Written examination: 15. Dec. 2024

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

Duration: 4 hours

Aids and facilities allowed: All, except access to Internet

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 ultimately determined during censoring.

The final answers should be given by filling in and submitting the 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										

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

Exercise	X.1	X.2	XI.1	XI.2	XII.1	XII.2	XII.3	XII.4	XII.5	XII.6
Question	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Answer										

The exam paper contains 27 pages.

Continue on page 2

Using Python in this exam: *This version is the Python-version of the exam. A version using R is also available.*
Note that we use the following libraries and abbreviations in all Python code in this exam. We recommend that you copy paste this into your own script.

```
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
```

Please be aware that certain characters ("~", "_", "^", etc.) may not transfer correctly if you choose to copy paste from the exam template. If you get error messages please check that all the special characters are correctly typed into your code (you may need to re-type manually).

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 corresponding built-in functions in Python.*

Exercise I

A team of researchers evaluate a deterministic simulation model by comparing the model simulations with experimental results. The researchers consider two factors: load (kg) and velocity (knots). The researchers propose the following model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},$$

where the errors are assumed to be independent and normally distributed with $E[\varepsilon_{ij}] = 0$ and $V[\varepsilon_{ij}] = \sigma^2$. In the model, Y_{ij} is the difference between the simulated and experimental results obtained using load level i and velocity level j, and consequently the parameters α_i and β_j refer to the load and velocity effects, respectively. The table below displays the obtained differences (experimental result minus simulation result):

	5 knots	10 knots	25 knots	50 knots
100 kg	-33.72	-26.95	29.11	-38.87
200 kg	-5.75	-3.00	-15.41	20.56
300 kg	29.96	-24.77	-12.05	1.52
400 kg	-4.72	5.72	24.39	43.16
500 kg	-22.36	23.99	-24.17	33.36

The data can be read into Python using the code chunk:

```python
df = pd.DataFrame({
'y': [-33.72, -26.95,  29.11, -38.87,
       -5.75,  -3.00, -15.41,  20.56,
       29.96, -24.77, -12.05,   1.52,
       -4.72,   5.72,  24.39,  43.16,
      -22.36,  23.99, -24.17,  33.36],
'knots': pd.Categorical([5, 10, 25, 50,
                         5, 10, 25, 50,
                         5, 10, 25, 50,
                         5, 10, 25, 50,
                         5, 10, 25, 50]),
'load': pd.Categorical([100, 100, 100, 100,
                        200, 200, 200, 200,
                        300, 300, 300, 300,
                        400, 400, 400, 400,
                        500, 500, 500, 500]),
})
```

Question I.1 (1)

What is the parameter estimate $\hat{\alpha}_3$ (i.e. for load level "300 kg")?

1 ☐ -1.335

2 ☐ -0.900

3 ☐ 0.374

4 ☐ 2.705

5 ☐ 17.138

Question I.2 (2)

According to the model, SS(load) is 2454.51, SS(velocity) is 1107.10, and the total sum of squares is 11867.74. What is the residual mean square (MSE)?

1 ☐ 415.3

2 ☐ 692.2

3 ☐ 2076.5

4 ☐ 2768.7

5 ☐ 8306.1

Question I.3 (3)

The researchers discard the experimental results due to a technical error. When they repeat the experiment, they find the parameter estimates given below:

Parameter	α_1	α_2	α_3	α_4	α_5
Estimate	1.00	2.00	3.00	4.00	5.00

Parameter	β_1	β_2	β_3	β_4	μ
Estimate	0.25	1.00	3.13	5.00	0.00

What is MS(load) according to the new parameter estimates?

1 ☐ 13.75

2 ☐ 35.83

3 ☐ 55.00

4 ☐ 220.00

5 ☐ The quantity cannot be determined without knowing the complete data set.

Continue on page 5

Exercise II

In a pass/fail course, a class of $n = 30$ students was evaluated, with the results presented below. A score of 0 indicates 'failed' and a score of 1 indicates 'passed'.

1	0	1	1	1	0	1	1	1	1
0	0	0	0	1	1	0	1	1	1
1	1	1	1	1	1	0	0	1	1

The data can be read into Python using the code chunk:

```
data = np.array([1,0,1,0,0,1,1,0,1,1,0,1,1,1,1,0,1,1,1,0,0,1,1,0,1,1,1,1,1,1])
```

Question II.1 (4)

What is the estimated probability of passing the course and its 95% confidence interval, assuming the usual assumptions are satisfied (Note: The result is based on the equation given in the textbook, but confidence intervals calculated using in-built functions in Python, may give slightly different results).

1 ☐ $\hat{p} = 0.70$ and $[0.49, 0.91]$

2 ☐ $\hat{p} = 0.70$ and $[0.54, 0.86]$

3 ☐ $\hat{p} = 0.76$ and $[0.57, 0.95]$

4 ☐ $\hat{p} = 0.70$ and $[0.61, 0.79]$

5 ☐ $\hat{p} = 0.76$ and $[0.51, 0.89]$

Question II.2 (5)

What is the standard error of $\hat{p}$ if the "Plus 2" approach is used in the calculation of the confidence interval?

