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: **02323 Introduction to Statistics**

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

This exercise is concerned with basic Python functionality.

Question I.1 (1)

What does the command `np.arange(10)` (`np` refers to the NumPy package) output?

1 ☐ The array `[1 2 3 4 5 6 7 8 9 10]`.

2 ☐ The array `[1 2 3 4 5 6 7 8 9]`.

3* ☐ The array `[0 1 2 3 4 5 6 7 8 9]`.

4 ☐ The array `[0 1 2 3 4 5 6 7 8 9 10]`.

5 ☐ An array with ten random numbers between zero and one.

6 ☐ Don't know / No answer

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

Correct answer is 3. NumPy provides efficient functions for computing descriptive statistics such as mean, variance, standard deviation, and percentiles.

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

Question I.2 (2)

Which function in the `scipy.stats`-package outputs random numbers?

1 ☐ The `cdf`-function

2 ☐ The `pdf`-function

3 ☐ The `pmf`-function

4 ☐ The `ppf`-function

5* ☐ The `rvs`-function

6 ☐ Don't know / No answer

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

The rvs function is commonly used in Python's scipy.stats module to generate random variates from a specified probability distribution

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

Continue on page 5

Exercise II

A DTU student wants to ensure that she is never late for an early morning lecture. Over the course of two weeks, she therefore measures the time it takes her to bike to campus. She uses Python and saves her measurements in the vector `x`.

Question II.1 (3)

How can she compute the sample variance of her biking times? (`np` refers to the NumPy package.)

1 ☐ `np.mean(x)`

2 ☐ `np.std(x)`

3 ☐ `np.std(x, ddof = 1)`

4 ☐ `np.var(x)`

5* ☐ `np.var(x, ddof = 1)`

6 ☐ Don't know / No answer

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

The correct option is number 5. It is important to specify ddof = 1.

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

Question II.2 (4)

The student wants to model her transportation times using a normal distribution. Which of the following statements is not true about a normal distribution?

1 ☐ It has two parameters: the mean μ and the variance σ^2.

2 ☐ It is symmetric about the mean.

3 ☐ The mean is equal to the median.

4* ☐ The mean is strictly positive.

5 ☐ The variance is strictly positive.

6 ☐ Don't know / No answer

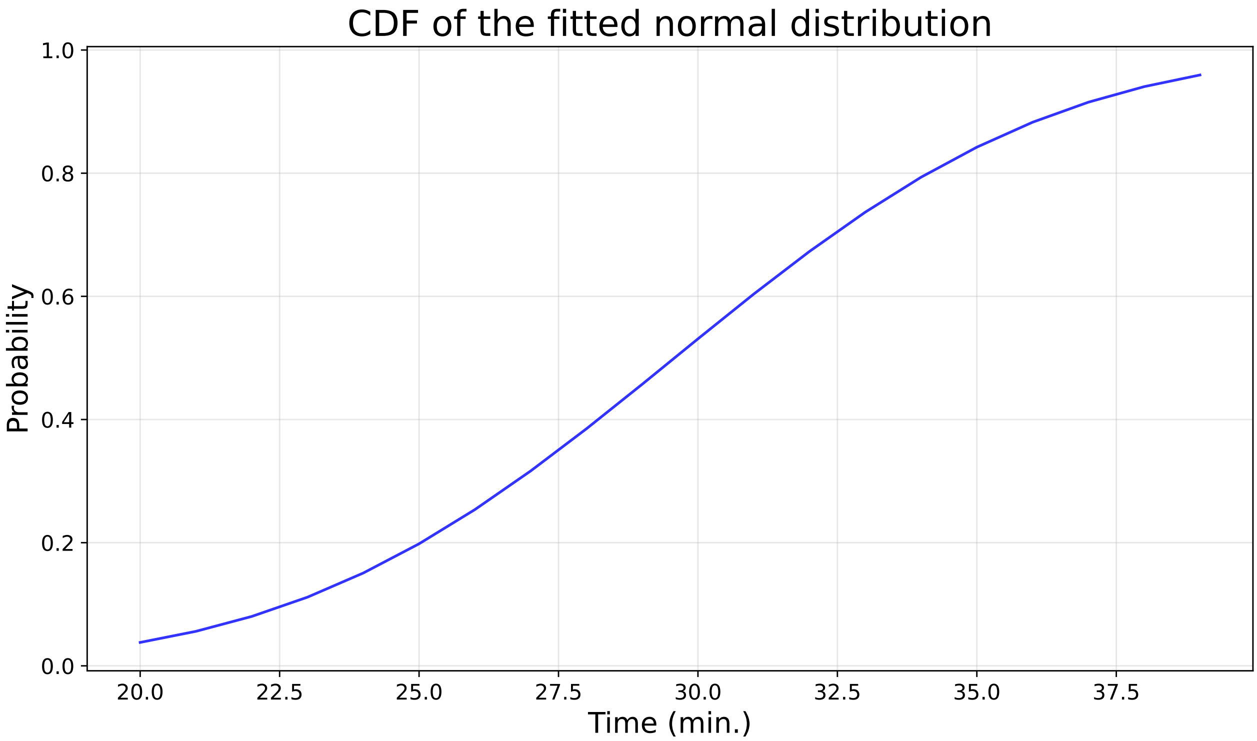
------------------------------------ FACIT-BEGIN ------------------------------------

