Written examination: 26.06.2025

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 15 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	IV.1	V.1	V.2	VI.1
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer	4	4	2	1	5	5	2	2	5	2

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

Exercise	X.1	XI.1	XI.2	XII.1	XII.2	XIII.1	XIV.1	XIV.2	XIV.3	XV.1
Question	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Answer	3	3	4	4	5	3	1	1	1	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 corresponding built-in functions in Python.*

Exercise I

Let X follow a normal distribution with mean 2 and variance 16.

Question I.1 (1)

What is the median of X?

1 □ median(X) = -4

2 □ median(X) = -2

3 □ median(X) = 0

4* □ median(X) = 2

5 □ median(X) = 4

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

The normal distribution is symmetric around its mean. Therefore, the median is exactly the mean value, which in this case is 2.

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

Question I.2 (2)

Which one of the following lines of code correctly calculates the probability $P(X \leq 3)$?

1 □ `stats.norm.pdf(3, loc = 2, scale = 16)`

2 □ `stats.norm.pdf(3, loc = 2, scale = 4)`

3 □ `stats.norm.cdf(3, loc = 2, scale = 16)`

4* □ `stats.norm.cdf(3, loc = 2, scale = 4)`

5 □ `1 - stats.norm.cdf(3, loc = 2, scale = 16)`

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

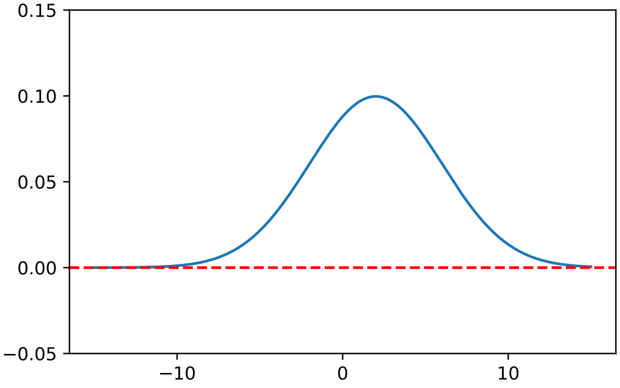
Note that $P(X \leq 3) = F(3)$, where F denotes the distribution function (the cdf) of X. Since the mean of X is 2, and the standard deviation of X is $\sqrt{16} = 4$, the location and scale arguments of the Python function should be 2 and 4, respectively.

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

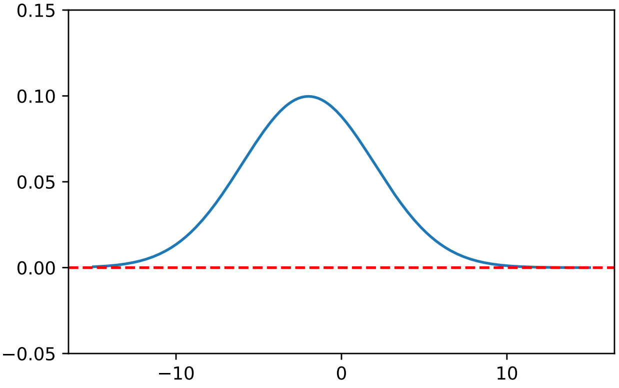
Question I.3 (3)

Let $Y = -X$. Which of the following plots shows the density function of Y?

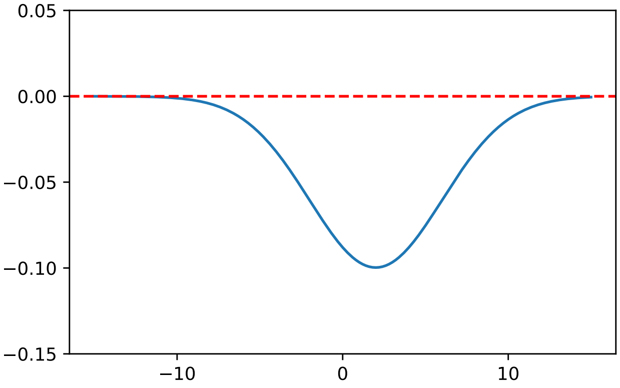
Plot A



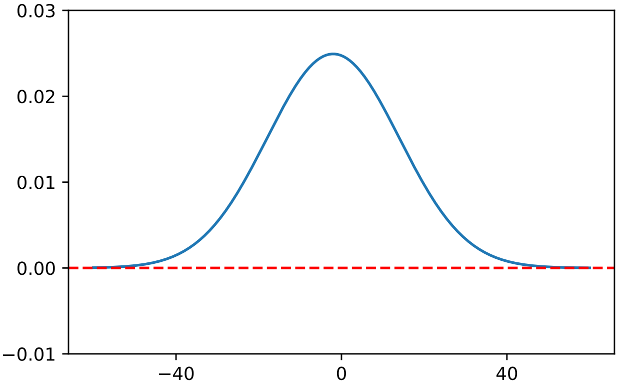
Plot B



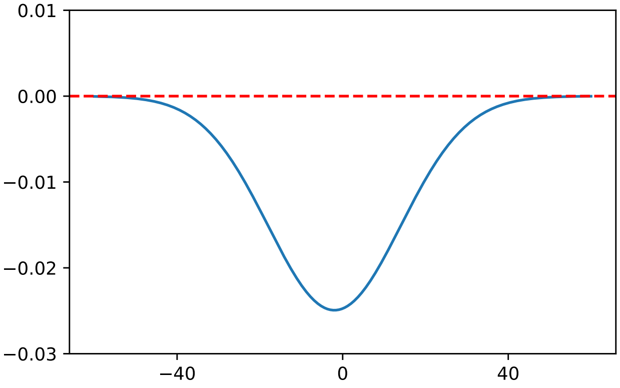
Plot C



Plot D



Plot E



1 □ Plot A

2* □ Plot B

3 □ Plot C

4 □ Plot D

5 □ Plot E

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

Plots C and E are not density functions, as densities are never negative. Plot A shows the density function of X i.e., a normal distribution with mean two and standard deviation four. Plot D shows the density function of a normal distribution with the correct mean of negative two, but with a standard deviation of 16. Finally, plot B shows the correct density function of a normal distribution with the correct mean and standard deviation.