1 ☐ $\hat{\sigma}_{\hat{p}} = 0.0786$

2 ☐ $\hat{\sigma}_{\hat{p}} = 0.0802$

3 ☐ $\hat{\sigma}_{\hat{p}} = 0.0868$

4 ☐ $\hat{\sigma}_{\hat{p}} = 0.0883$

5 ☐ $\hat{\sigma}_{\hat{p}} = 0.0918$

Continue on page 6

Exercise III

In a study examining the difference in taste between regular and decaffeinated coffee, a taster has 4 cups containing coffee. Each cup contains either regular or decaffeinated coffee. The taster knows that there are two cups of each. The taster chose two cups at random.

Question III.1 (6)

What is the probability that the taster selected regular coffee in one of the cups and decaffeinated coffee in the other one (not taking into account the order of which they were chosen)?

1 ☐ 1/4

2 ☐ 1/3

3 ☐ 1/2

4 ☐ 2/3

5 ☐ 3/4

Question III.2 (7)

In another study examining the ability to detect the difference between regular and decaffeinated coffee, 30 participants are given a cup of each type to taste. Past studies suggest a 85% probability ($p = 0.85$) that individuals can detect the difference between regular and decaffeinated. Let Y represent the number of participants out of 30 who can differentiate between the two types. What is the variance of Y?

1 ☐ $\mathrm{V}(Y) = 5.37$

2 ☐ $\mathrm{V}(Y) = 3.83$

3 ☐ $\mathrm{V}(Y) = 3.11$

4 ☐ $\mathrm{V}(Y) = 2.79$

5 ☐ $\mathrm{V}(Y) = 1.10$

Continue on page 7

Exercise IV

The lifetime of a certain type of battery, measured in years, follows an exponential distribution with a mean of 50 years.

Question IV.1 (8)

What is the probability that a battery will last less than 25 years?

1 ☐ $e^{-\frac{25}{50}}$

2 ☐ $1 - e^{-\frac{50}{25}}$

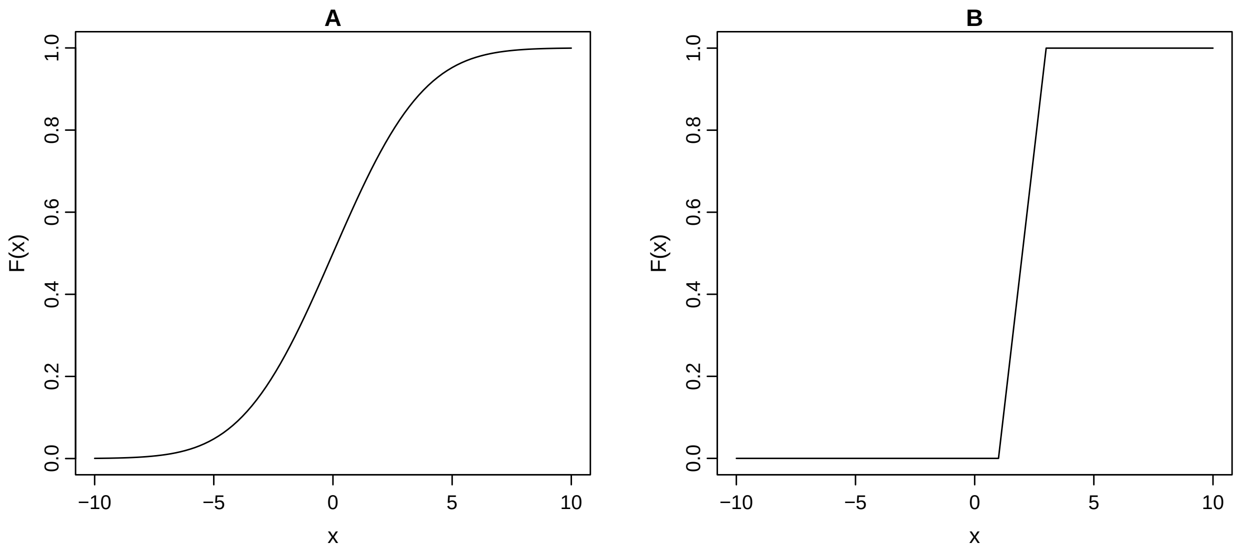
3 ☐ $e^{-\frac{50}{25}}$

4 ☐ $1 - e^{-\frac{25}{50}}$

5 ☐ $e^{-\frac{25}{50}} - e^{-\frac{50}{25}}$

Question IV.2 (9)

Below are two plots: one is a normal distribution CDF and the other is a uniform distribution CDF.



One of the statements is correct, judging from the plots, which one is that?

1 ☐ Plot A is a uniform distribution CDF with $a = 3$ and $b = 1$. Plot B is a normal distribution CDF with $\mu = -5$ and $\sigma = 10$.

2 ☐ Plot A is a uniform distribution CDF with $\mu = -5$ and $\sigma = 10$. Plot B is a normal distribution CDF with $a = 3$ and $b = 1$.

3 ☐ Plot A is a normal distribution CDF with $\mu = -5$ and $\sigma = 10$. Plot B is a uniform distribution CDF with $a = 3$ and $b = 1$.

4 ☐ Plot A is a normal distribution CDF with $\mu = 7$ and $\sigma = 1$. Plot B is a uniform distribution CDF with $a = -5$ and $b = 5$.

5 ☐ Plot A is a normal distribution CDF with $\mu = 0$ and $\sigma = 3$. Plot B is a uniform distribution CDF with $a = 1$ and $b = 3$.