The mean of a normal distribution can be any number, and therefore option four is correct.

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

Continue on page 7

The student fits a normal distribution to her measured biking times and uses this as a model for future biking times. The figure below shows the cumulative distribution function (cdf) of her model.



Question II.3 (5)

According to the model, what is the probability that she arrives in time for her lecture if she leaves 35 minutes before it starts?

1 ☐ Approximately 15%.

2 ☐ Approximately 45%.

3 ☐ Approximately 65%.

4* ☐ Approximately 85%.

5 ☐ Approximately 99%.

6 ☐ Don't know / No answer

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

The value of the cumulative distribution function in x = 35 is around 0.85, and therefore the correct answer is option 4.

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

Continue on page 8

Exercise III

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 III.1 (6)

What is the covariance between X and Y?

1 ☐ -25

2 ☐ -5

3* ☐ 0

4 ☐ 5

5 ☐ 25

6 ☐ Don't know / No answer

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

Since the random variables are independent, the covariance between the variables must be zero.

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

Question III.2 (7)

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

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

Since expectation is a linear operator, the expected value of Z is found as

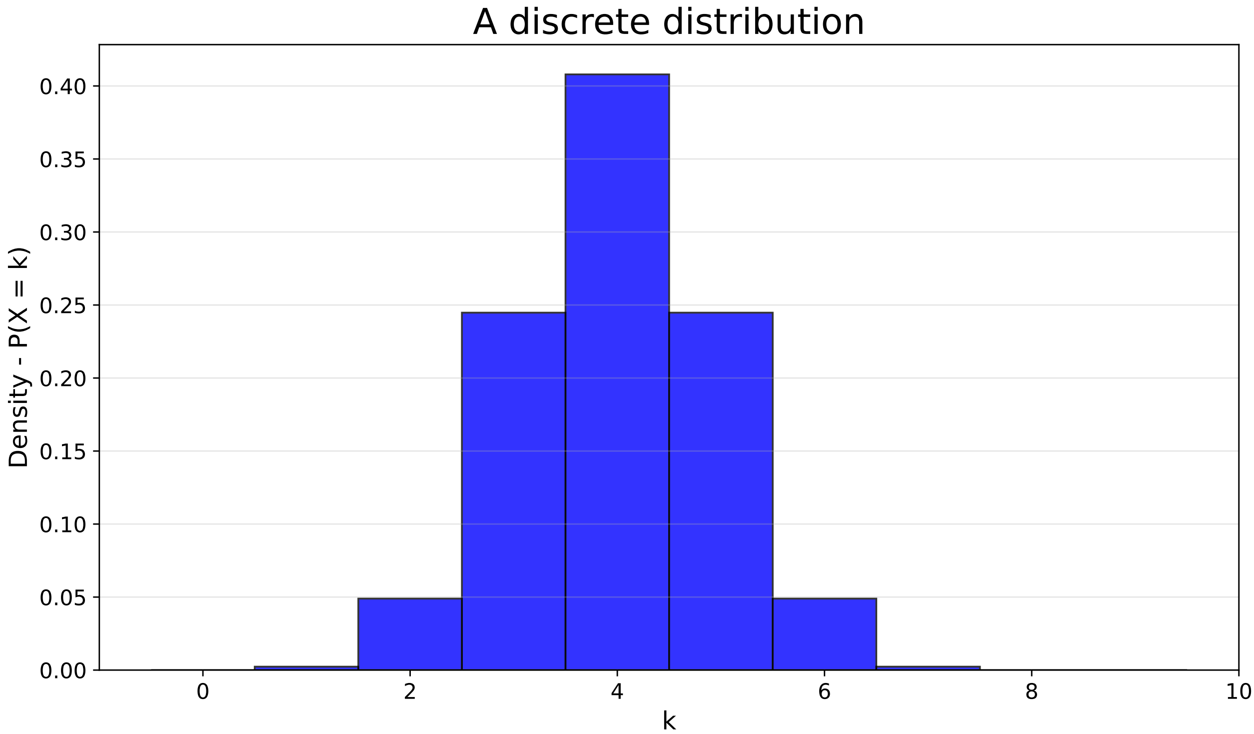
$$\mathbb{E}[Z] = \mathbb{E}[-4X + 3Y] = -4\mathbb{E}[X] + 3\mathbb{E}[Y] = -20 - 9 = -29.$$

