Written examination: 17.05.2025

Course name and number: **Introduction to Statistics (02323)**

Duration: 4 hours

Aids and facilities allowed: All aids - no internet access

The questions were answered by

____________________ ____________________ __________

(student number) (signature) (table number)

This exam consists of 30 questions of the "multiple choice" type, which are divided between 12 exercises. To answer the questions, you need to fill in the "multiple choice" form on exam.dtu.dk.

5 points are given for a correct "multiple choice" answer, and −1 point is given for a wrong answer. ONLY the following 5 answer options are valid: 1, 2, 3, 4, or 5. If a question is left blank or an invalid answer is entered, 0 points are given for the question. Furthermore, if more than one answer option is selected for a single question, which is in fact technically possible in the online system, 0 points are given for the question. The number of points needed to obtain a specific mark or to pass the exam is not predetermined.

The answers should be given by filling in and submitting the digital form. The table provided here is ONLY an emergency alternative. Remember to provide your student number if you do hand in on paper.

Exercise	I.1	I.2	I.3	II.1	II.2	III.1	III.2	IV.1	IV.2	V.1
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer	3	5	2	4	3	3	5	4	5	2

Exercise	V.2	V.3	V.4	V.5	VI.1	VI.2	VII.1	VII.2	VII.3	VIII.1
Question	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Answer	1	5	2	3	4	1	5	4	1	1

Exercise	VIII.2	VIII.3	IX.1	X.1	X.2	XI.1	XI.2	XII.1	XII.2	XII.3
Question	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Answer	5	4	3	1	2	3	1	2	2	3

The exam paper contains 39 pages.

Continue on page 2

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 your own result to the number of decimals given in the answer options before you choose your answer. Also remember that there may be slight discrepancies between the result of the book's formulas and the corresponding built-in functions in Python.*

Exercise I

This exercise covers fundamental statistical concepts and basic Python functionality.

Question I.1 (1)

Which function in the `scipy.stats`-package is used to find theoretical quantiles?

1 □ The `cdf`-function

2 □ The `pdf`-function

3*□ The `ppf`-function

4 □ The `rvs`-function

5 □ The `std`-function

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

CDF: Cumulative Distribution Function - The distribution function
PDF: Probability Density Function - The density function
PPF: Percent Point Function - The quantile function
RVS: Random variates - The random number generator
STD: Standard Deviation

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

Question I.2 (2)

Consider the sample $(1, 2, 3, 9, 10)$. Which of the following statements is false?

1 □ The sample lower quartile (Q1) is 2.

2 □ The sample median is 3.

3 □ The sample mean is 5.

4 □ The sample range is 9.

5*□ The sample variance is 16.

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

The summary statistics are calculated as shown in chapter 1.
The sample lower quartile:

$$q_{0.25} = x_{([0.25 \cdot 5])} = x_{([1.25])} = x_{(2)} = 2.$$

The sample median:

$$q_{0.5} = x_{([0.5 \cdot 5])} = x_{([2.5])} = x_{(3)} = 3.$$

The sample mean:

$$\bar{x} = \frac{1+2+3+9+10}{5} = \frac{25}{5} = 5.$$

The sample variance:

$$s^2 = \frac{(1-5)^2 + (2-5)^2 + (3-5)^2 + (9-5)^2 + (10-5)^2}{(5-1)} = \frac{70}{4} = 17.5.$$

And the sample range is the difference between the largest and smallest observation i.e., the range is $10 - 1 = 9$.

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

Question I.3 (3)

Which of the following quantities is not a random variable

1 □ The estimator $\hat{\beta}$

2*□ The population mean μ

3 □ The sample mean $\bar{X}$

4 □ The sample variance S^2

5 □ The test statistic T

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

Notice that an estimator is by definition always a random variable. Similarly, the sample mean and sample variance are random variable as indicated by the capital letters. Finally, the (unobserved) test statistic is a random variable, again indicated by the capital letter, while the population mean is always a constant.

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

Continue on page 5

Exercise II

Let Y follow an exponential distribution with rate parameter 2, and let U follow a continuous uniform distribution on the interval $[3, 6]$. The two random variables are independent.

Question II.1 (4)

What is the probability that Y exceeds 2?

1 □ $2^{-2} = 0.250$

2 □ $2^{-1} = 0.500$

3 □ $1 - e^{-4} = 0.982$

4* □ $e^{-4} = 0.018$

5 □ $2e^{-4} = 0.037$

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

The density function of $Y \sim \exp(\lambda)$ is given in definition 2.48, and the survival function of Y is then found through an application of eq. (2-42):

$$\mathbb{P}(Y > y) = \int_y^\infty \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_y^\infty = e^{-\lambda y}.$$

Hence, $\mathbb{P}(Y > 2) = e^{-2\lambda} = e^{-2 \cdot 2} = e^{-4} = 0.018$. The probability can also be found as:

```
rate = 2
print(1-stats.expon.cdf(2, loc=0, scale=1/rate))

0.01831563888873422
```

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

Question II.2 (5)

What is the standard deviation of U?