Continue on page 9

Exercise V

In an agricultural study, researchers are investigating the effectiveness of two different fertilizers, A and B, on increasing crop yield. They randomly select 20 plots of land and apply Fertilizer A to 10 plots and Fertilizer B to the remaining 10 plots. After the harvest, they record the yield (in units "bushels per acre" $= 6.73g/m^2$) from each plot. The researchers want to determine if there is a significant difference in the mean yield between the two fertilizers.

Yield data is recorded as follows:

$\text{Fertilizer}_A : 45, 48, 50, 42, 47, 49, 43, 44, 46, 41$
$\text{Fertilizer}_B : 51, 53, 52, 50, 55, 48, 54, 49, 56, 52$

All the measurements are assumed to be independent and the yield populations follow normal distributions.

Question V.1 (10)

What is the test statistic and 99% confidence interval for the difference in mean crop yield between fertilizers (fertilizer A minus fertilizer B) (both results must be correct)?

1 ☐ -5.17, [-8.39, -4.61]

2 ☐ -4.17, [-8.68, -4.32]

3 ☐ -5.76, [-9.14, -3.85]

4 ☐ -5.17, [-9.15, -3.85]

5 ☐ -5.17, [-10.13, -2.87]

Question V.2 (11)

Denoting the mean yield for fertilizer A as μ_A and the mean yield for fertilizer B as μ_B, what should be the conclusion for the following null hypothesis

$$H_0 : \mu_\text{A} - \mu_\text{B} = 0$$

on significance level $\alpha = 0.05$ (both conclusion and argument must be correct)?

1 ☐ The null hypothesis is accepted, since the p-value is 0.23.

2 ☐ The null hypothesis is rejected, since the p-value is 0.0023.

3 ☐ The null hypothesis is rejected, since the 95% confidence interval contains zero.

4 ☐ The null hypothesis is rejected, since the 99% confidence interval contains zero.

5 ☐ The null hypothesis is rejected, since the 95% confidence interval does not contain zero.

Continue on page 10

Exercise VI

A toy shop sells marbles made of glass. The marbles are approximately the same size with mean diameter (D) 1 cm, but the variance is only stated in terms of weight (W): $\sigma_W^2 = 0.03^2$. The marble weights follow a normal distribution.

The expression relating weight to diameter is

$$W = \rho \cdot \frac{4}{3} \cdot \pi \cdot \left(\frac{D}{2}\right)^3$$

and therefore the expression relating diameter to weight is

$$D = 2\left(\frac{3W}{4\pi\rho}\right)^{1/3}$$

Where $\rho = 2.6$ g/cm^3 is the density (equal to the density of glass). You can use $\pi = 3.14$, and $\mu_W = W(\mu_D)$.

Question VI.1 (12)

A customer wants to know the standard deviation of the diameter of the marbles (σ_D). Luckily, the customer has studied error propagation and knows how to approximate σ_D from σ_W. What is the standard deviation of the diameters of the marbles?