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

Continue on page 11

Exercise IV

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



Question IV.1 (8)

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

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

Since the distribution only has support on the set $\{1, ..., 7\}$ i.e., can only take values in said set, the distribution can neither be a binomial nor a Poisson distribution. In conclusion, option 3 is the correct answer.

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

Continue on page 13

Question IV.2 (9)

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

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

The probability in question can be calculated in multiple ways. The most direct calculation is:

$$\mathbb{P}(X < 5) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3) + \mathbb{P}(X = 4) \approx 0 + 0.05 + 0.25 + 0.40 = 0.70.$$

Thus, option 4 is the correct answer.

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

Question IV.3 (10)

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

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

The distribution is perfectly symmetric around 4, which suggests that $\mathbb{E}[X] = 4$. Then, according to definition 2.16:

$$\begin{aligned}\mathbb{V}[X] &\approx (1-4)^2 \cdot 0 + (2-4)^2 \cdot 0.05 + (3-4)^2 \cdot 0.25 + (4-4)^2 \cdot 0.40 \\ &+ (5-4)^2 \cdot 0.25 + (6-4)^2 \cdot 0.05 + (7-4)^2 \cdot 0 \\ &= 0 + 0.2 + 0.25 + 0 + 0.25 + 0.2 + 0 = 0.9.\end{aligned}$$

Hence, option 3 is the correct answer.

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

Continue on page 15

Exercise V

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 μ is the overall average sentence across all judges, α_i is the difference between the average sentence of judge i and the overall average, and ε_{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 V.1 (11)

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

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

This is a one-way ANOVA model, where the judges are considered as treatments. Since the error terms are assumed normally distributed, option 1 is incorrect. Notice then that

$$\mathbb{E}[Y_{ij}] = \mu + \alpha_i$$

and

$$\mathbb{V}[Y_{ij}] = \sigma^2,$$

which shows the mean sentences can vary between the judges as α_i refers to judge i specifically, while the sentence variances must be equal across the judges. In other words, the model allows judges to have different mean sentences, but not different sentence variances.

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

Question V.2 (12)

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

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

If the judges' average sentences differ substantially, this corresponds to large between-judge variation and hence a large SS(judges). Likewise, if each judge gives consistent sentences, within-judge variation is small, leading to a small SSE.

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

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 V.3 (13)

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

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

The ANOVA model does not assume that each judge must have made at least five sentences, nor that all judges must have contributed the same number of sentences. Furthermore, the data need only be normally distributed within each group. However, the model does assume that the judges have equal sentence variances, which is clearly not the case given the data. Therefore, the model assumptions are not satisfied because the judges exhibit considerably different sentence variances.