1 □ $\frac{3}{2}$

2 □ $\frac{3}{4}$

3* □ $\frac{\sqrt{3}}{2}$

4 □ $\frac{\sqrt{3}}{4}$

5 □ $\frac{9}{2}$

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

The variance of U is given by eq. (2-50) in theorem 2.36:

$$\sigma^2 = \frac{1}{12}(6-3)^2 = \frac{9}{12} = \frac{3}{4}.$$

Thus, the standard deviation of U is

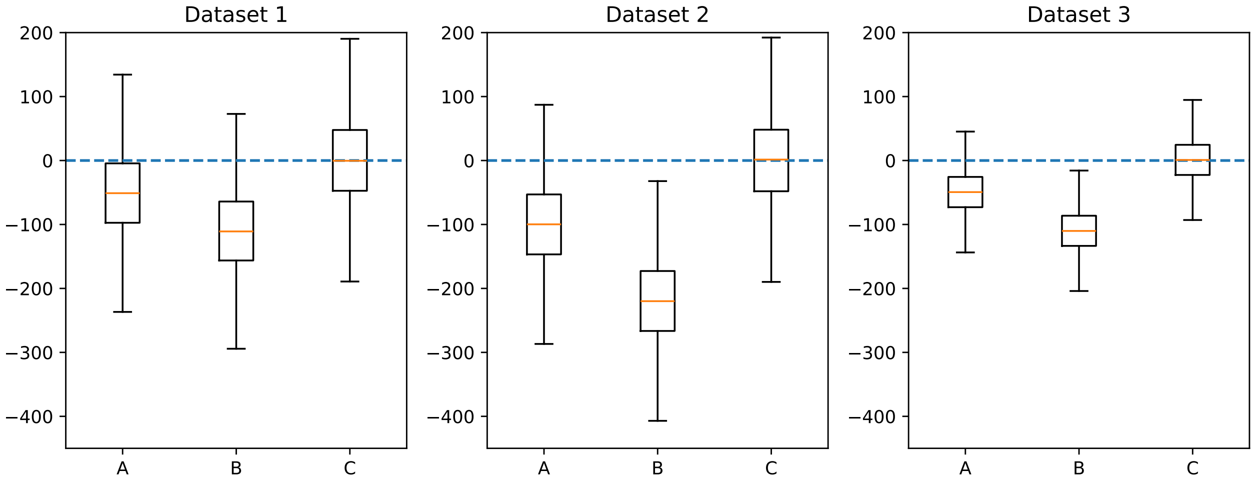
$$\sigma = \sqrt{\frac{3}{4}} = \frac{\sqrt{3}}{2}.$$

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

Continue on page 7

Exercise III

Three experiments compare three medications (A, B, and C) designed to regulate appetite. For each individual in each experiment, the decrease in calorie intake was measured. The results are stored in Dataset 1, Dataset 2, and Dataset 3 and are visualized in the boxplots below:



Question III.1 (6)

Which of the following statements is false?

1 □ The average effects of the medications seem the same in Datasets 1 and 3.

2 □ The SS(TR) - *Treatment Sum of Squares* - is largest in Dataset 2.

3* □ The SSE - *Sum of Squared Errors* - is largest in Dataset 3.

4 □ Within each dataset, the variance in each group (A, B, and C) is approximately equal.

5 □ In all three datasets, medication C could be a placebo (i.e., have no average effect).

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

The SSE represents the within-group variance, and since the variation within each group (A, B, C) is smallest in Dataset 3, statement number 3 is incorrect.

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

Question III.2 (7)

The sample size of each group is the same in all three datasets (i.e., all nine groups are of the same size). For each dataset, an analysis of variance (ANOVA) is performed, and the observed F-statistic and p-value are computed for the usual null hypothesis. Which of the following statements is correct?

1 □ The observed F-statistic is the same for all three datasets.

2 □ The test performed is a Welch t-test.

3 □ The p-value is the same for all three datasets.

4 □ The assumption of equal variances across groups appears violated in Dataset 1.

5*□ The p-value for Dataset 1 is larger than those for Datasets 2 and 3.

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

The p-value increases when SSE increases (indicating increased variation within groups) and SS(TR) decreases (indicating decreased variation between groups). Datasets 1 and 2 both have a larger SSE compared to Dataset 3. Datasets 1 and 3 have a smaller SS(TR) compared to Dataset 2. Therefore, Dataset 1 will give the largest p-value, and statement 5 is true.

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

Continue on page 9

Exercise IV

A fitness coach claims that the average weight loss for a new diet program is 4 kg. To test this claim, a random sample of 12 participants who completed the program was taken, and their weight losses (in kg) were recorded: 3.5, 4.2, 3.8, 4.4, 4.1, 3.9, 4.0, 4.3, 3.6, 3.7, 4.5, 4.1.

The researchers then perform a t-test to determine whether the average weight loss differs significantly from 4 kg at a significance level of 5%.

Question IV.1 (8)

Let t_{obs} be the observed test statistic for the t-test. What is the rejection rule?

1 □ Reject H_0 if $t_{\text{obs}} \leq t_{1-\alpha}$, where $t_{1-\alpha}$ is the $(1-\alpha)$ quantile of a t-distribution with 11 degrees of freedom.

2 □ Reject H_0 if $t_{\text{obs}} \leq t_{1-\alpha}$, where $t_{1-\alpha}$ is the $(1-\alpha)$ quantile of a t-distribution with 12 degrees of freedom.