1 ☐ $\sigma_D = 0.006$ cm

2 ☐ $\sigma_D = 0.086$ cm

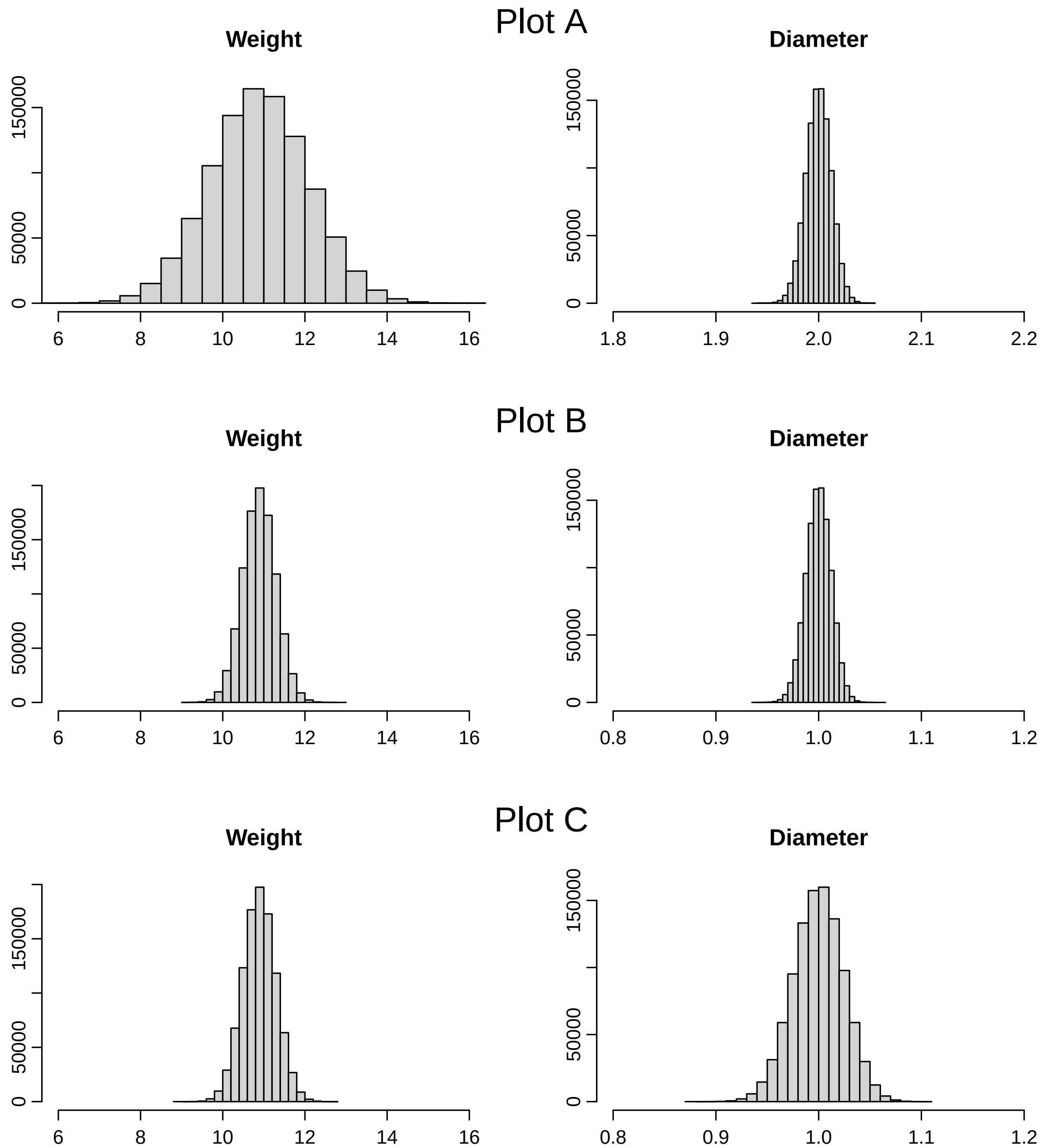
3 ☐ $\sigma_D = 0.015$ cm

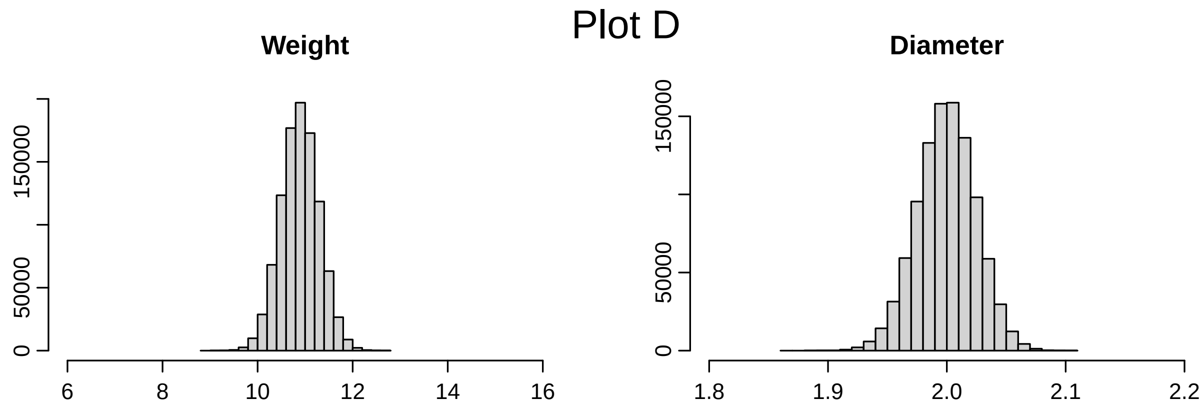
4 ☐ $\sigma_D = 0.007$ cm

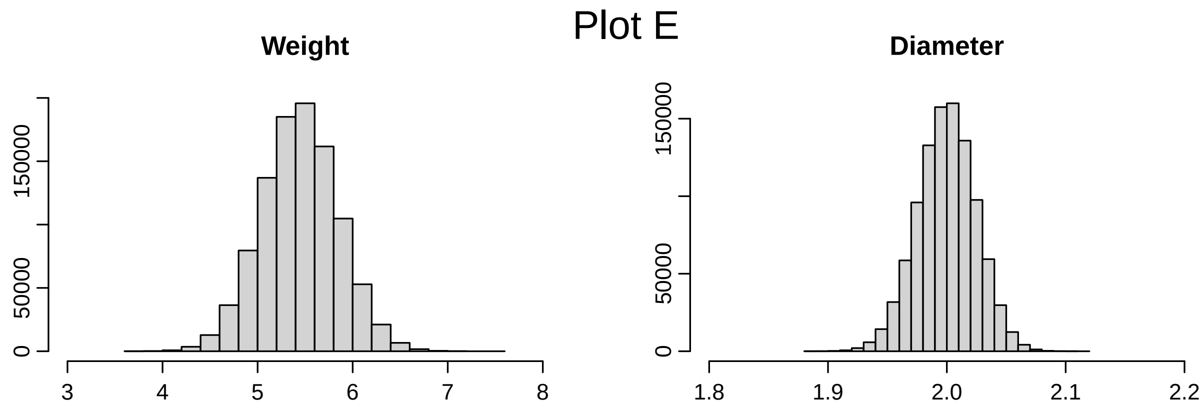
5 ☐ $\sigma_D = 0.04$ cm

Question VI.2 (13)

Another brand of marbles has mean diameter of 2 cm, $\sigma_W^2 = 0.4^2$ and density $\rho = 2.6$ g/cm^3. The weight of each marble may be calculated as $W = \rho \cdot \frac{4}{3} \cdot \pi \cdot \left(\frac{D}{2}\right)^3$. Which set of histograms matches the marbles from this other brand?