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

Exercise II

An engineer wants to test whether a new alloy has a tensile strength with a mean value of 500 MPa. A random sample of 30 specimens is tested, which gives a sample mean of 510 MPa and a sample standard deviation of 20 MPa. It is assumed that the observations are iid and normally distributed.

Question II.1 (4)

What is the corresponding p-value for the relevant hypothesis test with the following hypotheses:

$$H_0 : \mu = 500, \quad H_A : \mu \neq 500.$$

1* □ $p = 0.010$

2 □ $p = 0.621$

3 □ $p = 0.310$

4 □ $p = 0.006$

5 □ $p = 0.005$

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

Vi bruger en tosidet t-test for middelværdien med følgende teststatistik:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{510 - 500}{20/\sqrt{30}}$$

Beregning af p-værdien i Python:

```
teststatistic = (510-500)/(20/30**0.5)
print(teststatistic)
```

```
2.7386127875258306
```

```
result = 2*stats.t.cdf(-teststatistic, df=29)
print(result)
```

```
0.010437389498867 33
```

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

The engineer is now planning a new experiment where he wants to achieve a ”margin of error” of at most 2 MPa. He uses the observed standard deviation as a scenario and a significance level of $\alpha = 0.05$

Question II.2 (5)

What sample size should be taken to achieve a ”margin of error” of at most 2 MPa?

1 □ approx. 20 observations

2 □ approx. 40 observations

3 □ approx. 1538 observations

4 □ approx. 16 observations

5* □ approx. 385 observations

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

```
(stats.norm.ppf(0.975) * 20 / 2)**2
```

```
384.14588206941244
```

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

Continue on page 7

Exercise III

A coach wants to investigate whether there is a difference between different types of targeted training in terms of improving the time it takes to run up stairs. The coach collects data from 15 participants, who are (randomly) divided into three equally sized groups: Group A, Group B, and Group C. The coach has the participants perform targeted exercises over the next 4 weeks. Participants in the same group do the same exercises, but the coach assigns different exercises to the three groups. For each participant, data is collected on the improvement in the time it takes them to run up a staircase at the gym (the time improvement is measured in seconds).

The observed time improvements are:

Group:	time improvement (measured in seconds):
A	2.1, 2.5, 2.3, 2.4, 2.2
B	2.8, 2.9, 2.7, 3.0, 2.6
C	2.3, 2.4, 2.5, 2.2, 2.1

The average time improvement for all 15 participants is $\hat{\mu} = 2.467$, and the average time improvements within each group are given by: $\hat{\mu}_A = 2.30$, $\hat{\mu}_B = 2.80$, $\hat{\mu}_C = 2.30$. It can be assumed that all observations are independent and normally distributed.

Question III.1 (6)

What is the most appropriate statistical model and analysis when one wishes to examine whether there is a difference in the effect of the different types of training?

1 □ An appropriate model could be $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ ($\epsilon_{ij} \sim N(0, \sigma^2)$), where Y_{ij} is the time improvement of person number j in group number i. A relevant analysis would then be to perform a t-test that tests the null hypothesis $H_0 : \mu = 0$.

2 □ An appropriate model could be $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ ($\epsilon_i \sim N(0, \sigma^2)$), where x_i is the time improvement of person number i. A relevant analysis would then be to perform a t-test that tests the null hypothesis $H_0 : \beta_1 = 0$.

3 □ An appropriate model could be $Y_{ij} = \mu_i + \epsilon_{ij}$ ($\epsilon_{ij} \sim N(0, \sigma^2)$), where Y_{ij} is the time improvement of person number j in group number i. A relevant analysis would then be to perform an analysis of variance that tests the null hypothesis $H_0 : \mu_A = \mu_B = \mu_C = 0$.

4 □ An appropriate model could be $Y_{ij} = \beta_0 + \beta_i x_{ij} + \epsilon_{ij}$ ($\epsilon_{ij} \sim N(0, \sigma^2)$), where x_{ij} is the time improvement of person number j in group number i. A relevant analysis would then be an analysis of variance that tests the null hypothesis $H_0 : \beta_i = 0$ (that is, a total of 3 tests are performed – one for each group).

5* □ An appropriate model could be $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ ($\epsilon_{ij} \sim N(0, \sigma^2)$), where Y_{ij} is the time improvement of person number j in group number i. A relevant analysis would then be an analysis of variance that tests the null hypothesis $H_0 : \alpha_A = \alpha_B = \alpha_C = 0$.

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

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

Exercise IV

Assume that Y follows an exponential distribution with $E(Y) = 3$.

Question IV.1 (7)

What is $P(2 < Y < 4)$?

1 □ 0.49

2* □ 0.25

3 □ 0.61

4 □ 0.75

5 □ 0.0024

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

The probability can be calculated by

$$P(2 < Y < 4) = P(Y < 4) - P(Y < 2) = F(4) - F(2) \tag{1}$$

and in Python by

```
stats.expon.cdf(4,scale=3) - stats.expon.cdf(2,scale=3)

0.2498199809168653
```

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

Continue on page 9

Exercise V

A pet store wants to investigate what proportion of Danish households have a dog. They conduct a survey among 1000 of their customers, asking whether they have a dog. The store assumes that these 1000 customers represent 1000 households.

Of these, 320 respond that they have a dog.

Question V.1 (8)

What is the estimated proportion ($\hat{p}$) of households that have a dog, and what is the uncertainty (standard error, $s.e._{\hat{p}}$) of this proportion?