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

Continue on page 18

Exercise VI

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 VI.1 (14)

Which one of the following statements is false?

1 ☐ The model has three parameters.

2 ☐ The random variables ε_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

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

The 95% confidence interval for the mean response at a given point can never be wider than the 95% prediction interval for an individual response at the same point, cf. equations (5-59) and (5-60). Notice that the intervals are identical except for the additional one under the square root in eq. (5-60).

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

Question VI.2 (15)

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

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

The residual sum of squares (RSS) is calculated using eq. (5-8) as the sum of the squared residuals (as the name suggests)

$$\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{4} e_i^2 = 0^2 + 1^2 + (-2)^2 + 1^2 = 6.$$

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

Continue on page 20

Exercise VII

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 VII.1 (16)

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

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

Let N denote the number of drawn numbers on a ticket. Since the drawing is without replacement, the appropriate model for N is a hypergeometric distribution with $n = 7$ (numbers on a ticket), $a = 7$ (numbers drawn in the lottery), and $N = 36$ (the possible numbers to draw from). The chance of winning the jackpot is then given by eq. (2-25):

$$\mathbb{P}(N = 7) = \frac{\binom{7}{7}\binom{29}{0}}{\binom{36}{7}} = \frac{7!29!}{36!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{36 \cdot 35 \cdot 34 \cdot 33 \cdot 32 \cdot 31 \cdot 30} = \frac{1}{6 \cdot 5 \cdot 17 \cdot 11 \cdot 8 \cdot 31 \cdot 6} = \frac{1}{8347680}.$$

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

Exercise VIII

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 VIII.1 (17)

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

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

For $Y = f(X_1, X_2)$ with $f(x_1, x_2) = x_1^2 + x_2^2$, Method 4.3 yields that

$$\sigma_Y^2(x_1, x_2) = \sum_{i=1}^{2} \left(\frac{\partial f}{\partial x_i}(x_1, x_2) \right)^2 \sigma_{X_i}^2.$$

We derive that

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 2x_1, \quad \frac{\partial f}{\partial x_2}(x_1, x_2) = 2x_2,$$

such that

$$\sigma_Y^2(x_1, x_2) = (2x_1)^2 + (2x_2)^2 = 4x_1^2 + 4x_2^2.$$

Now we evaluate the function in the point $(\mathbb{E}[X_1], \mathbb{E}[X_2]) = (0, 0)$. Thus,

$$\sigma_Y^2(x_1, x_2) = 4(0^2) + 4(0^2) = 0.$$

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

Continue on page 22

Exercise IX

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

Question IX.1 (18)

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

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

Using the formula

$$\bar{x} \pm t_{0.975,df=14} \cdot \frac{s}{\sqrt{n}},$$

we get

$$82 \pm 2.145 \cdot \frac{6.2}{\sqrt{15}} \approx 82 \pm 3.4,$$

resulting in the interval $(78.6, 85.4)$. Therefore, the correct answer is 1.

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

Question IX.2 (19)

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

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

The p-value is less than 0.05 and the confidence interval does not include 75, indicating a statistically significant difference in means between the two groups. Therefore, the correct answer is 4.

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

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	···	13
Average preparation time with AI tutor (min.)	30	45	80	110	···	95
Average preparation time without AI tutor (min.)	30	45	100	120	···	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 IX.3 (20)

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 χ^2-test

5 ☐ A one-way ANOVA test

6 ☐ Don't know / No answer

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

This is a tricky question, because none of the suggested models actually work correctly in this setup. However, there still is one option that is more appropriate than the alternatives. The university considers two groups i.e., it has two samples, but it notes that the expected preparation time varies substantially across lectures, which suggests that the observations within the samples might have different means. Therefore, it makes sense to consider the observations paired by lectures to account for the expected lecture-to-lecture variation. Hence, the paired t-test is the appropriate *out of the possible options*. The standard two-sample test (the Welch test) fails to account for the lecture-to-lecture variation in expected preparation time, and the pooled test (which is equivalent to a one-way ANOVA with two groups) makes the additional mistake of assuming an equal variance across both groups and all lectures. The χ^2-tests are usually only used when working with proportions and is thus completely irrelevant in this context.