Plot A
Weight
Diameter
0
50000
150000
6
8
10
12
14
16
1.8
1.9
2.0
2.1
2.2
Plot B
Weight
Diameter
0.8
0.9
1.0
1.1
1.2
Plot C
Weight
Diameter





1 ☐ Plot A

2 ☐ Plot B

3 ☐ Plot C

4 ☐ Plot D

5 ☐ Plot E

Question VI.3 (14)

The toy shop gets new marbles delivered every month. Sometimes they find that some of the marbles are broken. The owner of the toy shop decides to take note of the deliveries that contain broken marbles. The time (measured in months) between deliveries containing broken marbles is stored in the variable x.

```
x = np.array([13, 4, 1, 17, 11, 2, 24, 25, 8, 4, 7, 7, 5, 6, 2, 13, 16, 3, 9, 11])
```

Use the book´s definition of sample quantiles to determine the IQR (*"Inter Quartile Range"*).

1 ☐ IQR = 4

2 ☐ IQR = 7.5

3 ☐ IQR = 9

4 ☐ IQR = 11

5 ☐ IQR = 12

Question VI.4 (15)

The owner of the toy shop decides that they will stop buying marbles from the given vendor if the incidents of deliveries with broken marbles becomes too frequent. Without assuming any distribution of time between deliveries with broken marbles, the owner of the toy shop makes a non-parametric 95% bootstrap confidence interval for the median time between such events. Which of the following Python codes calculates this confidence interval for the median correctly?

1 ☐
```
simsamples = np.random.choice(x, size=(10000, len(x)))
medians = np.median(simsamples, axis=1)
quantiles = np.quantile(medians, [0.05, 0.95], method="averaged_inverted_cdf")
print(quantiles)
```

2 ☐
```
simsamples = np.random.choice(x, size=(10000, len(x)))
medians = np.median(simsamples, axis=1)
quantiles = np.quantile(medians, [0.025, 0.975], method="averaged_inverted_cdf")
print(quantiles)
```

3 ☐
```
simsamples = stats.expon.rvs(size=(10000, len(x)), scale=np.mean(x))
medians = np.median(simsamples, axis=1)
quantiles = np.quantile(medians, [0.05, 0.95], method="averaged_inverted_cdf")
print(quantiles)
```

4 ☐
```
simsamples = stats.expon.rvs(size=(10000, len(x)), scale=np.mean(x))
medians = np.median(simsamples, axis=1)
quantiles = np.quantile(medians, [0.025, 0.975], method="averaged_inverted_cdf")
print(quantiles)
```

5 ☐
```
simsamples = np.random.choice(x, size=(10000, len(x)))
medians = np.median(simsamples, axis=1)
quantiles = np.quantile(medians, [0.005, 0.995], method="averaged_inverted_cdf")
print(quantiles)
```

Question VI.5 (16)

After some time the vendor makes an effort to increase the quality of the marbles by manually removing bags with broken marbles. Again the owner of the toy shop decides to take note of the deliveries that contain broken marbles. The time (measured in months) between deliveries containing broken marbles is stored in the variable y.

```
y = np.array([3,2,1,14,23,38,25,4,14,28,6,34,5,25,17,20,11,19,4,9])
```

Hereafter the following calculations are made in order to test whether the new effort to increase quality has made any difference:

```
simXsamples = stats.expon.rvs(size=(10000, len(x)), scale=np.mean(x))

simYsamples = stats.expon.rvs(size=(10000, len(y)), scale=np.mean(y))

simDiff = np.median(simXsamples, axis=1) - np.median(simYsamples, axis=1)



print(np.percentile(simDiff, [0.5, 99.5], method="averaged_inverted_cdf"))


[-15.8691798    5.10555541]

print(np.percentile(simDiff, [2.5, 97.5], method="averaged_inverted_cdf"))


[-12.40129205   3.13755755]

print(np.percentile(simDiff, [5, 95], method="averaged_inverted_cdf"))


[-10.80677812   1.87399112]
```

Which of the following statements is correct?

1 ☐ The analysis makes no assumptions about the distributions of x and y. At $\alpha = 1\%$ significance level it may be concluded that there is <u>no</u> significant difference in medians.

2 ☐ The analysis assumes that both x and y are normally distributed. At $\alpha = 5\%$ significance level it may be concluded that there is <u>no</u> significant difference in medians.

3 ☐ The analysis assumes that both x and y are exponentially distributed. At $\alpha = 1\%$ significance level it may be concluded that there is a significant difference in medians.

4 ☐ The analysis makes no assumptions about the distributions of x and y. At $\alpha = 5\%$ significance level it may be concluded that there is a significant difference in medians.

5 ☐ None of the statements above are correct.

Continue on page 17

Exercise VII

Suppose we have collected exam scores from two groups:

Group 1: 82, 91, 85, 89, 88

Group 2: 76, 84, 80, 82, 83

We assume that the exam scores follow normal distributions with equal variances. Additionally, we assume that the exam scores can be considered independent and identically distributed (i.i.d.), within each group.

