English assigment follows the Danish version

Exam question paper for:

Written examination: 20. December 2025

Course name and number: **02323 Introduction to 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 19 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

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

Theorem 3.14: The distribution of the sample mean approaches a normal distribution as the sample size n becomes sufficiently large, regardless of the shape of the population distribution.

The mean of the sampling distribution of the sample mean is equal to the population mean: $\mu_{\bar{X}} = \mu = 40$

and the standard deviation (standard error) of the sampling distribution is: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{150}}$

Therefore, the sample means are approximately normally distributed around 40, not skewed or uniform.

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

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

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

$$
\begin{aligned}
E[X] &= \sum x \cdot f(x) \\
&= 1 \cdot 0.1 + 3 \cdot 0.4 + 5 \cdot 0.3 + 6 \cdot 0.2 \\
&= 0.1 + 1.2 + 1.5 + 1.2 \\
&= 4.0
\end{aligned}
$$

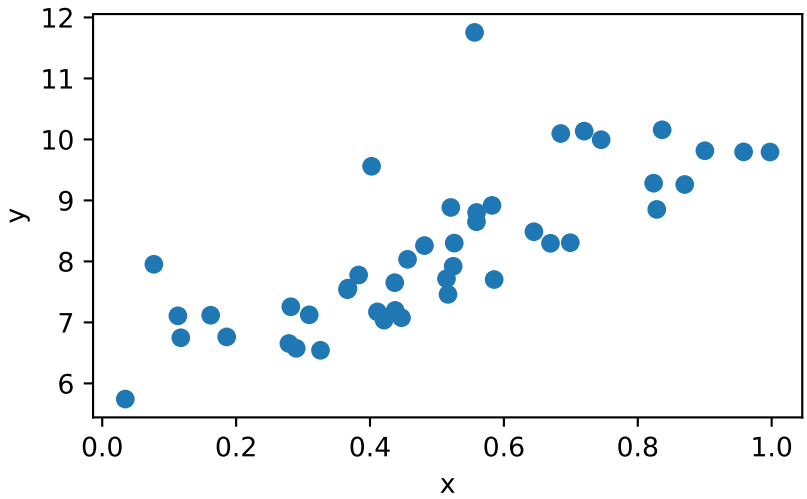
$$
\begin{aligned}
V[X] &= \sum (x - E[X])^2 \cdot f(x) \\
&= (1-4)^2 \cdot 0.1 + (3-4)^2 \cdot 0.4 + (5-4)^2 \cdot 0.3 + (6-4)^2 \cdot 0.2 \\
&= (-3)^2 \cdot 0.1 + (-1)^2 \cdot 0.4 + 1^2 \cdot 0.3 + 2^2 \cdot 0.2 \\
&= 9 \cdot 0.1 + 1 \cdot 0.4 + 1 \cdot 0.3 + 4 \cdot 0.2 \\
&= 0.9 + 0.4 + 0.3 + 0.8 \\
&= 2.4
\end{aligned}
$$

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

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

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

A is the slope 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 is the standard error of A.

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

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

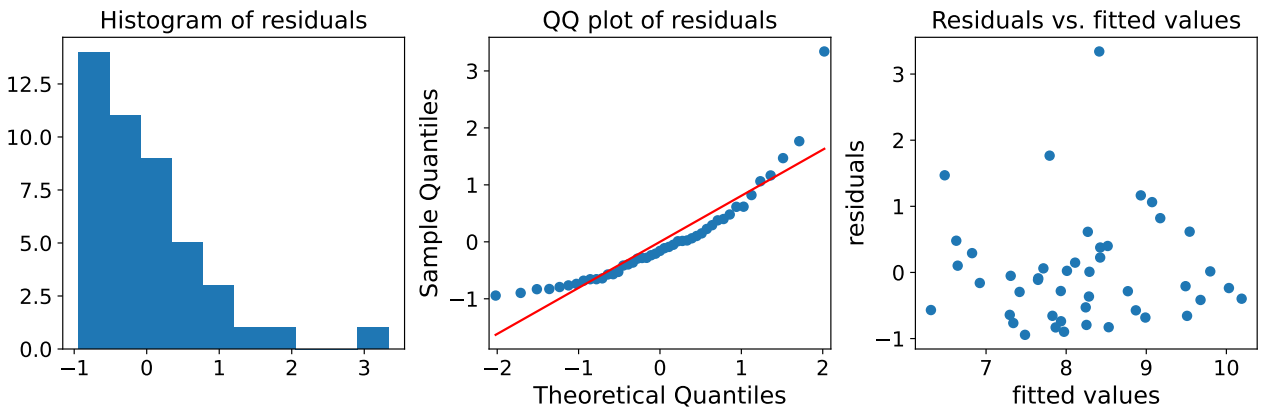
---------- FACIT-BEGIN ----------

From the provided information we estimate $A = \hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{Sxx} = \frac{Sxy}{Sxx} = 4.0273$

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

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

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

The model assumes that $\varepsilon_i \sim N(0, \sigma^2)$ and i.i.d.