To realize why the paired test might fail, we formalize the construction of the data. Therefore, let X_{ij} denote the preparation time of student j with the AI tutor for lecture i, and Y_{ij} denote the preparation time of student j using traditional preparation methods for lecture i. The university then takes the average for each lecture i.e.,

$$X_i = \frac{1}{50}\sum_{j=1}^{50} X_{ij}, \quad Y_i = \frac{1}{300}\sum_{j=1}^{300} X_{ij}.$$

Notice that the central limit theorem implies that we can assume

$$X_i \sim \mathcal{N}\left(\mu_{1i}, \sigma_{1i}/50\right), \quad Y_i \sim \mathcal{N}\left(\mu_{2i}, \sigma_{2i}/300\right).$$

This ensures that the differences $D_i = X_i - Y_i$ are normally distributed, but their mean values and variances are not guaranteed to be the same. This would imply that the differences are not identically distributed, which is assumed in the paired test.

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

Continue on page 25

Question IX.4 (21)

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

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

Since there truly is a difference in mean preparation time between the two groups, the null hypothesis is false. The test therefore reaches the correct conclusion, if it correctly rejects the null hypothesis, The probability of this is exactly the power of the test, which is 90%. Therefore, the correct answer is 2.

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

Question IX.5 (22)

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

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

Since data is right-skewed, we are seeking transformations that transforms large values into smaller values. The exponential transform has the opposite effect and should thus not be considered here. Therefore, the correct answer is 2. You could argue that the reciprocal transformation might also fail to improve the situation if the sample contains many very small preparation times, since such values would be exaggerated by the transformation. However, given only the information that the data are right-skewed, we should not rule out the reciprocal transformation a priori.

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

Continue on page 27

Exercise X

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 X.1 (23)

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

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

Expected count:

$$e = \frac{40 \cdot 80}{240} = 13.33$$

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

Continue on page 28

Question X.2 (24)

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

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

Identify the observed proportion $\hat{p}$
From the data:
Column Total = 80 and Row Total = 120
Total number of students $n = 240$
So the sample proportion is:

$$\hat{p} = \frac{80 + 120}{240} = \frac{200}{240} = 0.833$$

Determine the z-value for 95% confidence
For a 95% confidence level, the critical value is:

$$z_{1-\alpha/2} = z_{0.975} = 1.96$$

Plug into the confidence interval formula

$$\begin{aligned} \text{CI} &= \hat{p} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= 0.833 \pm 1.96 \cdot \sqrt{\frac{0.833 \cdot (1 - 0.833)}{240}} \\ &= 0.833 \pm 1.96 \cdot \sqrt{\frac{0.833 \cdot 0.167}{240}} \\ &= 0.833 \pm 1.96 \cdot \sqrt{\frac{0.139111}{240}} \\ &= 0.833 \pm 1.96 \cdot \sqrt{0.00057963} \\ &= 0.833 \pm 1.96 \cdot 0.02407 \\ &= 0.833 \pm 0.0472 \end{aligned}$$

Final result

$$95\% \text{ Confidence Interval} = [0.833 - 0.0472,\ 0.833 + 0.0472] = [0.7858,\ 0.8802]$$

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

Question X.3 (25)

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

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

$$\hat{p}_{VBL} = \frac{32+30}{80}, \quad \hat{p}_{IS} = \frac{22+52}{84}, \quad \hat{p} = \frac{62+74}{164}$$

$$\hat{p}_{VBL} = 0.775, \quad \hat{p}_{IS} = 8.881, \quad \hat{p} = 0.829$$

$$z = \frac{\hat{p}_{VBL} - \hat{p}_{IS}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{80} + \frac{1}{84}\right)}}$$

$$= \frac{0.775 - 0.881}{\sqrt{0.829(1-0.829)\,(0.0125 + 0.0119)}} \approx -1.80$$

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

Continue on page 31

Question X.4 (26)