1 □ $\hat{p} = 0.32$ and $s.e._{\hat{p}} = 0.00022$

2*□ $\hat{p} = 0.32$ and $s.e._{\hat{p}} = 0.015$

3 □ $\hat{p} = 0.32$ and $s.e._{\hat{p}} = 0.047$

4 □ $\hat{p} = 0.32$ and $s.e._{\hat{p}} = 0.32$

5 □ $\hat{p} = 0.32$ and $s.e._{\hat{p}} = 0.010$

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

```
ph = 320/1000
ph

0.32

np.sqrt(ph*(1-ph)/1000)

0.014751271131668619
```

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

Question V.2 (9)

Official figures indicate that about 20% of Danes have a dog. The pet store had therefore expected that their survey would result in a proportion closer to 0.20. Is it likely that their result – that as many as 32% of households have a dog – is due to random variation? And could it be true that the true proportion of Danish households with a dog is actually around 20%?

1 □ Yes, the pet store has randomly selected a sample where more than expected have a dog. This is likely due to random variation, and the true proportion could well be around 20%.

2 □ No, it is unlikely that the pet store's result is due to random variation. The p-value for the relevant test is 0.0015, so we would reject the null hypothesis that the true proportion is 0.20. Thus, we must conclude that the true proportion is probably not 20%.

3 □ No, it is unlikely that the pet store's result is due to random variation. The p-value for the relevant test is 0.0015, so we would reject the null hypothesis that the true proportion is 0.20. However, it is doubtful whether the sample is representative, so the true proportion could still be 20%.

4 □ No, it is unlikely that the pet store's result is due to random variation. The p-value for the relevant test is $2 \cdot 10^{-21}$, so we would reject the null hypothesis that the true proportion is 0.20. Since the sample is clearly representative, we must conclude that the true proportion of households with a dog is probably not 20%.

5*□ No, it is unlikely that the pet store's result is due to random variation. The p-value for the relevant test is $2 \cdot 10^{-21}$, so we would reject the null hypothesis that the true proportion is 0.20. However, it is doubtful whether the sample is representative, so the true proportion could still be 20%.

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

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

Continue on page 11

Exercise VI

In a study, data from 4 different groups are available:

Group 1:	89, 102, 94, 90, 100
Group 2:	78, 46, 65, 72, 69
Group 3:	83, 89, 81, 89, 90
Group 4:	82, 101, 93, 88, 104

The table can be entered into Python using the following code.

```
y = np.array([89, 102, 94, 90, 100,
              78, 46, 65, 72, 69,
              83, 89, 81, 89, 90,
              82, 101, 93, 88, 104])
Group = pd.Categorical([1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4,4,4])
D = pd.DataFrame({'y': y, 'Group': Group})
```

It can be assumed that the data can be described by the following model: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0, \sigma^2)$.

Question VI.1 (10)

What is the between group variation, MS(Group), and the within group variation, MSE?

1 □ MS(Group) = 190.3 and MSE = 1122

2*□ MS(Group) = 894.5 and MSE = 70.15

3 □ MS(Group) = 2683 and MSE = 1122

4 □ MS(Group) = 894.5 and MSE = 2683

5 □ MS(Group) = 190.3 and MSE = 70.15

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

```
fit = smf.ols('y ~ Group', data=D).fit()
anova = sm.stats.anova_lm(fit)
print(anova)

            df    sum_sq  mean_sq          F    PR(>F)
Group      3.0   2683.35   894.45  12.750535  0.000164
Residual  16.0   1122.40    70.15        NaN       NaN
```

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

Continue on page 12

Question VI.2 (11)

Which statement about the model above is NOT correct?

1 □ Y_{ij} is observation number j in group number i. $\hat{\alpha}_i$ is group i's average deviation from the overall mean $\hat{\mu}$.

2*□ The total variance of the data (i.e. $\frac{1}{N-1}SST$) cannot be greater than $\hat{\sigma}^2$.

3 □ MSE represents the variance within each group, and since we assume it is the same across all groups, we also have $MSE = \hat{\sigma}^2$.

4 □ If the variance of the α_i's is large compared to the MSE, this means that there is a difference between the groups.

5 □ In the data above (for Group 1, $i = 1$), we obtain $\hat{\alpha}_1 = 9.75$.

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

Continue on page 14

Exercise VII

Let $X \sim \mathcal{N}(50, 2^2)$ and $Y \sim \mathcal{N}(45, 5^2)$ be independent and form the random variables $V = 3X + Y$ and $U = \max(X, Y)$ i.e., U is the maximum of X and Y. Use simulation to answer the following three questions.

Question VII.1 (12)

What is the probability $P(Y > X)$?

1* □ Approximately 17.7%

2 □ Approximately 42.2%

3 □ Approximately 50.0%

4 □ Approximately 57.8%

5 □ Approximately 82.3%

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

Based on the simulation in Python:

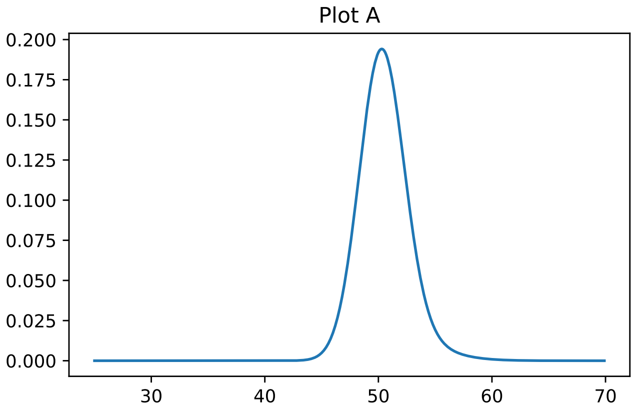
```
# Set seed
np.random.seed(123)

# Number of simulations
n = 100000

# Simulate independent realizations of X and Y:
x = stats.norm.rvs(size=n,loc=50,scale=2)
y = stats.norm.rvs(size=n,loc=45,scale=5)

# Calculate the probability
print(np.sum(1*(y > x))/n)

0.17734
```