Thus we expect a histogram of the residuals to resemble a normal distribution, but the histogram shown in the first plot shows a very skewed distribution of residuals indicating that the assumption about normality is violated (the residuals do not follow a normal distribution).

The conclusion from the qq-plot agrees with the above stated. We expect the quantiles of the residuals to follow the quantiles of a normal distribution (in which case the dots would follow a straight line), but there seems to be a clear systematic deviation from the straight line in the qq-plot. From the qq-plot we would therefore also conclude that the residuals do not seems to follow a normal distribution.

The plot with residuals vs. fitted values does not indicate any clear violation of assumptions, although also here it is apparent that he distribution of residuals is skewed (with more residuals below zero than above). However we do not find any pattern indicating that the residuals are not independent.

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

Continue on page 10

Exercise IV

An industrial engineer is analyzing the production efficiency scores (on a 0–100 scale) of a machine across 30 consecutive days (the measurements can be assumed to be independent). The goal is to assess how consistently the machine performs and to detect possible outliers of unusually high or low efficiency. The quartiles for the daily efficiency scores are as follows:

$$Q_1 = 72, \quad Q_2 = 78, \quad Q_3 = 88.$$

Question IV.1 (6)

Which of the following statements about the Inter Quartile Range (IQR) is correct?

1 ☐ The IQR measures the total spread between the smallest and largest efficiency scores in the dataset.

2 ☐ The IQR is calculated as $Q_1 - Q_3 = 72 - 88 = -16$, which represents the left-skewness of the data.

3* ☐ The IQR is $Q_3 - Q_1 = 88 - 72 = 16$, representing the range of the middle 50% of efficiency scores.

4 ☐ The IQR equals the median ($Q_2 = 78$), which divides the dataset into two equal halves.

5 ☐ The IQR is used only for normally distributed data to calculate the standard deviation.

6 ☐ Don't know / No answer

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

The Interquartile Range (IQR) quantifies the variability of the central 50% of the data and is defined as: $IQR = Q_3 - Q_1$. Here,

$IQR = 88 - 72 = 16$.

This means that on typical days, the machine's efficiency lies between 72 and 88 points. A large IQR indicates inconsistent production performance, while a smaller IQR implies the machine operates with stable efficiency. Outlier days can be identified using:

Lower bound: $Q_1 - 1.5 \times IQR = 72 - 24 = 48$,
Upper bound: $Q_3 + 1.5 \times IQR = 88 + 24 = 112$.
Since efficiency scores above 100 are impossible, any day below 48 would be considered an outlier of poor performance.

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

Continue on page 11

Exercise V

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 V.1 (7)

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

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

The total number of observations is $n = 50 + 150 + 150 + 50 = 400$.

We compute the test statistic by completing the ANOVA table:

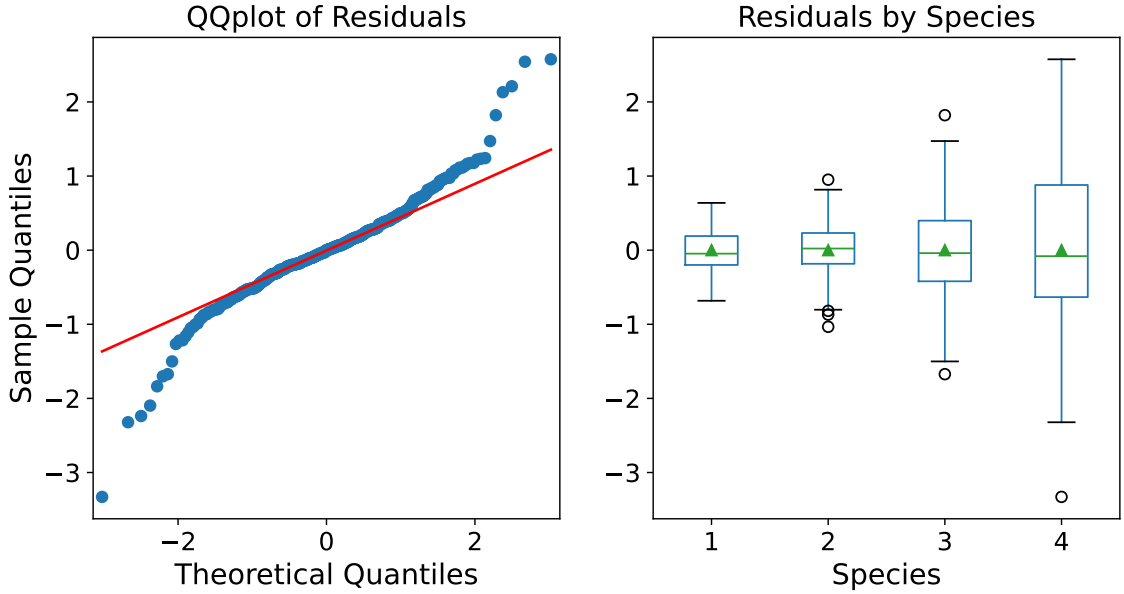
Source	df	SS	MS	F-value	p-value
Species	$4 - 1 = 3$	6.048	$\frac{6.048}{3} = 2.016$	$\frac{2.016}{0.3969} = 5.079$	0.00184
Residual	$400 - 4 = 396$	$163.234 - 6.048 = 157.19$	$\frac{157.19}{396} = 0.3969$		
Total	$400 - 1 = 399$	163.234			