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

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

$$\hat{p}_1 = \hat{p}_{IS} = \frac{52}{84}, \quad \hat{p}_2 = \hat{p}_{GLP} = \frac{38}{76}$$

$$CI = \hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$CI = 0.619 - 0.5 \pm 1.96 \cdot \sqrt{\frac{0.619(1-0.619)}{84} + \frac{0.5(1-0.5)}{76}}$$
$$= [-0.034, 0.272]$$

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

Question X.5 (27)

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 χ^2-distribution with 9 degrees of freedom

4 ☐ A χ^2-distribution with 6 degrees of freedom

5* ☐ A χ^2-distribution with 4 degrees of freedom

6 ☐ Don't know / No answer

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

According to Method 7.22, the chi-squared test statistic for an $r \times c$ contingency table is

$$\chi^2_{\text{obs}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

and the degrees of freedom are calculated as:

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

In this case, with $r = 3$ (learning outcomes) and $c = 3$ (training types), we get:

$$df = (3-1)(3-1) = 2 \cdot 2 = 4$$

Exam score	VBL	GLP	IS	Row Total
Below Average	$O_{11} = 18$	$O_{12} = 12$	$O_{13} = 10$	40
Average	$O_{21} = 32$	$O_{22} = 26$	$O_{23} = 22$	80
Above Average	$O_{31} = 30$	$O_{32} = 38$	$O_{33} = 52$	120
Column Total	80	76	84	240

$$\begin{aligned}
e_{1,1} &= \frac{40 \cdot 80}{240}, \quad e_{1,2} = \frac{40 \cdot 76}{240} \quad e_{1,3} = \frac{40 \cdot 84}{240} \\
e_{2,1} &= \frac{80 \cdot 80}{240}, \quad e_{2,2} = \frac{80 \cdot 76}{240}, \quad e_{2,3} = \frac{80 \cdot 84}{240} \\
e_{3,1} &= \frac{120 \cdot 80}{240}, \quad e_{3,2} = \frac{120 \cdot 76}{240}, \quad e_{3,3} = \frac{120 \cdot 84}{240}
\end{aligned}$$

$$\begin{bmatrix} 13.33 & 12.67 & 14 \\ 26.67 & 25.33 & 28 \\ 40.00 & 38.00 & 42 \end{bmatrix}$$

$$\begin{aligned}
\chi^2_{obs} &= \frac{(18-13.33)^2}{13.33} + \frac{(12-12.67)^2}{12.67} + \frac{(10-14)^2}{14} + \frac{(32-26.67)^2}{26.67} \\
&+ \frac{(26-25.33)^2}{25.33} + \frac{(22-28)^2}{28} + \frac{(30-40)^2}{40} + \frac{(38-38)^2}{38} + \frac{(52-42)^2}{42} \\
&= 10.062
\end{aligned}$$

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

Continue on page 34

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

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

We simply identify the parameter estimates from the Python output. In the model, the intercept is v_0 and slope is a, so the estimates are $\hat{v}_0 = 4.470$ and $\hat{a} = 4.006$, respectively.

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

Continue on page 36

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

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

We use Method 5.15 to find the confidence interval. The parameter estimate and standard error are given in the Python output, and only the quantile from the t-distribution is needed to calculate the interval. Since the output states that there are 100 observations in the sample, and the significance level is given as $\alpha = 0.01$, the relevant quantile is the 99.5% quantile of the t-distribution with $100 - 2 = 98$ degrees of freedom, which is given as 2.6269. The 99% confidence interval for v_0 is thus

$$\hat{v}_0 \pm t_{1-0.01/2}(98) \cdot \hat{\sigma}_{v_0} = 4.4695 \pm 2.6269 \cdot 0.421 = [3.364, 5.575].$$

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

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_{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

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

Method 5.14 and Theorem 5.12 state that under the null hypothesis $\mathcal{H}_0 : v_0 = 5$, the test statistic is calculated as

$$T = \frac{\hat{v}_0 - 5}{\hat{\sigma}_{v_0}}.$$

The Python output shows that $\hat{v}_0 = 4.4695$ and $\hat{\sigma}_{v_0} = 0.421$, i.e.

$$t_{\text{obs}} = \frac{4.4695 - 5}{0.421} = -1.260.$$

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

The exam is finished.