Question VII.1 (17)

What is the estimate of the pooled variance?

1 ☐ 9.00

2 ☐ 27.10

3 ☐ 11.25

4 ☐ 10.00

5 ☐ 8.00

Question VII.2 (18)

What is the minimum number of observations required in each group (same number of observations in both groups) to achieve a power of 99% for detecting a difference in means of at least 4 between the two groups, assuming the variance is 20 (equal variances in both groups) and a significance level of 1%?

1 ☐ At least 56 (or 55, depending on calculation method)

2 ☐ At least 39 (or 38, depending on calculation method)

3 ☐ At least 62 (or 61, depending on calculation method)

4 ☐ At least 32 (or 31, depending on calculation method)

5 ☐ At least 82 (or 79, depending on calculation method)

Continue on page 18

Exercise VIII

In preparation for a conference, organizers need to plan coffee breaks efficiently. They estimate that the number of attendees needing coffee will follow a Poisson distribution and that, on average, 200 attendees will need coffee every hour. The organizers set up enough coffee stations to serve 240 attendees per hour.

Question VIII.1 (19)

What is the probability that, during a randomly selected hour, the number of attendees needing coffee exceeds the capacity?

1 ☐ 0.0027

2 ☐ 0.023

3 ☐ 0.11

4 ☐ 0.24

5 ☐ 0.0045

Continue on page 19

Exercise IX

Let $X_i \sim N(0, \sigma^2)$ be i.i.d. random variables and define

$$Q = \frac{1}{n} \sum_{i=1}^{n} X_i^2.$$

Question IX.1 (20)

For $0 < \alpha < 1$, for what q do we have $P(Q > q) = 1 - \alpha$?

1 ☐ $q = \frac{\sigma^2}{n} \chi^2_\alpha$, where χ^2_α is the α quantile of a χ^2-distribution with n degrees of freedom.

2 ☐ $q = \frac{\sigma^2}{n} \chi^2_{1-\alpha}$, where $\chi^2_{1-\alpha}$ is the $(1-\alpha)$ quantile of a χ^2-distribution with n degrees of freedom.

3 ☐ $q = \frac{\sigma^2}{n-1} \chi^2_\alpha$, where χ^2_α is the α quantile of a χ^2-distribution with n degrees of freedom.

4 ☐ $q = \frac{\sigma^2}{1-n} \chi^2_{1-\alpha}$, where $\chi^2_{1-\alpha}$ is the $(1-\alpha)$ quantile of a χ^2-distribution with $(n-1)$ degrees of freedom.

5 ☐ $q = \frac{\sigma^2}{n} \chi^2_\alpha$, where χ^2_α is the α quantile of a χ^2-distribution with $(n-1)$ degrees of freedom.

Continue on page 20

Exercise X

A technology company has recorded its monthly sales figures over a period of three years (36 months). The monthly sales numbers are summarized in the below table showing the average monthly sales and the sample standard deviation of the monthly sales for each of the three years.

Year	2021	2022	2023
Average monthly sales (M DKK)	391.2	402.5	429.4
Standard deviation of monthly sales (M DKK)	22.3	27.5	26.7

The engineers at the company then formulates a one-way ANOVA model for the data using the monthly sales as the response variable and the year as *the treatment*.

Question X.1 (21)

In the ANOVA model, what is the residual mean square (MSE)?

1 ☐ 25.50

2 ☐ 162.56

3 ☐ 407.70

4 ☐ 655.48

5 ☐ 1966.43

Question X.2 (22)

The engineers pre-planned to calculate pairwise confidence intervals for $\mu_{2022} - \mu_{2021}$ and $\mu_{2023} - \mu_{2022}$ using an overall significance level of 10%, where μ_i refers to the mean monthly sales for year i. Which quantile from the t-distribution must be used in the calculations of the confidence intervals?

1 ☐ The 90% quantile of the t-distribution with 33 degrees of freedom

2 ☐ The 95% quantile of the t-distribution with 33 degrees of freedom

3 ☐ The 95% quantile of the t-distribution with 34 degrees of freedom

4 ☐ The 97.5% quantile of the t-distribution with 33 degrees of freedom

5 ☐ The 97.5% quantile of the t-distribution with 34 degrees of freedom

Continue on page 21

Exercise XI

To study crime in Denmark, researchers are interested in the number of individuals placed in pretrial detention after their arrest. These figures are recorded and available through Statistics Denmark. The annual counts from 2015 to 2022 are categorized into three age groups: "Young" (ages 15-29), "Mid-age" (ages 30-39), and "Old" (ages 40 and above). The data is read into Python using the following code:

```
tbl = np.array([[2048, 1072, 821],

                [2208, 998, 836],

                [2359, 1092, 853],

                [2138, 1093, 880],

                [1984, 935, 799],

                [1777, 872, 860],

                [1604, 818, 729],

                [1564, 943, 753]])

tbl = pd.DataFrame(tbl,

                   index=['2015','2016','2017','2018',

                          '2019','2020','2021','2022'],

                   columns=['Young','Midage','Old'])
```

Question XI.1 (23)

Consider the null hypothesis that the age distribution of individuals placed in pretrial detention does not change over the years. What is the result and conclusion of the appropriate test (both argument and conclusion must be correct)?