3 □ Reject H_0 if $t_{\text{obs}} \geq t_{1-\alpha}$, where $t_{1-\alpha}$ is the $(1-\alpha)$ quantile of a t-distribution with 11 degrees of freedom.

4* □ Reject H_0 if $t_{\text{obs}} \leq -t_{\alpha/2}$ or if $t_{\text{obs}} \geq t_{1-\alpha/2}$, where $t_{\alpha/2}$ and $t_{1-\alpha/2}$ are the $\alpha/2$ and $(1-\alpha/2)$ quantiles of a t-distribution with 11 degrees of freedom, respectively.

5 □ Reject H_0 if $t_{\text{obs}} \leq -t_{\alpha/2}$ or if $t_{\text{obs}} \geq t_{1-\alpha/2}$, where $t_{\alpha/2}$ and $t_{1-\alpha/2}$ are the $\alpha/2$ and $(1-\alpha/2)$ quantiles of a t-distribution with 10 degrees of freedom, respectively.

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

The critical values for a (two-sided) one-sample t-test are given in definition 3.31. Thus, the critical values are the $\alpha/2$ and $(1-\alpha/2)$ quantiles of a t-distribution with $n-1=11$ degrees of freedom. Method 3.32 and the subsequent paragraphs describe the rejection rules for the hypothesis test and show that the null hypothesis is rejected if the observed test statistic is not in the interval $(t_{\alpha/2}, t_{1-\alpha/2})$.

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

Question IV.2 (9)

What is the value of the observed test statistic (t_{obs})?

1 □ $t_{\text{obs}} = 2.83$

2 □ $t_{\text{obs}} = 1.09$

3 □ $t_{\text{obs}} = 1.03$

4 □ $t_{\text{obs}} = 0.32$

5*□ $t_{\text{obs}} = 0.09$

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

The test statistic is defined in method 3.23, specifically in eq. (3-21).

```
# Data
weight_loss = np.array([3.5,4.2,3.8,4.4,4.1,3.9,4.0,4.3,3.6,3.7,4.5,4.1])

# Number of observations
n = len(weight_loss)

# Estimated sample mean and sample standard deviation
sample_mean = np.mean(weight_loss)
sample_std = np.std(weight_loss, ddof=1)

# Mean value under the null hypothesis
claimed_mean = 4

# Calculate the t-statistic
t_stat = (sample_mean - claimed_mean) / (sample_std / np.sqrt(n))

# Output the result
print(f"Test statistic: {t_stat}")
```

```
Test statistic: 0.09090909090909542
```

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

Continue on page 11

Exercise V

A computer engineer wants to test the efficiency of an algorithm. She records the execution time (`time`) for jobs of varying complexity (`complex`).

To analyze the data, she runs the code below, where `dat` contains the recorded data.

```python
fit1 = smf.ols(formula = 'time ~ complex', data = dat).fit()
print(fit1.summary(slim=True))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   time   R-squared:                       0.971
Model:                            OLS   Adj. R-squared:                  0.971
No. Observations:                  50   F-statistic:                     1631.
Covariance Type:            nonrobust   Prob (F-statistic):           1.03e-38
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.1732      0.007    -23.620      0.000      -0.188      -0.158
complex        0.0006   1.47e-05     40.391      0.000       0.001       0.001
==============================================================================
```

Question V.1 (10)

Consider the test statistic for the null hypothesis that the intercept is zero. Which distribution is the test statistic compared to in order to obtain the associated p-value?

1 □ A $t(50)$-distribution

2* □ A $t(48)$-distribution

3 □ An $F(1, 50)$-distribution

4 □ A $\mathcal{N}(0, 1^2)$-distribution

5 □ A $\mathcal{N}(0, 0.007^2)$-distribution

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

Theorem 5.12 states that the test statistic for the null hypothesis follows a t-distribution with $n - 2$ degrees of freedom. Since the output shows that there are $n = 50$ observations in the sample, the test statistic should be compared with a $t(48)$-distribution.

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

Question V.2 (11)

Based on the output above, which statement about the validity of the model (i.e., the model assumptions) is correct?

1* □ The validity of the model assumptions cannot be assessed based on the output.

2 □ The assumptions must be satisfied because the R^2-value is close to 1.

3 □ The model should be extended with additional terms because the p-values are small.

4 □ The assumptions must be satisfied because the p-values are close to zero.

5 □ The model should be extended with additional terms because the R^2-value is close to 1.

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

The validity of the model assumptions cannot be assessed based on the output considered. Instead, the output provides information about the number of observations, the method of fitting, and the goodness of fit. The test statistics and p-values provided in the output pertain to hypothesis testing, not to the validity of the model assumptions.

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

For the following questions, the numbers below may be useful:

```
print(np.mean(dat["complex"]))
```

```
484.58
```

```
print(np.var(dat["complex"], ddof=1))
```

```
12603.187346938777
```

Question V.3 (12)

Let $\hat{\sigma}$ denote the estimated residual standard deviation. What is the 95% confidence interval for the mean (execution) time of a job with complexity 300 according to the model?

