historical class mean score.

4 ☐ The mean score of students using the alternative platform is significantly different from the historical class mean score.

5 ☐ There is a 1.8% probability that the mean score of students using the alternative platform is significantly higher than the historical class mean score.

6 ☐ Don't know / No answer

The university has developed an AI tutor designed to help students study and prepare for lectures. To examine whether students spend less time on lecture preparation when using the AI tutor compared with ordinary preparation methods, the university has followed two groups of students attending the same lectures: a small group of 50 students using the AI tutor and a larger group of 300 students using traditional preparation methods. For each lecture, the university has recorded the average preparation time for both groups:

| Lecture | 1 | 2 | 3 | 4 | $\cdots$ | 13 |
|---|---|---|---|---|---|---|
| Average preparation time with AI tutor (min.) | 30 | 45 | 80 | 110 | $\cdots$ | 95 |
| Average preparation time without AI tutor (min.) | 30 | 45 | 100 | 120 | $\cdots$ | 100 |

The university notes that the expected preparation time varies substantially across lectures (as evident from the table) and that the unequal group sizes may lead to different group variances.

**Question VIII.3 (18)**

Considering the experimental design, which one of the following options represents the most appropriate test for assessing a mean difference in preparation time between the two groups?

1 ☐ A Welch two-sample $t$-test

2 ☐ A pooled two-sample $t$-test

3 ☐ A paired $t$-test

4 ☐ A $\chi^2$-test

5 ☐ A one-way ANOVA test

6 ☐ Don't know / No answer

Continue on page 14

**Question VIII.4 (19)**

The university eventually decides to use a different type of test and selects a significance level of 5%. For the given parameters, the test has a statistical power of 90%. Assuming that the model assumptions are satisfied and that there truly is a difference in mean preparation time between the two groups, what is the probability that the test reaches the correct conclusion?

1 ☐ 95%

2 ☐ 90%

3 ☐ 10%

4 ☐ 5%

5 ☐ 2.5%

6 ☐ Don't know / No answer

**Question VIII.5 (20)**

In the alternative test, you calculate adjusted preparation times (these remain strictly positive), which the test requires to be approximately normally distributed. However, a histogram of the adjusted preparation times reveals that they are right-skewed. Which of the following transformations should not be considered for making the adjusted preparation times approximately normally distributed?

1 ☐ The cube root transformation i.e., transform $x$ into $x^{1/3}$.

2 ☐ The exponential transformation i.e., transform $x$ into $\exp(x)$.

3 ☐ The logarithmic transformation i.e., transform $x$ into $\log(x)$.

4 ☐ The reciprocal transformation i.e., transform $x$ into $1/x$.

5 ☐ The square root transformation i.e., transform $x$ into $\sqrt{x}$.

6 ☐ Don't know / No answer

Continue on page 15

**Exercise IX**

A university wants to investigate whether the type of digital training technology affects students' exam scores. 240 students were randomly assigned to one of three digital training technologies: Video-based Learning (VBL), Gamified Learning Platform (GLP), and Interactive Simulations (IS). Their exam scores were categorized into three levels: Below Average, Average, and Above Average.

| Exam score | VBL | GLP | IS | Row Total |
|---|---|---|---|---|
| Below Average | 18 | 12 | 10 | 40 |
| Average | 32 | 26 | 22 | 80 |
| Above Average | 30 | 38 | 52 | 120 |
| Column Total | 80 | 76 | 84 | 240 |

Students with Average or Above Average exam scores are considered as "successful learners". The below table of quantiles from the standard normal distribution is needed to solve some of the problems in this exercise.

| Quantile | $q_{0.90}$ | $q_{0.95}$ | $q_{0.975}$ | $q_{0.99}$ |
|---|---|---|---|---|
| Value | 1.282 | 1.645 | 1.960 | 2.326 |

**Question IX.1 (21)**

Under the null hypothesis that the distribution of exam scores is the same across all three technologies, what is the expected number of students using Video-based Learning (VBL) with a Below Average exam score?

1 ☐ 12.90

2 ☐ 13.33

3 ☐ 14.00

4 ☐ 15.12

5 ☐ 18.50

6 ☐ Don't know / No answer

Continue on page 16

**Question IX.2 (22)**

What is the 95% confidence interval for the overall proportion of successful learners based on the data?

1 ☐ $[0.786, 0.880]$

2 ☐ $[0.701, 0.812]$

3 ☐ $[0.692, 0.833]$

4 ☐ $[0.688, 0.784]$

5 ☐ $[0.622, 0.754]$

6 ☐ Don't know / No answer

**Question IX.3 (23)**

Is there a significant difference between the proportions of successful learners among students using VBL and students using IS at the 5% significance level? (Hint: Calculate the test statistic under $H_0 : p_{VBL} - p_{IS} = 0$, where $p_{VBL}$ is the proportion of students using VBL who are considered successful learners, and $p_{IS}$ is the corresponding proportion for students using IS.)

1 ☐ There is no significant difference, as the observed test statistic $z_{\text{obs}} = -1.56 > -1.96$.

2 ☐ There is no significant difference, as the observed test statistic $|z_{\text{obs}}| = |-2.12| > 1.96$.

3 ☐ There is no significant difference, as the observed test statistic $z_{\text{obs}} = -1.80 > -1.96$.