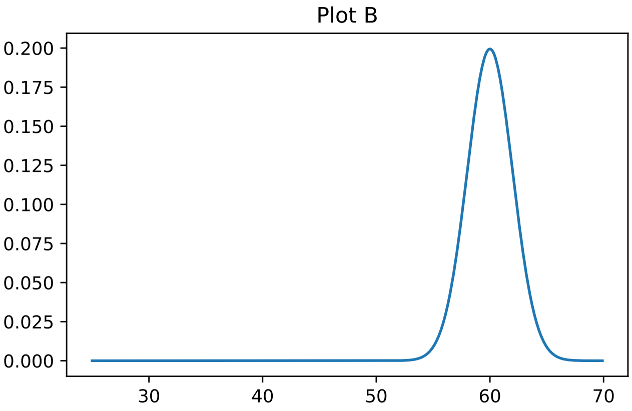
```
------------------------------------ FACIT-END ------------------------------------
```

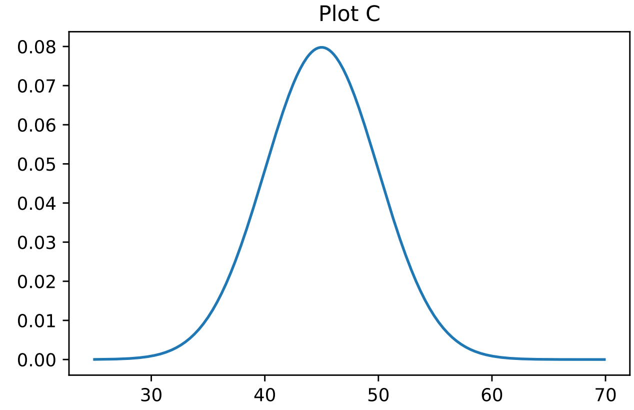
Continue on page 15

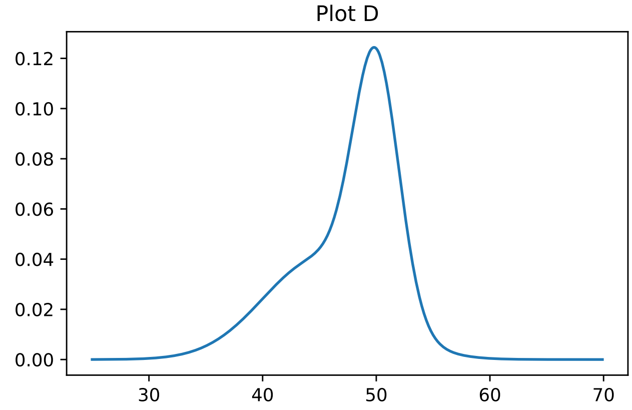
Question VII.2 (13)

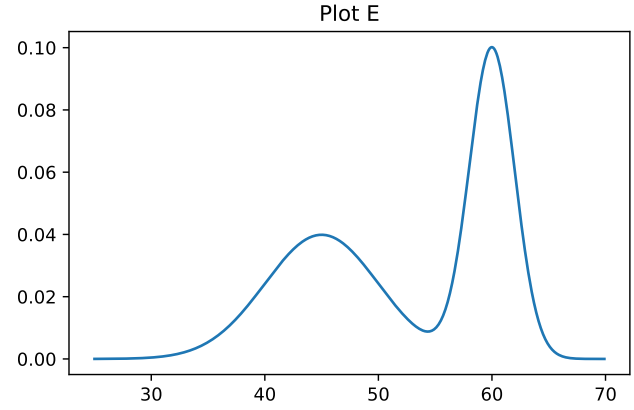
Which of the following plots shows the probability density function (pdf) of U?











1* □ Plot A

2 □ Plot B

3 □ Plot C

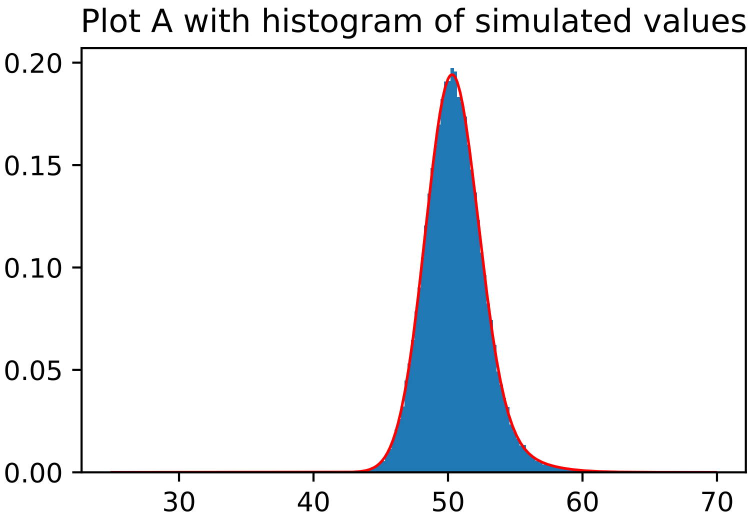
4 □ Plot D

5 □ Plot E

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

Plot A shows the correct density function of U. If we perform a simulation with 100,000 realizations of U and overlay Plot A with a histogram of the simulated values, we see that the empirical distribution matches the density almost perfectly.