1 □ $0.18 \pm 3.32\hat{\sigma}$

2 □ $0.0068 \pm 2.08\hat{\sigma}$

3 □ $0.18 \pm 0.82\hat{\sigma}$

4 □ $0.0068 \pm 1.27\hat{\sigma}$

5* □ $0.0068 \pm 0.55\hat{\sigma}$

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

The confidence interval (band) for the predicted values (the regression line) is given in method 5.18. The parameter estimates are given in the computer output.

```
# The number of observations
n = 50

# Parameter estimates
B0 = -0.1732
B1 = 0.0006

# The value of x (Complexity)
xnew = 300
```

In addition, the $1 - \alpha/2 = 0.975$ quantile of the $t(50 - 2)$-distribution and S_{xx} are needed to calculate the confidence interval. The latter can be derived from its definition in theorem 5.4

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = (n-1)\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right),$$

i.e. S_{xx} is the product of $(n - 1)$ and the sample variance of the x-values (it is given above):

```
# The 0.975 quantile of the t(48)-distribution
tval = stats.t.ppf(1-0.05/2,df=n-2)

# The sample variance of the x-values
SVX = 12603.187346938777

# The Sxx
Sxx = (n-1)*SVX
```

Thus, the confidence interval is centered at

```
# The predicted value (Center of CI)
B0 + B1*xnew

0.0068000000000000005
```

and the associated margin of error is $\hat{\sigma}$ times

```
# Average of the x-values
MX = 484.58

# Margin of error factor
tval*np.sqrt(1/n+(xnew-MX)**2/Sxx)

0.5512539671493902
```

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

Question V.4 (13)

What is the estimate of the residual standard deviation ($\hat{\sigma}$)?

1 □ $2 \cdot 10^{-5}$

2* □ 0.01

3 □ 5.5

4 □ 0.97

5 □ 0.007

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

The estimate of the residual standard deviation is found rewriting equation (5-44):

$$\hat{\sigma} = \hat{\sigma}_{\beta_1}\sqrt{S_{xx}}.$$

The standard deviation (std. error) of the estimator $\hat{\beta}_1$ is given in the computer output, and S_{xx} was found above (in the solution to the previous question):

```
# The Sxx
Sxx = 49*12603.19

# The std. error of the estimator
stdB1 = 1.47*10**(-5)

# Estimate of the residual standard deviation
stdB1*np.sqrt(Sxx)
```

```
0.01155195840660362
```

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

The engineer figures that the execution time is proportional to the third power of complexity and formulates the following model:

$$Y_i = \beta_0 x_i^{\beta_1} \varepsilon_i; \quad \varepsilon_i \sim LN(0, \sigma^2),$$

where Y_i and x_i are the time and complexity of job i, respectively. She hence tests the hypothesis $\mathcal{H}_0 : \beta_1 = 3$ against a two-sided alternative by running the below code chunk:

```
ltime = np.log(dat[’time’])
lcomplex = np.log(dat[’complex’])
dat2 = pd.DataFrame({’ltime’ : ltime, ’lcomplex’ : lcomplex})
fit2 = smf.ols(formula = ’ltime ~ lcomplex’,data=dat2).fit()
print(fit2.summary(slim=True))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  ltime   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.999
No. Observations:                  50   F-statistic:                 8.032e+04
Covariance Type:            nonrobust   Prob (F-statistic):           4.87e-79
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -18.0015      0.055   -325.241      0.000     -18.113     -17.890
lcomplex       2.5459      0.009    283.406      0.000       2.528       2.564
==============================================================================
```

Question V.5 (14)

Based on the output above, what is the test statistic (t_{obs}) for the usual test of $\mathcal{H}_0$, and what is the conclusion of the test at a significance level of 5%?

1 □ The hypothesis is rejected as $t_{\text{obs}} = -5.645$.

2 □ The hypothesis is accepted as $t_{\text{obs}} = 13.84$.

3* □ The hypothesis is rejected as $t_{\text{obs}} = -50.46$.

4 □ The hypothesis is rejected as $t_{\text{obs}} = 283.4$.

5 □ The hypothesis is accepted as $t_{\text{obs}} = -5.645$.

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

The model fitted in the code chunk is given as

$$\ln(Y_i) = \ln\left(\beta_0 x_i^{\beta_1} \varepsilon_i\right) = \ln(\beta_0) + \beta_1 \ln(x_i) + \ln(\varepsilon_i), \quad \ln(\varepsilon_i) \sim \mathcal{N}(0, \sigma^2)$$

since $\varepsilon_i \sim \text{LN}(0, \sigma^2)$. The output from the code chunk gives the parameter estimate $\hat{\beta}_1 = 2.5459$ and the estimated standard deviation (std. error) $\hat{\sigma}_{\beta_1} = 0.009$. The test statistic is then calculated using theorem 5.12

$$t_{obs} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}} = \frac{2.5459 - 3}{0.009} = -50.46.$$

This is verified numerically as:

```
# Parameter estimate
B1hat = 2.5459

# Parameter value under H0
B10 = 3

# Estimated std. deviation of parameter
sigB1 = 0.009

# Test statistic
tobs = (B1hat-B10)/sigB1
print(tobs)
```

```
-50.455555555555556
```

Since the numerical value of the observed test statistic exceeds the $1 - \alpha/2 = 0.975$ quantile of the $t(48)$-distribution (as $n = 50$ leads to $n - 2 = 48$):

```
# The 0.975 quantile of the t(48)-distribution
tval = stats.t.ppf(1-0.05/2,df=n-2)
print(tval)

2.010634757624232
```

the null hypothesis is rejected at the 5% significance level.