1 ☐ The p-value is 0.24 and the conclusion is that there is no significant change in distribution across the years.

2 ☐ The p-value is $0.24 \cdot 10^{-10}$ and the conclusion is that there is a significant change in distribution in every year across all years.

3 ☐ The p-value is $0.24 \cdot 10^{-10}$ and the conclusion is that there is a significant change in distribution at least in one of the years.

4 ☐ The p-value is $4.1 \cdot 10^{-10}$ and the conclusion is that there is a significant change in distribution in every year across all years.

5 ☐ The p-value is $4.1 \cdot 10^{-10}$ and the conclusion is that there is a significant change in distribution at least in one of the years.

Question XI.2 (24)

Under the null hypothesis of no change in distribution, what is the expected number of individuals placed in pretrial detention in the "Young" category if the total number of such placements in a specific year is 3000?

1 ☐ 978

2 ☐ 1364

3 ☐ 1566

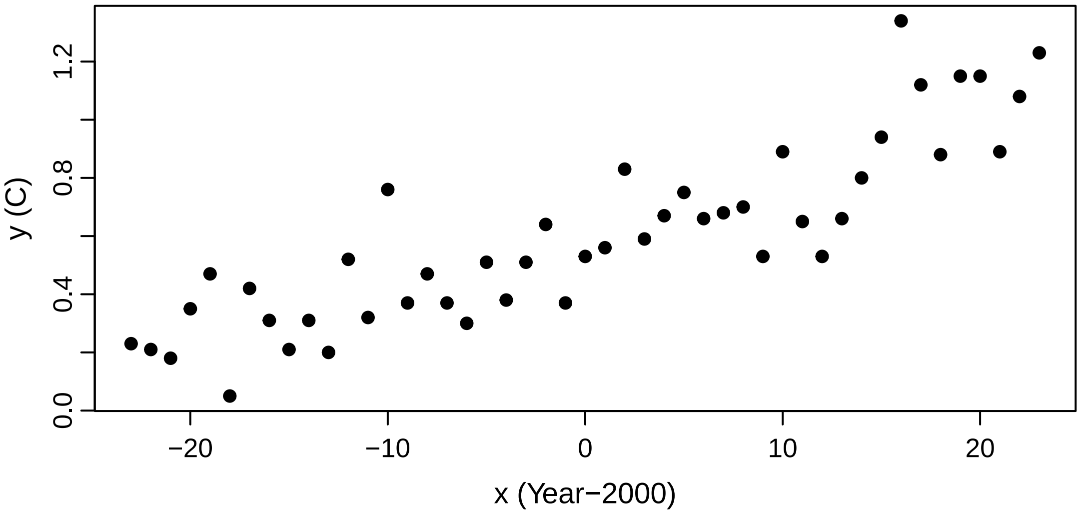
4 ☐ 1960

5 ☐ 2048

Continue on page 23

Exercise XII

The figure below shows the average global temperature anomaly, which is the temperature minus average over the period 1900-2000 in [°C] as a function of time. The period is the years 1977 to 2023 (the x-axes is Year-2000).



As a first approach a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

is fitted to the data. In the model Y_i is the temperature anomaly and x_i is the year (minus 2000), of observation i. The result is given below:

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

print(fit.summary(slim=True))


                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.754
Model:                            OLS   Adj. R-squared:                  0.748
No. Observations:                  47   F-statistic:                     137.7
Covariance Type:            nonrobust   Prob (F-statistic):           2.76e-15
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.6015      0.023     26.639      0.000       0.556       0.647
x              0.0195      0.002     11.734      0.000       0.016       0.023
==============================================================================
```