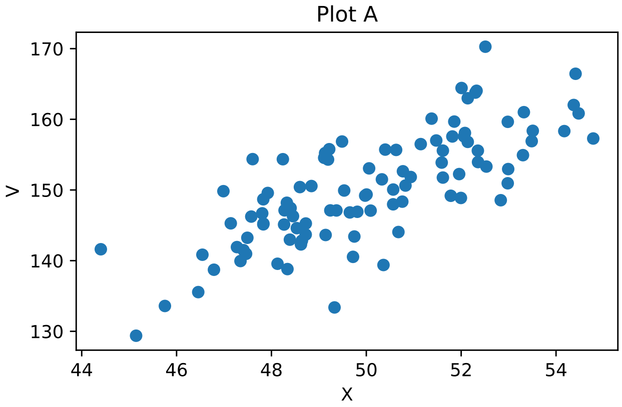
```
# Plot A
plt.plot(xvals,yvalsA,color='red',linewidth=1)

# Find realizations of U and make histogram
I = 1*(y>x)
u = I*y + (1-I)*x
plt.hist(u,density=True,bins=100)
plt.title("Plot A with histogram of simulated values")

plt.show();
```

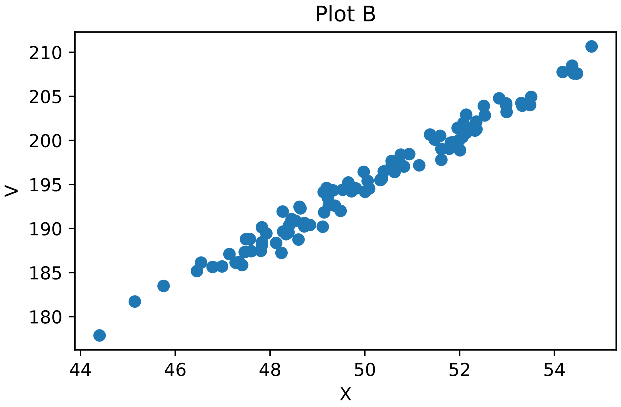


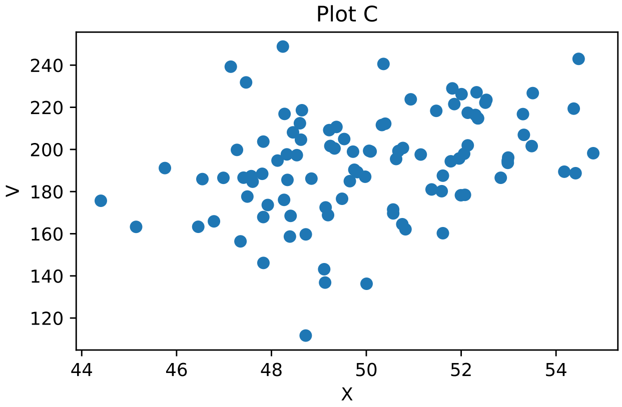
----------------------------------- FACIT-END -----------------------------------

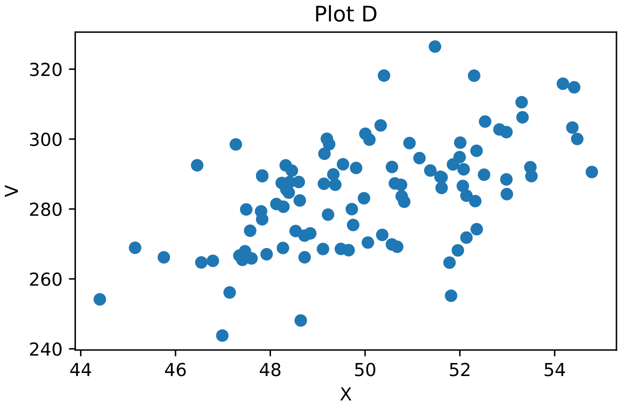
Question VII.3 (14)

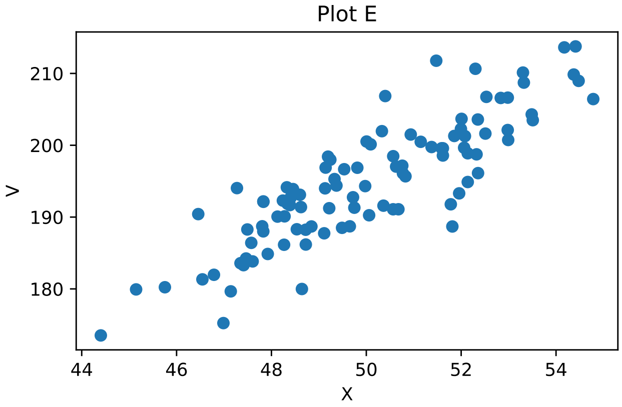
Which of the following plots most likely shows a scatter plot of 100 observations of (X, V)?











1 □ Plot A

2 □ Plot B

3 □ Plot C

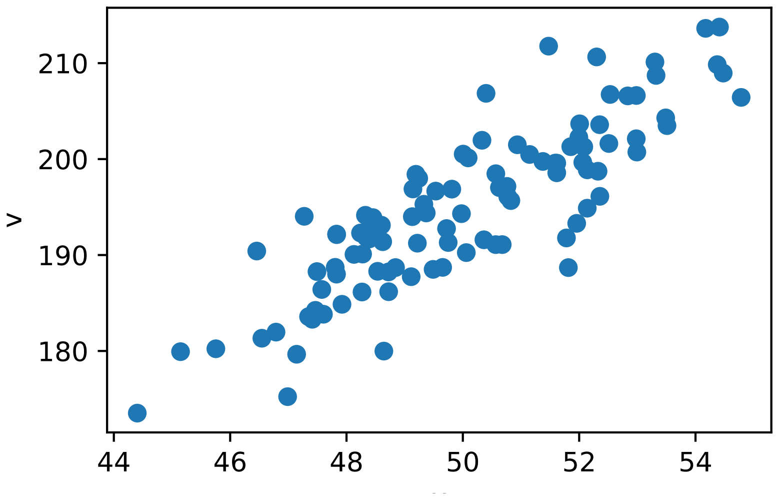
4 □ Plot D

5* □ Plot E

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

Plot E is a scatter plot of (X, V). Notice that $\mathbb{E}[X] = 50$ and $\mathbb{E}[V] = \mathbb{E}[3X + Y] = 3\mathbb{E}[X] + \mathbb{E}[Y] = 3{\cdot}50{+}45 = 195$. Since both X and V are normally distributed, the scatter plot should be approximately centered around the point $(50, 195)$. This leaves options B, C, and E. To identify plot E as the correct option, note the vertical variance around fixed x-values. In Plot B, the spread (standard deviation) is approximately one, while in Plot C, it is approximately 25. As the variance of V for a given value of X, say 50, should be $\mathbb{V}[V|X = 50] = \mathbb{V}[3{\cdot}50{+}Y] = \mathbb{V}[Y] = 5^2$, the vertical spread of V for a fixed value of x should be 5, which matches the spread in Plot E. These properties can also be observed by simulating the values and making a scatter plot as seen below:

```
# Number of observations
n = 100

# Find simulated values of V based on X and Y
x = x[:n]
y = y[:n]
v = 3*x + y

# Scatter plot
plt.scatter(x,v)
plt.xlabel("x")
plt.ylabel("v")
plt.show();
```



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

Exercise VIII

The police set up a traffic checkpoint where they randomly select cars for inspection. On average, they check 10 cars per hour, and the number of cars selected for inspection in an hour follows a Poisson distribution. Let X denote the total number of cars selected during a five-hour period.