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

Continue on page 19

Exercise VI

The two questions in this exercise must be solved using simulation.

The number of customers shopping in a supermarket on a randomly selected day follows a Poisson distribution with a mean of 5000. The amount a randomly selected customer spends is exponentially distributed, with an average of 200 DKK. It is assumed that customers act independently of each other and independently of the number of customers.

The following code snippet can be used to simulate k daily revenues, but you need to complete the missing parts of the code by specifying the parameters of the distributions:

```
# Set seed
np.random.seed(2025)

# Number of samples
k = 100000

# Number of customers (array of customer counts for the k days)
N = stats.poisson.rvs(mu=____,size=k)

# The revenues (array with revenues for the k days)
Y = np.array([np.sum(stats.expon.rvs(scale=_____, size=n)) for n in N])
```

Let Y denote the supermarket's revenue (or turnover) on a randomly selected day.

Question VI.1 (15)

What is the expectation and the standard deviation of Y?

1 □ $\mathbb{E}[Y]$ = 1,500,000 DKK and SD(Y) = 1,000,000 DKK

2 □ $\mathbb{E}[Y]$ = 1,500,000 DKK and SD(Y) = 20,000 DKK

3 □ $\mathbb{E}[Y]$ = 1,000,000 DKK and SD(Y) = 1,000,000 DKK

4* □ $\mathbb{E}[Y]$ = 1,000,000 DKK and SD(Y) = 20,000 DKK

5 □ $\mathbb{E}[Y]$ = 500,000 DKK and SD(Y) = 1,000,000 DKK

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

Denote the number of customers shopping in the supermarket on the selected day by N, and let X_i denote the amount that customer i spends on the selected day. Then the model is

$$Y = \sum_{i=1}^{N} X_i.$$

One can show that

$$\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X_i] = 5000 \cdot 200 = 1000000$$

and that

$$\mathbb{V}[Y] = \mathbb{E}[N]\mathbb{V}[X_i] + \mathbb{E}[X_i]^2\mathbb{V}[N] = 5000 \cdot 200^2 + 200^2 \cdot 5000 = 10000 \cdot 200^2 = (100 \cdot 200)^2 = 20000^2.$$

This is however beyond the scope of the course, and the estimates must be obtained through simulation. For example:

```
# Set seed
np.random.seed(2025)

# Number of samples
k = 100000

# Number of customers
N = stats.poisson.rvs(mu=5000,size=k)

# The revenues
Y = np.array([np.sum(stats.expon.rvs(scale=200, size=n)) for n in N])

# Find expectation and standard deviation
print(Y.mean())
```

```
1000033.2324169412
```

```
print(Y.std(ddof=1))
```

```
19981.067769057736
```

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

Question VI.2 (16)

What is the probability that the supermarket makes more than 1,500,000 DKK in revenue on a randomly selected day?

1* □ Almost 0%

2 □ Approximately 11%

3 □ Approximately 22%

4 □ Approximately 33%

5 □ Approximately 50%

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

The distribution of Y is rather difficult to determine, however Chebyshev's inequality yields that

$$\mathbb{P}(Y > 1,500,000) \leq \mathbb{P}(|Y - \mu_Y| > 500,000) = \mathbb{P}(|Y - \mu_Y| > 25\sigma_Y) \leq \left(\frac{1}{25}\right)^2 = 0.16\%.$$

Hence, the probability is almost zero. This is however beyond the scope of the course, and the estimates must be obtained through simulation. For example:

```
# Probability
1*(Y > 1500000).mean()

0.0
```

We use that the probability of an event is the expectation of the indicator variable of that event, i.e.

$$\mathbb{P}(Y > 1,500,000) = \mathbb{E}[1_{\{Y>1,500,000\}}].$$

See page 188 under section 4.1 for further explanation.

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

Continue on page 22

Exercise VII

A shipping company has hired a sales representative to recruit new clients. The sales representative visits n (independent) clients each month, and historical data suggests that the probability of recruiting a client after a visit is p. Let X denote the number of recruited clients in a randomly selected month.

Question VII.1 (17)

Which one of the following models is the most appropriate?

1 □ $X \sim U(0, n)$

2 □ $X \sim H(n, np, n)$

3 □ $X \sim \mathcal{N}(np, p^2)$

4 □ $X \sim \text{Pois}(np)$

5* □ $X \sim \text{Bin}(n, p)$

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

The sales representative performs a sequence of independent and identical Bernoulli experiments. They visit n independent clients, and each client has probability p of being recruited (considered a success here). Therefore, the number of recruited clients (number of successes) follows a binomial distribution with parameters n and p, cf. section 2.3.1. Thus, $X \sim \text{Bin}(n, p)$ is the most appropriate model.

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

The sales representative earns a salary of 5000 USD per month and receives an additional bonus of 100 USD for each client recruited during the month. The sales representative's total compensation in a randomly selected month (Y) is therefore given by

$$Y = 100X + 5000,$$

where X denotes the number of clients recruited during that month. You may assume that the standard deviation of X is 2.

Question VII.2 (18)

What is the variance of Y?