```
print(round(np.sqrt(fit.scale),4))
```

```
0.1548
```

```
print(fit.pvalues)
```

```
Intercept    3.420248e-29
x            2.758270e-15
dtype: float64
```

Question XII.1 (25)

Which of the following statements about the assumptions of the model is not correct?

1 ☐ $\varepsilon_i \sim N(0, \sigma^2)$.

2 ☐ ε_i and ε_j are independent for $i \neq j$.

3 ☐ $V(\varepsilon_i) = V(\varepsilon_j)$ for all (i, j).

4 ☐ Y_i and ε_i are independent.

5 ☐ $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

Question XII.2 (26)

What is the conclusion (using significance level $\alpha = 0.05$) for the relationship between time (in years) and temperature based on the model (both conclusion and argument must be correct)?

1 ☐ The temperature changes significantly with time (`x`), since $0.0195 < 0.05$.

2 ☐ The temperature changes significantly with time (`x`), since $0.002 < 0.05$.

3 ☐ Time (`x`) have a significant effect on the temperature, since $0.002 < 0.05$.

4 ☐ The temperature changes significantly with time (`x`), since $2.758 \cdot 10^{-15} < 0.05$.

5 ☐ The temperature is a function of time (`x`), since $0.0195 < 0.05$.

Continue on page 25

Question XII.3 (27)

The model can be written in matrix–vector notation as

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

where the first column in $\boldsymbol{X}$ is a vector of ones and the second column is the vector $[-23, -22, ..., 22, 23]^T$.

Based on the model what is the 95% confidence interval for global average temperature anomaly in 2100?

1 ☐ [2.09, 3.01]

2 ☐ [1.19, 3.92]

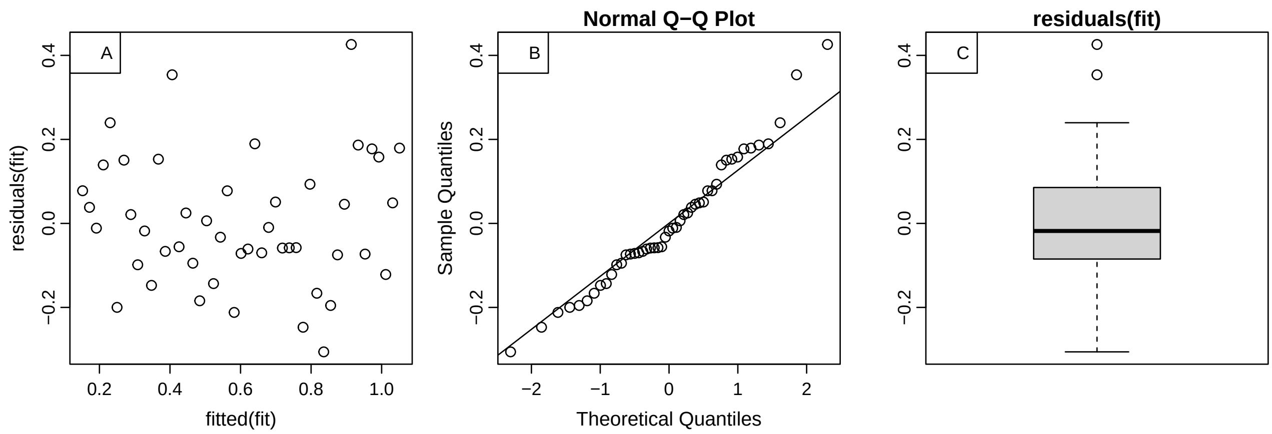
3 ☐ [2.51, 2.60]

4 ☐ [2.17, 2.94]

5 ☐ [2.22, 2.89]

Question XII.4 (28)

In order to validate the model the following residual plots have been created.



Based on the plots which of the following statements is correct (both the conclusion and the figure reference from which this can be concluded must be correct)?

1 ☐ The residuals seems to be independent, as seen on Plot B.

2 ☐ The residuals are clearly not identically distributed, as seen on Plot C.

3 ☐ There does not seem to be any systematic patterns in the residuals, as seen on Plot A.

4 ☐ There is clearly missing a quadratic term in the model, as seen on Plot C.

5 ☐ The variance homogeneity property is clearly violated, as seen on Plot B.

Question XII.5 (29)

Regardless of the conclusions in the previous questions, it is decided to fit a quadratic model

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

in the Python-code below `x2` represents `x`^{2}, further parts of the output from summary is removed, and some numbers are replaced by characters.

```
fit = smf.ols(formula = 'y ~ x + x2', data = dat).fit()
print(fit.summary(slim=True))

                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.779
Model:                            OLS   Adj. R-squared:                  0.769
No. Observations:                  47   F-statistic:                     77.49
Covariance Type:            nonrobust   Prob (F-statistic):           3.82e-15
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   0.5472833  0.0324647          A          D           -           -
x           0.0195317  0.0015949          B          E           -           -
x2          0.0002946  0.0001315          C          F           -           -
==============================================================================
```

In order to conclude if the quadratic term should be included in the model, which of the following conclusions is correct at a significance level $\alpha = 0.05$?

1 ☐ `C`=6.6 and $\hat{\beta}_2$ is significantly different from 0 as the critical value is 2.02.

2 ☐ `C`=2.2 and $\hat{\beta}_2$ is significantly different from 0 as the critical value is 2.02.

3 ☐ `B`=11.7 and $\hat{\beta}_1$ is significantly different from 0 as the critical value is 1.96.

4 ☐ `A`=26.6 and $\hat{\beta}_1$ is significantly different from 0 as the critical value is 1.96.

5 ☐ `C`=2.2 and $\hat{\beta}_2$ is <u>not</u> significantly different from 0 as the critical value is 2.02.

Continue on page 27

Question XII.6 (30)

The standard errors in the summary above can be obtained from the matrix $\mathbf{\Sigma}_\beta$, such that $(\mathbf{\Sigma}_\beta)_{11} = 0.0324647^2$, $(\mathbf{\Sigma}_\beta)_{22} = 0.0015949^2$, and $(\mathbf{\Sigma}_\beta)_{33} = 0.0001315^2$, but which elements of the matrix $\mathbf{\Sigma}_\beta$ is equal to zero?

1 ☐ $(\mathbf{\Sigma}_\beta)_{12}$, $(\mathbf{\Sigma}_\beta)_{21}$, $(\mathbf{\Sigma}_\beta)_{23}$, and $(\mathbf{\Sigma}_\beta)_{32}$

2 ☐ None

3 ☐ All but the diagonal elements

4 ☐ $(\mathbf{\Sigma}_\beta)_{12}$, $(\mathbf{\Sigma}_\beta)_{21}$, $(\mathbf{\Sigma}_\beta)_{13}$, and $(\mathbf{\Sigma}_\beta)_{31}$

5 ☐ $(\mathbf{\Sigma}_\beta)_{13}$, and $(\mathbf{\Sigma}_\beta)_{31}$

The exam is finished. Enjoy the vacation!