Question VIII.1 (15)

What is the most appropriate model for X?

1 □ X follows a binomial distribution with parameters $n = 5$ and $p = 0.1$.

2 □ X follows a binomial distribution with parameters $n = 5$ and $p = 0.5$.

3 □ X follows a Poisson distribution with rate $\lambda = 2$.

4 □ X follows a Poisson distribution with rate $\lambda = 10$.

5* □ X follows a Poisson distribution with rate $\lambda = 50$.

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

Due to the scalling property of the Poisson distribution, see pp. 63-64, X should follow a Poisson distribution with a rate five times the hourly rate i.e. with a rate $\lambda = 50$.

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

Question VIII.2 (16)

What is the probability that no cars are selected for inspection during a given hour?

1* □ Approximately 0%

2 □ Approximately 10%

3 □ Approximately 20%

4 □ Approximately 80%

5 □ Approximately 90%

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

Let Y denote the number of cars selected during an hour. The density function of a Poisson(10) distributed random variable is given by definition 2.27:

$$P(Y = 0) = e^{-10} \approx 0.$$

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

Continue on page 20

Question VIII.3 (17)

Historical data suggests that 40% of inspections result in a citation. If, during a given hour, the police randomly select three cars for inspection, what is the probability that exactly two of the inspections result in a citation?

1 □ 0.096

2 □ 0.144

3* □ 0.288

4 □ 0.432

5 □ 0.720

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

Let Z denote the number of inspections that result in a citation. Then Z follows a binomial distribution with $n = 3$ (here the inspections constitute the trials) and $p = 0.4$ (the probability that each independent inspection results in a citation). Thus,

$$P(Z = 2) = \binom{3}{2} 0.4^2 (1 - 0.4)^{(3-2)} = 3 \cdot 0.16 \cdot 0.6 = 0.288,$$

cf. eq. (2-20).

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

Continue on page 22

Exercise IX

A couple is planning to buy a house and wants to estimate the expected price. They collect data on recent property sales in the area and use Python to fit a multiple linear regression model with property price as the response variable and the sizes of the house and lot (in square meters) as explanatory variables.

They are particularly interested in two very similar properties. Property A has a house of 175 square meters and a lot of 800 square meters, while Property B has the same lot size but a house that is 10 square meters smaller i.e., 165 square meters.

The couple obtains the following output from their Python model, where the price is given in tkr. (tusinde kroner - DKK thousands).