1 □ 5200

2 □ 20000

3 □ 25000

4* □ 40000

5 □ 45000

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

Using theorem 2.54, specifically eq. (2-72), the variance of Y is found as

$$\mathbb{V}[Y] = \mathbb{V}[100X + 5000] = 100^2\mathbb{V}[X] = 100^2\text{SD}[X]^2 = 100^2 \cdot 2^2 = 40000.$$

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

Question VII.3 (19)

Which one of the following statements about the correlation between X and Y is true?

1* □ The correlation between X and Y is one: $\rho(X, Y) = 1$.

2 □ The correlation between X and Y is between one and zero: $0 < \rho(X, Y) < 1$.

3 □ The correlation between X and Y is zero: $\rho(X, Y) = 0$.

4 □ The correlation between X and Y is negative: $-1 \leq \rho(X, Y) < 0$.

5 □ The correlation cannot be determined without additional information.

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

Since all observations (X_i, Y_i) must be on a straight line with a positive slope, the correlation must be equal to one, i.e. $\rho(X, Y) = 1$, cf. remark 1.21. Alternatively, the correlation can be calculated explicitly using definition 2.62:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\text{SD}[X]\text{SD}[Y]} = \frac{\text{Cov}(X, 100X + 5000)}{2 \cdot \sqrt{40000}} = \frac{\text{Cov}(X, 100X)}{400} = \frac{100\mathbb{V}[X]}{400} = \frac{100 \cdot 2^2}{400} = 1.$$

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

Continue on page 25

Exercise VIII

A municipality tested a service that allows citizens to report damages to traffic infrastructure (e.g., roads and traffic lights) via an app. The reported issues were categorized into three groups: needs immediate repair, can be postponed, or can be ignored.

During a one-year trial, $n = 165$ independent issues were reported. The counts for each issue type were compiled by season and entered into the below table:

```
# Reading the data into Python
data = np.array([[44, 10, 32, 18], [20, 8, 4, 2], [11, 2, 10, 4]])
# Converting to a Pandas dataframe
data = pd.DataFrame(data, index=['Immediate', 'Postpone', 'Ignore'],
columns=['Spring', 'Summer', 'Fall', 'Winter'])
print(data)

           Spring  Summer  Fall  Winter
Immediate      44      10    32      18
Postpone       20       8     4       2
Ignore         11       2    10       4
```

The municipality plans to test if the distribution of fault types depends on the season. Specifically, it tests the null hypothesis

$$H_0 : p_{i1} = p_{i2} = p_{i3} = p_{i4} = p_i \text{ for } i = 1, 2, 3,$$

where p_{ij} represents the proportion of observations in column j that fall into row i i.e., the proportion of fault type i out of all faults in season j, and p_i represents the proportion of all observations that are in row i.

Question VIII.1 (20)

Under the null hypothesis, what is the estimated proportion of faults requiring immediate repair?

1* □ 0.6303

2 □ 0.4231

3 □ 0.2061

4 □ 0.7939

5 □ None of the above answers are correct.

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

The dataset constitutes a 3×4 frequency table (see Chapter 7.5 of the book). Under the null hypothesis, the distribution of fault types does not depend on the season. Therefore, the correct estimate of the proportion of faults requiring immediate repair is given by:

$$\hat{p}_1 = \frac{44 + 10 + 32 + 18}{165} = \frac{104}{165} = 0.6303.$$

This represents the proportion of all faults (ignoring the season) that require immediate repair.

```
# Rowsums
data.sum(axis=1)

Immediate    104
Postpone      34
Ignore        27
dtype: int64

# Estimated proportion
data.iloc[0].sum(axis=0)/np.sum(data.values)

0.6303030303030303
```

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

Question VIII.2 (21)

At a significance level of $\alpha = 0.05$, what is the correct conclusion for the usual test of the null hypothesis? (Both argument and conclusion must be correct.)

1 □ The null hypothesis is rejected, indicating a significant difference in distribution between the seasons, as the p-value is 0.04.

2 □ The null hypothesis is rejected, indicating a significant difference in distribution between the seasons, as the p-value is 0.08.

3 □ The null hypothesis is accepted, indicating no significant difference in distribution between the seasons, as the p-value is 0.04.

4 □ The null hypothesis is accepted, indicating no significant difference in distribution between the seasons, as the p-value is 0.08.

5* □ No conclusion can be drawn, since some expected cell counts under the null hypothesis are too low i.e., $e_{ij} < 5$ for some cells.

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

The validity of the test relies on an assumption that is only valid when the expected cell count is at least five for all cells. Although method 7.22 does not explicitly state this (it appears only on the week 10 slides), the approximation scheme is the same as in section 7.4, and therefore the same rules of validity apply. The expected cell counts are calculated using eq. (7-53) from method 7.22 or with Python as:

```
# Test
chi2, p_val, dof, expected = stats.chi2_contingency(data,
correction=False)

# Expected cell counts
print(pd.DataFrame(expected, index=['Immediate', 'Postpone', 'Ignore'],
columns=['Spring', 'Summer', 'Fall', 'Winter']))

              Spring     Summer       Fall     Winter
Immediate  47.272727  12.606061  28.993939  15.127273
Postpone   15.454545   4.121212   9.478788   4.945455
Ignore     12.272727   3.272727   7.527273   3.927273
```

The output indicates that the expected number of e.g. faults classified as ”can be ignored” during both summer and winter is below 5.