Thus, the value of the test statistic is $F = 5.079$ (you can verify that this does indeed produce a p-value of 0.00184). Since the p-value is just 0.00184, the null hypothesis is rejected at the 0.01 (1%) significance level, but cannot be rejected at the 0.001 (0.1%) significance level.

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

Question V.2 (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

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

We will go through the different plots one at a time. First, note that the researchers are working with a one-way ANOVA model, and the answer options must therefore be evaluated in relation to the assumptions of this model.

The box plots of the residuals suggest the residual variance increases from Species 1 to Species 4. considering the relatively large sample sizes, these plots suggest a clear violation of the homoscedasticity (equal residual variance) assumption. Moreover, number of outliers does not seem disproportionate (given the large sample sizes, especially in Species 2 and 3).

Next, the normal-QQ plot shows considerable departures from the theoretical line, indicating that the assumption of normality is violated.

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

Continue on page 16

Exercise VI

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 VI.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

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

The hypotheses are: $H_0 : \mu = 6, \quad H_1 : \mu \neq 6$.

Because $p = 0.03 < \alpha = 0.05$, we reject the null hypothesis and conclude that the mean daily screen time of university students is not equal to 6 hours.

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

Continue on page 17

Exercise VII

A lecturer collects satisfaction ratings (on a scale from 1–10) for three interactive tools used during lectures — Kahoot, Vevox, and Socrative — to enhance student engagement and conceptual understanding. The summary statistics (mean, standard deviation and coefficient of variation) of student ratings are shown below:

Software	Mean	St. Dev.	CV × 100%
Kahoot	9.0	0.6	6.7%
Vevox	8.2	1.0	12.2%
Socrative	7.0	1.8	25.7%

Question VII.1 (10)

Which of the following statements correctly interprets these results?

1* ☐ Kahoot has the highest mean rating and lowest relative variation, indicating that students rated it consistently high.

2 ☐ Vevox has the smallest standard deviation, indicating it received the best ratings.

3 ☐ Socrative's CV of 25.7% suggests it has the highest student satisfaction and the most consistent ratings.

4 ☐ Kahoot has the highest mean rating but also the largest fluctuations in individual ratings.

5 ☐ The CV values cannot be compared because the sample sizes are different.

6 ☐ Don't know / No answer

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

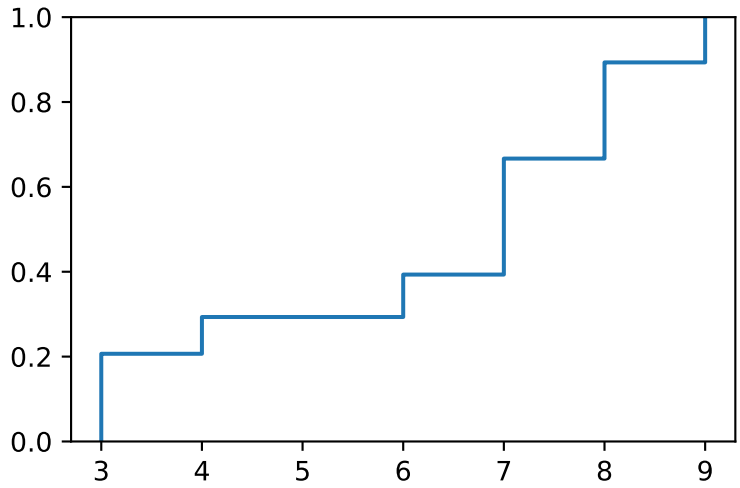
Kahoot has the highest mean rating (mean = 9) and lowest relative variation (small CV = 6.7%), indicating that students rated it consistently high.