4 ☐ There is a significant difference, as the observed test statistic $z_{\text{obs}} = -2.27 < -1.96$.

5 ☐ There is a significant difference, as the observed test statistic $z_{\text{obs}} = -2.13 < -1.96$.

6 ☐ Don't know / No answer

Continue on page 17

**Question IX.4 (24)**

What is the 95% confidence interval for the difference in the proportions of Above Average outcomes for students in the IS and GLP groups?

1 ☐ $[0.113, 0.279]$

2 ☐ $[0.105, 0.293]$

3 ☐ $[0.091, 0.286]$

4 ☐ $[0.082, 0.312]$

5 ☐ $[-0.034, 0.272]$

6 ☐ Don't know / No answer

**Question IX.5 (25)**

In the usual test of independence between the type of digital training technology assigned to students and their exam score, which distribution does the test statistic follow under the null hypothesis of independence?

1 ☐ An $F$-distribution with 3 and 3 degrees of freedom

2 ☐ An $F$-distribution with 2 and 2 degrees of freedom

3 ☐ A $\chi^2$-distribution with 9 degrees of freedom

4 ☐ A $\chi^2$-distribution with 6 degrees of freedom

5 ☐ A $\chi^2$-distribution with 4 degrees of freedom

6 ☐ Don't know / No answer

Continue on page 18

**Exercise X**

In a two-way ANOVA model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

where $\alpha_i$ represents the treatment effect and $\beta_j$ the block effect, an analysis is conducted to assess the significance of both factors.

**Question X.1 (26)**

Which of the following statements is false?

1 ☐ The blocks are included to reduce the residual variation.

2 ☐ The treatment effect and the block effect are tested using separate $F$-tests based on different degrees of freedom.

3 ☐ The errors in the model are assumed to be normally distributed and independent.

4 ☐ The blocks are included to increase the number of treatment groups.

5 ☐ The total variation in the data can be decomposed into treatment, block, and residual components.

6 ☐ Don't know / No answer

Consider the ANOVA table from the two-way ANOVA analysis:

| **Source** | **Df** | **F value** | **Pr(>F)** |
|---|---|---|---|
| Treatment | 2 | 74.40 | 0.0001 |
| Block | 3 | 6.37 | 0.0270 |
| Residuals | 6 | – | – |

Continue on page 19

**Question X.2 (27)**

Which of the following interpretations is most appropriate?

1 ☐ Both the treatment and block effects are statistically significant at the 1% level.

2 ☐ Both the treatment and block effects are statistically significant at the 5% level.

3 ☐ The block effect is not statistically significant at the 5% level.

4 ☐ The residuals are statistically significant at the 5% level, indicating model failure.

5 ☐ The treatment effect is not statistically significant at the 1% level.

6 ☐ Don't know / No answer

Continue on page 20

**Exercise XI**

In classical mechanics, the motion of an object under constant acceleration is governed by the equation

$$v(t) = v_0 + at,$$

where $v(t)$ is the velocity at time $t$, $v_0$ is the initial velocity (velocity at time $t = 0$), and $a$ is the constant acceleration.

In an experiment, a set of velocity measurements is recorded at different time points for an object moving under constant acceleration. Due to measurement errors, the observed velocities deviate from the theoretical model. To account for this, we introduce an error term, leading to the linear regression model:

$$v_i = v_0 + at_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where the errors are assumed to be independent.

The following output is given:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      v   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
No. Observations:                 100   F-statistic:                 3.071e+05
Covariance Type:            nonrobust   Prob (F-statistic):          3.89e-173
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.4695      0.421     10.629      0.000       3.635       5.304
t              4.0064      0.007    554.190      0.000       3.992       4.021
==============================================================================
```

**Question XI.1 (28)**

What are the estimated parameter values?

1 ☐ $\hat{v}_0 = 554.190$ and $\hat{a} = 10.629$

2 ☐ $\hat{v}_0 = 10.629$ and $\hat{a} = 554.190$

3 ☐ $\hat{v}_0 = 0.421$ and $\hat{a} = 0.007$

4 ☐ $\hat{v}_0 = 4.470$ and $\hat{a} = 4.006$

5 ☐ $\hat{v}_0 = 4.006$ and $\hat{a} = 4.470$

6 ☐ Don't know / No answer

Continue on page 21

The following numbers can be used in the next question:

```
print(np.round(stats.t.ppf(q = [0.995,0.99,0.975,0.95,0.90], df = 100-2),4))

[2.6269 2.365  1.9845 1.6606 1.2902]
```

**Question XI.2 (29)**

What is the 99% confidence interval for $v_0$?

1 ☐ $[3.364, 5.575]$

2 ☐ $[3.474, 5.465]$

3 ☐ $[3.635, 5.304]$

4 ☐ $[3.990, 4.023]$

5 ☐ $[3.992, 4.021]$

6 ☐ Don't know / No answer

**Question XI.3 (30)**

Consider the null hypothesis $\mathcal{H}_0 : v_0 = 5$ against a two-sided alternative hypothesis $\mathcal{H}_1 : v_0 \neq 5$. What is the observed test statistic ($t_{\text{obs}}$) under the null hypothesis?

1 ☐ -2.360

2 ☐ -1.260

3 ☐ 1.260

4 ☐ 10.629

5 ☐ 554.190

6 ☐ Don't know / No answer

The exam is finished.