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

Another purpose was to test an AI system for automatically handling the reported faults, comparing its performance to manual evaluation, where a human categorized the faults. Based on expert knowledge, the number of discrepancies between the two methods is expected to be 21 (out of the 165 observed faults).

Question VIII.3 (22)

Let $X \sim \text{Bin}(n, p)$ be the count of how many faults were evaluated differently (You need to estimate p). What is the estimate of the standard deviation of X/n?

1 □ $\hat{\sigma} = 0.3343$

2 □ $\hat{\sigma} = 0.1273$

3 □ $\hat{\sigma} = 0.1118$

4* □ $\hat{\sigma} = 0.0259$

5 □ $\hat{\sigma} = 0.0007$

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

Initially, the parameters of the distribution must be determined. The parameter n was given as 165 (the number of independent issues that were reported), which yields the estimate

$$\hat{p} = \frac{x}{n} = \frac{21}{165}$$

according to eq. (7-2). The standard deviation can now be calculated using eq. (7-6) with $\hat{p}$ instead of p (see p. 276 and example 7.6):

$$\hat{\sigma} = \sqrt{\text{Var}\left(\frac{X}{n}\right)} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{21/165(1-21/165)}{165}} = 0.0259.$$

In Python, the answer is found as:

```python
# The observed number of discrepancies between the two methods
x = 21

# The number of reported issues
n = 165

# The observed proportion
phat = x/n

# The estimated variance
```

```
varhat = phat*(1-phat) / n

# The estimated standard deviation
np.sqrt(varhat)

0.025945675200460937
```

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

Continue on page 30

Exercise IX

A researcher designs an experiment to estimate the mean time of a chemical reaction. The experiment should be able to detect a difference in mean of 0.5 seconds assuming a standard deviation of 1.2 seconds for the reaction time (previous studies suggest this is the best guess of σ). The researcher would like to use a 95% confidence level and a power of 80% for this experiment.

Question IX.1 (23)

What is the minimum sample size required to achieve the desired specifications? (Note that the facit is calculated with Python. Thus, facit can differ slightly from results obtained with formulas.)

1 □ 38

2 □ 47.17

3* □ 48

4 □ 63

5 □ 79

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

This is a power and sample size calculation for a setup with one sample. The required code can thus be found in chapter 3.3.2 of the book:

```
import statsmodels.stats.power as smp

# The sample size for power=0.80
delta = 0.5
sd = 1.2
alpha = 0.05
power = 0.8
smp.TTestPower().solve_power(effect_size=delta/sd, alpha=alpha,
power=power)

47.16622572477531
```

Round-up the value 47.17 gives 48, which is the minimum sample size required.

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

Continue on page 31

Exercise X

In an experiment, $n = 85$ individuals are tested for a binary response (positive or negative). Let the binomially distributed random variable

$$X \sim \text{Bin}(n, p)$$

denote the number of positive outcomes.

The following null hypothesis is considered

$$H_0 : p = 0.2.$$

Question X.1 (24)

Under the null hypothesis (i.e., assuming it is true), what is the probability of observing 25 or more positive outcomes?

1* □ 0.025

2 □ 0.030

3 □ 0.061

4 □ 0.132

5 □ 0.987

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

The probability of observing 25 or more positive outcomes in a binomial distribution:

$$P(X \geq 25) = 1 - P(X \leq 24) = 1 - F(24)$$

with $p = 0.2$ is calculated by:

```
1-stats.binom.cdf(24,n=85,p=0.2)

0.024703301590213167
```

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

Question X.2 (25)

If 25 positive outcomes are actually observed, what is the observed test statistic in the usual test of the null hypothesis?

1 □ 8.000

2* □ 2.169

3 □ 1.904

4 □ 0.057

5 □ 0.030

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

The test statistic for a hypothesis test with one proportion is described in method 7.11:

$$z_{obs} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{25 - 85 \cdot 0.2}{\sqrt{85 \cdot 0.2 \cdot 0.8}} = 2.169.$$

In Python, the hypothesis can be conducted using the code:

```
z_obs,p_value = smprop.proportions_ztest(25, 85, value=0.2, prop_var=0.2)
print(z_obs)

2.1693045781865616

print(p_value)

0.030059567892412428
```

Notice that the p-value differs from the one calculated in the former question. This is due to the fact that method 7.11 invokes a normal approximation, whereas the quantity calculated before is an exact probability.

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

Continue on page 33

Exercise XI

A study is conducted to compare the average scores of two groups of students in a math test. Group A consists of 12 students with a mean score of 78 and a standard deviation of 10, while Group B consists of 15 students with a mean score of 82 and a standard deviation of 8. Researchers want to perform a two-sample t-test to determine if there is a significant difference between the means of the two groups at a 5% significance level i.e., they want to test the null hypothesis

$$H_0 : \mu_A - \mu_B = 0$$

against a two-sided alternative hypothesis.

Question XI.1 (26)

What is the observed test statistic in the appropriate test?

1 □ -3.422

2 □ -1.155

3* □ -1.127

4 □ -0.765

5 □ -0.317

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

The appropriate test is a Welch two-sample t-test, see method 3.49 (note that there is a print error in the book). The observed test statistic is thus calculated as

$$t_{obs} = \frac{\bar{x}_A - \bar{x}_B - \delta_0}{\sqrt{s_A^2/n_A + s_B^2/n_B}} = \frac{78 - 82 - 0}{\sqrt{10^2/12 + 8^2/15}} = -1.127.$$

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

Question XI.2 (27)

The test statistic follows a t-distribution with how many degrees of freedom?