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

Continue on page 18

Exercise VIII

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 VIII.1 (11)

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

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

Only discrete values are simulated, hence the answer is one of the options with `np.random.choice`. The "jump heights" in the ecdf plot corresponds to the probabilities of the values and hence we see that option 4 is the correct answer (option 1 would give equal jumps for each value of the sample space).

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

Continue on page 19

Exercise IX

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 IX.1 (12)

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

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

Since the *Margin of Error* is given by:

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

The formula for the required sample size n is:

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

Insert the known values:

$$n = 0.6(1-0.6)\left(\frac{1.96}{0.05}\right)^2 = 0.24(39.2)^2 = 0.24 \times 1536.64 = 368.7$$

Thus, the researcher should plan for a sample size of **369 students** to achieve the desired precision. (This is of course quite many students, and probably the researcher will have to aim for a larger *Margin of Error*.)

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

Question IX.2 (13)

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

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

When no estimate of p is available, we assume the maximum variance case $p = 0.5$, giving $p(1-p) = 0.25$:

$$n = 0.5(1-0.5)\left(\frac{1.96}{0.05}\right)^2 = 0.25(39.2)^2 = 0.25 \times 1536.64 = 384.16$$

Therefore, a minimum of **385 students** should be included in the study to achieve the same margin of error with 95% confidence, even without prior information about p.

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

Question IX.3 (14)

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

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

We have $\hat{p}_1 = 30/50 = 0.6$ and $\hat{p}_2 = 75/150 = 0.5$. For the null hypothesis $H_0 : p_1 = p_2 = p$ the best estimate of p is: $\hat{p} = \frac{x_1+x_2}{n_1+n_2} = 105/200 = 0.525$.

We compute the test statistic: $z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

$$z_{obs} = \frac{0.60 - 0.50}{\sqrt{0.525(1-0.525)\left(\frac{1}{50}+\frac{1}{150}\right)}} = \frac{0.10}{\sqrt{0.249 \times 0.0267}} = \frac{0.10}{0.0815} = 1.23.$$

Since $|z_{obs}| = 1.23 < 1.96$, we **fail to reject** H_0.

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

Continue on page 22

Question IX.4 (15)

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

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

This is a case/scenario with two variables with 3 categorical outcomes, a 3×3 contingency table. According to Method 7.22:
If $\chi^2_{\text{obs}} > \chi^2_{1-\alpha}$ with $(r-1)(c-1)$ degrees of freedom, then reject the null hypothesis.

Degrees of freedom: $df = (r-1)(c-1) = (3-1)(3-1) = 4$.

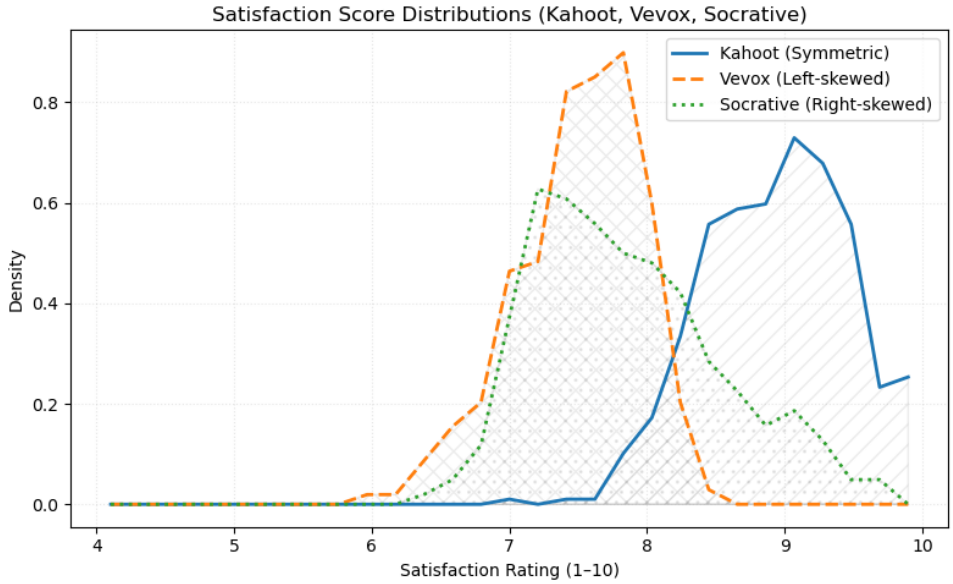
The correct code is: `critical_value = stats.chi2.ppf(1 - alpha, df=4)`

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

Continue on page 23

Exercise X

The plot below shows the distributions of student satisfaction scores (1–10 scale) for three interactive learning tools used during lectures: Kahoot, Vevox, and Socrative. Each curve represents how students rated the tools in terms of engagement and conceptual understanding.