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.892
Model:                            OLS   Adj. R-squared:                  0.887
No. Observations:                  45   F-statistic:                     173.8
Covariance Type:            nonrobust   Prob (F-statistic):           4.84e-21
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    585.8165    613.968      0.954      0.345    -653.220    1824.853
House         43.9742      2.757     15.951      0.000      38.411      49.538
Lot            3.1214      0.275     11.334      0.000       2.566       3.677
==============================================================================
```

Question IX.1 (18)

What is the expected price of Property A according to the model?

1 □ Approximately DKK 10.193 M

2* □ Approximately DKK 10.778 M

3 □ Approximately DKK 20.200 M

4 □ Approximately DKK 35.726 M

5 □ Approximately DKK 36.311 M

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

The parameter estimates are found in the Python output, and the expected (predicted) price of Property A is then found by plugging in the features of Property A into the estimated model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 175 + \hat{\beta}_2 \cdot 800 = 585.8165 + 43.9742 \cdot 175 + 3.1214 \cdot 800 = 10778.4215.$$

The expected price of Property A according to the model is thus approximately DKK 10.778 M.

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

Continue on page 24

A real estate agent claims that Property B should cost 400,000 less than Property A, implying that each square meter of house is worth 40,000. The couple wants to evaluate whether the data support this claim (null hypothesis) based on the fitted model.

Question IX.2 (19)

The usual test of the real estate agent's claim (null hypothesis) results in which p-value?

1 □ $p = 0.922$

2 □ $p = 0.692$

3 □ $p = 0.345$

4* □ $p = 0.157$

5 □ $p < 0.001$

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

The null hypothesis of the real estate agent can be formulated as $\mathcal{H}_0 : \beta_1 = 40$. Method 6.4 shows that the relevant test statistic becomes

$$t_{\text{obs},\beta_1} = \frac{43.9742 - 40}{2.757} = 1.4415.$$

The p-value thus becomes

$$p = 2P(T > |t_{\text{obs}}|) = 2P(T > 1.4415),$$

where T follows a t-distribution with $n - (p + 1) = 45 - (2 + 1) = 42$ degrees of freedom.

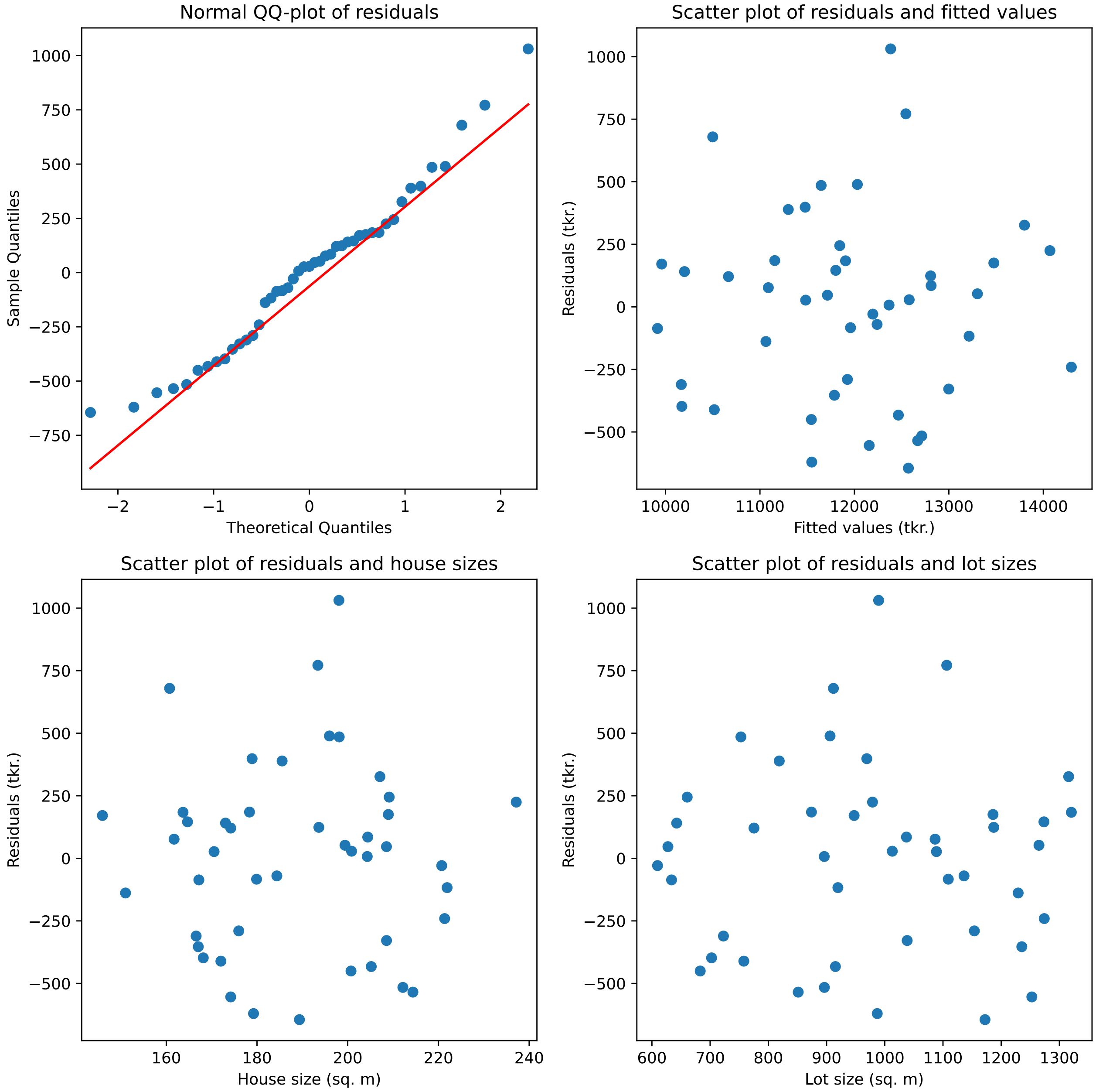
```python
# Observed test statistic
tobs = (43.9742-40)/2.757
print(tobs)

1.441494377947045

# p-value
p = 2*(1-stats.t.cdf(tobs,df=42))
print(p)

0.15686012308774466
```

Thus, the p-value is 0.157.

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

Continue on page 25

The couple validates the model and generates the following diagnostic plots:



Continue on page 26

Question IX.3 (20)

Which of the following statements is false?

1 □ The normal QQ-plot of the residuals suggests that the residuals may be slightly right-skewed but shows no obvious violation of the normality assumption.

2* □ The normal QQ-plot of the residuals suggests that the residuals may be inter-dependent but shows no obvious violation of the independence assumption.

3 □ The scatter plot of residuals versus fitted values suggests that the variance of the residuals may increase with the fitted values but shows no obvious violation of the homoscedasticity (variance homogeneity) assumption.

4 □ The scatter plot of residuals versus house size suggests that the variance of the residuals may increase with house size but shows no obvious violation of the homoscedasticity (variance homogeneity) assumption.

5 □ The scatter plot of residuals versus lot size suggests that the variance of the residuals may increase with lot size but shows no obvious violation of the homoscedasticity (variance homogeneity) assumption.

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

You cannot determine whether the residuals are inter-dependent (i.e., dependent on each other) based on a QQ-plot.

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

Continue on page 27

Exercise X

One wants to compare the means in two samples, A and B. Both samples contain 50 independent measurements, which are assumed to be normally distributed. It is stated that the 95% confidence interval for the mean in each group is:

95% CI for $\hat{\mu}_A = [15.2, 17.8]$

95% CI for $\hat{\mu}_B = [13.0, 15.5]$.

Question X.1 (21)

Which of the following statements is correct?

1 □ Since the confidence intervals overlap, the two underlying populations could have the same mean. Therefore, we can easily see that the difference between the two sample means is not significantly different from zero (at a 5% significance level).

2 □ Since the confidence intervals overlap, the difference between the two sample means is not statistically significantly different from zero (at a 5% significance level). Thus, the two underlying populations have the same distribution.

3* □ The sample means for sample A and B are 16.50 and 14.25, respectively, and there is a significant difference between these means at a 5% significance level (but not at a 1% significance level).

4 □ The sample means for sample A and B are 16.50 and 14.25, respectively, and there is a significant difference between these means (at a 1% significance level).

5 □ The sample means for sample A and B are 16.50 and 14.25, respectively, but there is no significant difference between these means (at a 5% significance level).

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

Den korrekte konklusion er, at der er en signifikant forskel mellem gennemsnittene af stikprøve A og stikprøve B, ved et signifikansniveau på 5% (men ikke på 1%, da den relevante p-værdi bliver 0.013).