1* □ 20.853

2 □ 22.382

3 □ 24.998

4 □ 25

5 □ 26

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

Theorem 3.50 states the distribution of the test statistic, and equation (3-50) gives the formula for the degrees of freedom of the t-distribution. In Python, this is calculated as

```
# Variances
vs = np.array([10**2,8**2])
# Sample sizes
ns = np.array([12,15])

# The degrees of freedom
nu = ((vs[0]/ns[0]+vs[1]/ns[1])**2)/((vs[0]/ns[0])**2/(ns[0]-1)
+(vs[1]/ns[1])**2/(ns[1]-1))

print(nu)
```

```
20.852573482028465
```

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

Continue on page 35

Exercise XII

Consider the random variable Y and the sample $\mathbf{y} = (6, 5, 4, 4, 1, 10, 7, 6, 5, 3)$.

Question XII.1 (28)

Using non-parametric bootstrapping with 100,000 replications (bootstrap samples), what is the 90% confidence interval for the mean of Y? (Note: The facit uses `np.random.seed(2025)` and your answer may vary according to the seed.)

1 □ [3.7,6.6]

2* □ [3.9,6.3]

3 □ [4.2,6.0]

4 □ [5.0,5.2]

5 □ [5.1,5.1]

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

According to section 4.3.2, the one-sample non-parametric bootstrap confidence interval is found as:

```
# Set seed
np.random.seed(2025)

# Data
y = np.array([6, 5, 4, 4, 1, 10, 7, 6, 5, 3])

# Number of simulations/replications (bootstrap samples)
k = 100000

# Simulate k bootstrap samples with replacement
simsamples = np.random.choice(y,size=(k,len(y)))

# Compute the mean in each of the k bootstrap samples
simmeans = simsamples.mean(axis=1)

# Find the two relevant quantiles of the k generated means
print(np.quantile(simmeans,[0.05,0.95],
method='averaged_inverted_cdf'))

[3.9 6.3]
```

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

Question XII.2 (29)

Assume that Y follows a Poisson distribution. What is the 90% confidence interval for the mean of Y using parametric bootstrapping with 100,000 replications (bootstrap samples)? (Note: The facit uses `np.random.seed(2025)` and your answer may vary according to the seed.)

1 □ [3.7,6.6]

2* □ [4.0,6.3]

3 □ [4.2,6.0]

4 □ [5.0,5.2]

5 □ [5.1,5.1]

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

According to section 4.2.2, the one-sample parametric bootstrap confidence interval is found as:

```
# Set seed
np.random.seed(2025)

# Data
y = np.array([6, 5, 4, 4, 1, 10, 7, 6, 5, 3])

# Parameter estimate
mu = y.mean()

# Number of simulations/replications (bootstrap samples)
k = 100000

# Simulate k bootstrap samples from a Poisson distribution
simsamples = stats.poisson.rvs(mu,size=(k,len(y)))

# Compute the mean in each of the k bootstrap samples
simmeans = simsamples.mean(axis=1)

# Find the two relevant quantiles of the k generated means
print(np.quantile(simmeans,[0.05,0.95],
method='averaged_inverted_cdf'))
```

```
[4.  6.3]
```

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

The following chunk of code has been run:

```
# Set seed
np.random.seed(2025)

# Data
y = np.array([6, 5, 4, 4, 1, 10, 7, 6, 5, 3])

# Estimated mean
mu = y.mean()

# Simulate bootstrap samples
simsamples = stats.poisson.rvs(mu,size=(1000,len(y)))
simsamples = pd.DataFrame(simsamples)

# Compute medians
simmedians = simsamples.median(axis=1)

# Find confidence interval
print(np.quantile(simmedians,[0.05,0.95],
method=’averaged_inverted_cdf’))

[3.5 6.5]
```

Question XII.3 (30)

Consider the null hypothesis $\mathcal{H}_0$: median$(Y) = 5$. What is the correct conclusion about $\mathcal{H}_0$ based on the code chunk?

1 □ Based on the parametric bootstrapping procedure, $\mathcal{H}_0$ cannot be rejected on a significance level of 90%.

2 □ Based on the parametric bootstrapping procedure, $\mathcal{H}_0$ is rejected on a significance level of 10%.

3* □ Based on the parametric bootstrapping procedure, $\mathcal{H}_0$ cannot be rejected on a significance level of 10%.

4 □ Based on the non-parametric bootstrapping procedure, $\mathcal{H}_0$ is rejected on a significance level of 5%.

5 □ Based on the non-parametric bootstrapping procedure, $\mathcal{H}_0$ cannot be rejected on a significance level of 10%.

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

Since the bootstrap samples are generated via sampling from a Poisson distribution, we have made an assumption on the distribution of the data, and the code thus runs a parametric bootstrapping procedure. Since the code uses the 5% and 95% quantiles, the code returns a 90% confidence interval, i.e. the applied significance level is 10%. Finally, the null hypothesis cannot be rejected at this significance level as the produced confidence interval contains the value 5.

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

The exam is finished. Enjoy the vacation!