Question X.1 (16)

Which of the following statements correctly interprets the relationship between the mean and median ratings for Vevox and Socrative?

1 ☐ The distribution of ratings for Vevox is left-skewed, meaning that the mean is greater than the median, and the distribution of ratings for Socrative is right-skewed, meaning that the mean is less than the median.

2* ☐ The distribution of ratings for Vevox is left-skewed, meaning that the mean is less than the median, and the distribution of ratings for Socrative is right-skewed, meaning that the mean is greater than the median.

3 ☐ For both distributions the mean and median are identical because satisfaction scores are uniformly distributed.

4 ☐ Vevox's left-skewed distribution means its mean is larger than Socrative's means.

5 ☐ Since both distributions are skewed (not symmetrical) we cannot know if the means are greater or less than the median.

6 ☐ Don't know / No answer

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

- In a left-skewed distribution (Vevox), the long tail on the left pulls the mean downward, so Mean < Median.
- In a right-skewed distribution (Socrative), the long tail on the right pulls the mean upward, so Mean > Median.
(- In a symmetric distribution, the mean and median are approximately equal.)

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

Question X.2 (17)

Which of the following statements correctly describes the main purpose of each of the following visualization methods: box plot, histogram, pie chart, bar plot, and a cumulative distribution plot?

1 ☐ A box plot displays categorical proportions, a histogram shows percentages of categories, a pie chart illustrates numerical spread, a bar plot shows data symmetry, and a cumulative distribution plot identifies outliers.

2* ☐ A box plot summarizes the data distribution using quartiles and highlights outliers; a histogram shows how quantitative data are distributed across intervals; a pie chart illustrates proportions of categories in a whole; a bar plot compares categorical frequencies; and a cumulative distribution plot shows the fraction of frequencies up to each value.

3 ☐ A box plot visualizes frequency counts; a histogram compares group averages; a pie chart is best for continuous data; a bar plot represents quartiles; and a cumulative distribution plot displays discrete counts without accumulation.

4 ☐ A box plot and histogram both represent categorical data; a pie chart and bar plot are for continuous data; and a cumulative distribution plot is for comparing groups.

5 ☐ All five techniques display the same information but differ only in color and layout.

6 ☐ Don't know / No answer

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

Each visualization technique has a distinct purpose in descriptive statistics:

- Box Plot: Summarizes a quantitative variable through its quartiles (Q1, median, Q3), minimum, maximum, and identifies potential outliers.
- Histogram: Shows the frequency or density distribution of continuous (quantitative) data across defined intervals or "bins."

- Pie Chart: Represents categorical data as proportions of a whole, with each slice corresponding to a category's percentage contribution.
- Bar Plot: Compares frequencies, counts, or percentages among discrete categories, where each bar's height reflects its value.
- Cumulative Distribution Plots: Plots the accumulated fraction of frequencies, increasing from 0 to 1, showing how observations accumulate across a range of values.

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

Continue on page 26

Exercise XI

Let X and Y be random variables that both follow a continuous uniform distribution on the interval $[0, 2]$:

$X \sim \mathcal{U}(0, 2)$ and $Y \sim \mathcal{U}(0, 2)$

Furthermore, X and Y are independent.

Question XI.1 (18)

Let $Z = X - Y$. What is the variance of Z?

1 ☐ The variance of Z is: $\mathbf{V}[Z] = 4$

2 ☐ The variance of Z is: $\mathbf{V}[Z] = 2$

3* ☐ The variance of Z is: $\mathbf{V}[Z] = \frac{2}{3}$

4 ☐ The variance of Z is: $\mathbf{V}[Z] = \frac{2}{12}$

5 ☐ The variance of Z cannot be determined with the available information.

6 ☐ Don't know / No answer

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

First we need to calculate the variance of X and Y.

for a random variable that follows a uniform distribution $\mathcal{U}(\alpha, \beta)$ the variance is given by $\frac{1}{12}(\beta - \alpha)^2$, which in this case gives:

$$\mathbf{V}[X] = \mathbf{V}[Y] = \frac{1}{12}(2 - 0)^2 = \frac{4}{12} = \frac{1}{3}$$

To calculate the variance of Z we use (2-74) (the *variance rule*):

$$\mathbf{V}[Z] = \mathbf{V}[X - Y] = (1)^2\mathbf{V}[X] + (-1)^2\mathbf{V}[Y] = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

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

Question XI.2 (19)

Let X^* be the standardized version of X:

$$X^* = \frac{X - \mathbf{E}[X]}{\sqrt{\mathbf{V}[X]}}$$

What is the distribution of X^*?

1 ☐ A continuous uniform distribution on the interval $[0, 1]$.

2 ☐ A continuous uniform distribution on the interval $[-1, 1]$.

3* ☐ A continuous uniform distribution on the interval $[-\sqrt{3}, \sqrt{3}]$.

4 ☐ A standard normal distribution.

5 ☐ None of the above distributions.

6 ☐ Don't know / No answer

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

Translating (add/subtract) and scalling (multiply/divide) a uniform distribution will change the range, but not overall shape of the distribution and hence not the fact that the distribution is uniform (flat).
The standardized random variable X^* will follow a uniform distribution with zero mean ($\mathbf{E}[X^*] = 0$) and variance equal to one ($\mathbf{V}[X^*] = 1$).

Therefore, we must have $\alpha = -\beta$ (such that the uniform distribution is centered around zero).

Using the variance of a uniform distribution we have:

$$\mathbf{V}[X^*] = 1 = \frac{1}{12}(\beta - \alpha)^2 + = \frac{1}{12}(2\beta)^2$$

and solving this for β yields:

$$\beta = \sqrt{\frac{12}{2^2}} = \sqrt{3}$$

(and $\alpha = -\sqrt{3}$)

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

Continue on page 28

Exercise XII

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 XII.1 (20)

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

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

Recall that the confidence interval for σ is: $\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}},\ \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}\right]$

Given $n = 10$, $s = 4.2$, and degrees of freedom $\nu = 9$: $(n-1)s^2 = 9(4.2)^2 = 158.76$
The 95% confidence interval is: $\left[\sqrt{\frac{158.76}{19.02}},\ \sqrt{\frac{158.76}{2.70}}\right] = \left[\sqrt{8.35},\ \sqrt{58.80}\right] = [2.89,\ 7.67]$

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

Continue on page 29

Exercise XIII

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 XIII.1 (21)

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

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

The same participants are measured on the same participants before and after the intervention, so the two samples are dependent (paired).

Therefore, the correct test is the **paired t-test**, implemented in `scipy.stats` as:

`stats.ttest_1samp(Before - After, pop_mean=0)`
(or `stats.ttest_rel(Before, After)`, not shown)

Furthermore, the test should be to test if the average difference between heart rate before and after the program is significantly different than zero, therefore we should use `pop_mean=0`.

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

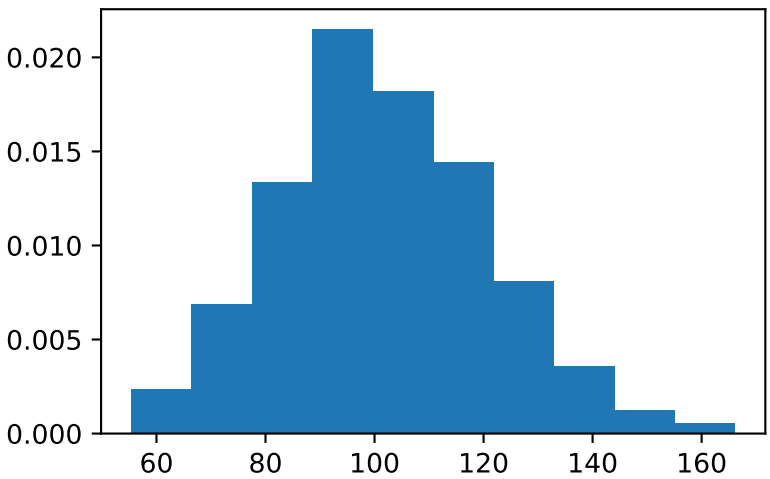
Continue on page 30

Exercise XIV

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 XIV.1 (22)

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

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

A is uniformly distributed between $\alpha_A = 0$ and $\beta_A = 5$ as can be seen from the code:

```
A = stats.uniform.rvs(size=50, loc=0, scale = 5)
```

B follows a normal distribution with mean $\mu_B = 10$ and standard deviation $\sigma_B = 1$, as can be seen from the code:

```
B = stats.norm.rvs(size=50, loc=10, scale = 1)
```

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

Question XIV.2 (23)

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

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

Using error propagation:

$$\begin{aligned}
\sigma_C &= \sqrt{\left(\frac{\partial C}{\partial A}\right)^2 \sigma_A^2 + \left(\frac{\partial C}{\partial B}\right)^2 \sigma_B^2} \\
&= \sqrt{(1)^2 (5/\sqrt{12})^2 + (2B)^2 1^2} \\
&= \sqrt{(1)^2 (25/12) + (20)^2 1} \\
&= \sqrt{25/12 + 400} \approx 20
\end{aligned}$$

Which also matches with the histogram above.

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

Continue on page 34

Exercise XV

A data analyst in a software company examines the relationship between developer experience (in years) and their average code quality score (rated from 0–100). The sample correlation coefficient between the two variables is found to be $r = 0.45$, indicating a moderate positive correlation. However, when plotting the data, the analyst observes that developers with very low and very high experience levels both have lower code quality scores, while those with moderate experience perform best — forming a curved (parabolic) pattern on the scatter plot.

Question XV.1 (24)

Which of the following statements best describes this situation?

1 ☐ The sample correlation coefficient of $r = 0.45$ correctly captures the strong nonlinear relationship between experience and code quality.

2 ☐ The moderate correlation suggests a weak relationship, and the scatter plot confirms there is no pattern in the data.

3* ☐ The curved scatter plot indicates a nonlinear relationship, so the sample correlation coefficient is not a good estimator for the true association between experience and code quality.

4 ☐ The sample correlation coefficient of $r = 0.45$ proves that the relationship between experience and code quality is linear and moderately strong.

5 ☐ The sample correlation coefficient becomes meaningless unless both variables are categorical.

6 ☐ Don't know / No answer

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

The sample correlation coefficient measures the strength of a linear relationship between two quantitative variables. It does not capture nonlinear patterns such as curved or parabolic relationships.

In this case, although $r = 0.45$ suggests a moderate positive linear relationship, the scatter plot reveals a nonlinear trend — developers with mid-level experience perform best, while both beginners and veterans perform slightly worse.

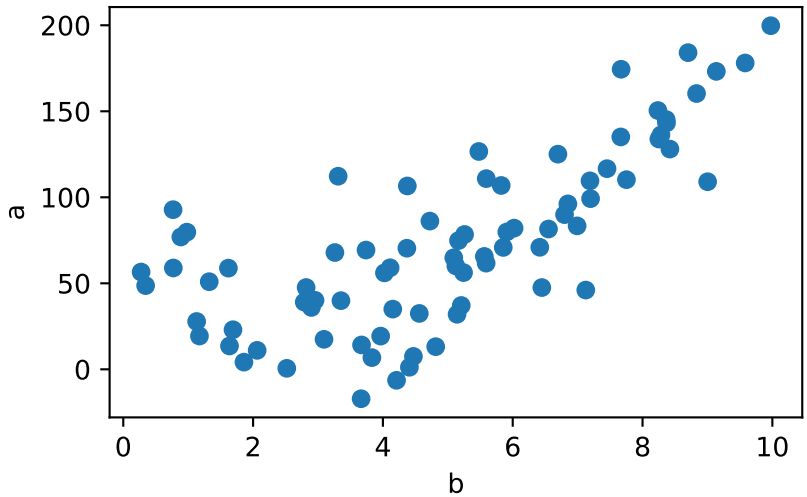
Thus, the correlation underestimates the true association because it ignores the nonlinearity evident in the data.

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

Continue on page 35

Exercise XVI

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



The data was stored in Python in a DataFrame called "D", containing the columns "a", "b" and "b2", where the "b2" column contains squared values of "b" ($b2 = b^2$).

A linear regression model was fitted to the data using the following command in Python:

```
fit = smf.ols(formula= 'a ~ b + b2', data=D).fit()
```

The resulting regression table and some extra information is printed below:

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

                            OLS Regression Results
==============================================================================
Dep. Variable:                      a   R-squared:                       0.692
Model:                            OLS   Adj. R-squared:                  0.684
No. Observations:                  80   F-statistic:                     86.59
Covariance Type:            nonrobust   Prob (F-statistic):           1.99e-20
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     58.9849     11.158      5.286      0.000      36.766      81.204
b            -16.4646      4.894     -3.364      0.001     -26.209      -6.720
b2             3.1350      0.479      6.539      0.000       2.180       4.090
==============================================================================
```

```
print(fit.scale)
795.9021043987004
```

```
print(fit.pvalues)
Intercept    1.131837e-06
b            1.198447e-03
b2           6.124791e-09
dtype: float64
```

Question XVI.1 (25)

What statistical model is fitted and what is the corresponding estimate of the residual standard deviation?

1 ☐ The fitted statistical model is: $a_i = \mu + b_i + b_i^2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$ (and ε_i's are independent). The estimated residual standard deviation is $\hat{\sigma} = 795.9$.

2 ☐ The fitted statistical model is: $a_i = \beta_0 \cdot b_i + \beta_1 \cdot b_i^2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$ (and ε_i's are independent). The estimated residual standard deviation is $\hat{\sigma} = 28.21$.

3 ☐ The fitted statistical model is: $a_i = \beta_0 + \beta_1 \cdot b_i + \beta_2 \cdot b_i^2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$ (and ε_i's are independent). The estimated residual standard deviation is $\hat{\sigma} = 795.9$.

4* ☐ The fitted statistical model is: $a_i = \beta_0 + \beta_1 \cdot b_i + \beta_2 \cdot b_i^2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$ (and ε_i's are independent). The estimated residual standard deviation is $\hat{\sigma} = 28.21$.

5 ☐ The fitted statistical model is: $a_i = \mu + b_i + b_i^2 + \varepsilon_i, \quad \varepsilon_i \sim N(10, \sigma^2)$ (and ε_i's are independent). The estimated residual standard deviation is $\hat{\sigma} = 28.21$.

6 ☐ Don't know / No answer

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

The fitted statistical model is:

$$a_i = \beta_0 + \beta_1 \cdot b_i + \beta_2 \cdot b_i^2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

(and ε_i's are independent).

The estimated residual standard deviation is $\hat{\sigma} = \sqrt{795.9} = 28.21$.

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

Question XVI.2 (26)

Calculate a 99%-confidence interval for the parameter listed in the regression table as coefficient (`coef`) "b".

To answer this question you may need one of the following quantiles found using Python:

```
print(stats.t.ppf(0.975, df=80))
1.990063421028384
```

```
print(stats.t.ppf(0.99, df=80))
2.373868271947045
```

```
print(stats.t.ppf(0.995, df=80))
2.63869059374035
```

```
print(stats.t.ppf(0.975, df=78))
1.990847068555052
```

```
print(stats.t.ppf(0.99, df=78))
2.3751109570702686
```

```
print(stats.t.ppf(0.995, df=78))
2.6403400123362197
```

```
print(stats.t.ppf(0.975, df=77))
1.9912543951146038
```

```
print(stats.t.ppf(0.99, df=77))
2.375756992900897
```

```
print(stats.t.ppf(0.995, df=77))
2.6411976082359923
```

1 ☐ $[-26.21; -6.719]$

2 ☐ $[-28.09; -4.844]$

3* ☐ $[-29.39; -3.539]$

4 ☐ $[-19.11; -13.82]$

5 ☐ $[-29.38; -3.551]$

6 ☐ Don't know / No answer

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

The confidence interval is calculated as follows:

$$\begin{aligned} \hat{\beta}_{\hat{b}} &\pm t_{1-\alpha/2}\hat{\sigma}_{\hat{\beta}_{\hat{b}}} \\ -16.4646 &\pm t_{0.995} \cdot 4.894 \\ -16.4646 &\pm 2.6412 \cdot 4.894 \end{aligned}$$

which gives the interval:

$$[-29.39; -3.539]$$

In the calculation we have used $\alpha = 0.01$, so we need the 0.995% quantile in a t-distribution with $df = n - 3 = 77$ degrees of freedom (the number of observations, $n = 80$, is found in the regression table).

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

Continue on page 40

Exercise XVII

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 XVII.1 (27)

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

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

For a one-sample test, the sample size is given approximately by:

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

where $\mu_0 - \mu_1$ is the difference in means: $\mu_0 - \mu_1 = 5$ mmHg.

Insert the known values into the formula:

$$n = \left(\frac{8(0.84+1.96)}{5}\right)^2 = \left(\frac{8\times 2.8}{5}\right)^2 = (4.48)^2 = 20.07$$

Since we cannot have 20.07 observations we need to round up.
The required sample size per group is 21 participants.

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

Continue on page 42

Exercise XVIII

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 XVIII.1 (28)

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

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

The estimated proportion is $\hat{p} = 30/50 = 0.6$. The confidence interval is given by:

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$0.6 \pm 1.96\sqrt{\frac{0.6(1-0.6)}{50}}$$
$$= [0.46; 0.74]$$

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. Since we have both $n\hat{p} = 30 \geq 15$ and $n(1-\hat{p}) = 20 \geq 15$ we believe that the normal distribution is a good approximation in this case.

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

Continue on page 44

Exercise XIX

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 XIX.1 (29)

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

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

The probability in question is calculated using equation (2-30).

$$P(N < 2) = P(N = 0) + P(N = 1) = \frac{2^0}{0!}e^{-2} + \frac{2^1}{1!}e^{-2} = 3e^{-2} \approx 40.6\%.$$

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

Question XIX.2 (30)

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

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

The time between two consecutive claims, say X, follows an exponential distribution with a mean of half a month i.e., with rate $\lambda = 2$. Thus, the density function is given by

$$f(x) = 2\exp(-2x), \quad x \geq 0,$$

according to eq. (2-68). Equation (2-38) then yields

$$P(X > 1) = \int_1^\infty f(x)dx = \int_1^\infty 2\exp(-2x)dx.$$

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

The exam is finished.