```python
# Gennemsnit de to materialer
mean_A, mean_B = (15.2 + 17.8)/2, (13.0 + 15.5)/2

# Beregner margin of error i de to prøver
ME_A = (17.8 - 15.2) / 2
ME_B = (15.5 - 13.0) / 2

print(mean_A, mean_B)
```

```
16.5 14.25

# Standard error of the mean beregnes ud fra margin of error:
sem_A = ME_A / stats.t.ppf(0.975, df = 50-1)
sem_B = ME_B / stats.t.ppf(0.975, df = 50-1)
print(sem_A, sem_B)

0.6469028757823986 0.6220219959446137

# Estimere forskellen mellem gennemsnittene og konfidensinterval for forskellen
diff = mean_A - mean_B
se_diff = np.sqrt(sem_A**2 + sem_B**2)
print(diff, se_diff)

2.25 0.8974378497335949

teststatistic = diff / se_diff
nu = (sem_A**2 + sem_B**2)**2 / (sem_A**4/49 + sem_B**4/49)

print(2*stats.t.cdf(-teststatistic, df = nu))

0.013818332076165151
```

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

Continue on page 29

Exercise XI

Capture-recapture is a method in which a number of individuals (animals) are captured, tagged, and released. After a period of time, a number of individuals are captured and it is examined how many individuals are tagged. The method can be used to estimate population sizes.

A biologist has captured $n_1 = 150$ fish in a lake, tagged them, and released them again. The biologist now plans to return and capture $n_2 = 200$ fish from the same lake.

Question XI.1 (22)

If we denote the total number of fish in the lake by N (and assume that N is the same when released and recaptured), what distribution will the number of tagged fish (Y) then follow at recapture (it is assumed that all tagged fish survive and that it is completely random which of the N fish are captured)?

1 □ A binomial distribution with $p = \frac{150}{N}$, and $n = 200$, i.e. $Y \sim B\left(200, \frac{150}{N}\right)$.

2 □ A normal distribution with paramters $\mu = \frac{200 \cdot 150}{N}$, and $\sigma^2 = \frac{200 \cdot 150}{N}\left(1 - \frac{150}{N}\right)$.

3* □ A hypergeometric distribution with parameters $n = 200$, $a = 150$, and N, i.e. $Y \sim H(200, 150, N)$.

4 □ A Poisson distribution with parameter $\lambda = \frac{150 \cdot 200}{N}$, i.e. $Y \sim Pois\left(\frac{150 \cdot 200}{N}\right)$.

5 □ An exponential distribution with parameter $\lambda = \frac{N}{150 \cdot 200}$, i.e. $Y \sim Exp\left(\frac{N}{150 \cdot 200}\right)$.

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

At recapture the number of marked fish is $a = 150$ (out of a total of N fish) and the number of fish captured is $n = 200$, this is a hypergeometric distribution with the mentioned parameters. You may think of this as $N - a$ white balls (unmarked fish) and a black balls (number of marked fish), and $n = 200$ balls (fish) are chosen at random. Now count the number of black balls Y (marked fish).

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

The length of the caught fish is measured in order to provide an estimate of their age. Thus, fish between 6 and 10 cm. are classified as 1-year-old, while fish of more than 10 cm. are classified as older. It is assumed that the length of a one-year-old fish follows a normal distribution with mean $\mu = 8$ cm. and standard deviation $\sigma = 1$ cm.

Question XI.2 (23)

What is the probability that a one-year-old fish is classified as older than one year?

1 □ 0.159

2 □ 0.5

3 □ 0.841

4* □ 0.0228

5 □ 0.977

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

The length, Y of a 1-year old fish is assumed to follow a $N(8,1)$-distribution and we are looking for the probability

$$P(Y > 10) = 1 - P(Y < 10) = 1 - F(10) \tag{2}$$

which can be calculated in Python by

```
1-stats.norm.cdf(10,8,1)

0.022750131948179209
```

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

Continue on page 31

Exercise XII

A consumer organization wants to investigate how often a parcel delivery company delivers packages to the nearest parcel shop. They collected data from 750 parcel deliveries, distributed across 5 regions. The results are summarized in the following table:

Region	**Delivered to the nearest parcel shop**	**Delivered elsewhere**	**Total**
Capital region	40	110	150
Central Jutland	95	55	150
Southern Denmark	80	70	150
North Jutland	85	65	150
Zealand	70	80	150
Total	370	380	750

A χ^2-test is now performed to investigate whether the proportion of packages delivered to the nearest parcel shop is the same across all 5 regions.

Question XII.1 (24)

What are the expected values in each cell of the table under the null hypothesis?

1 □

Region	**Delivered to the nearest parcel shop**	**Delivered elsewhere**
Capital region	75	75
Central Jutland	75	75
Southern Denmark	75	75
North Jutland	75	75
Zealand	75	75

2 □

Region	**Delivered to the nearest parcel shop**	**Delivered elsewhere**
Capital region	80	70
Central Jutland	70	80
Southern Denmark	60	90
North Jutland	50	100
Zealand	40	110

3 □

Region	**Delivered to the nearest parcel shop**	**Delivered elsewhere**
Capital region	100	0
Central Jutland	100	0
Southern Denmark	100	0
North Jutland	100	0
Zealand	100	0

4*□

Region	Delivered to the nearest parcel shop	Delivered elsewhere
Capital region	74	76
Central Jutland	74	76
Southern Denmark	74	76
North Jutland	74	76
Zealand	74	76

5□

Region	Delivered to the nearest parcel shop	Delivered elsewhere
Capital region	40	110
Central Jutland	95	55
Southern Denmark	80	70
North Jutland	85	65
Zealand	70	80

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

```
370 * 150 / 750

74.0

380 * 150 / 750

76.0
```

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

Question XII.2 (25)

A χ^2-test is performed to investigate whether the proportion of packages delivered to the nearest parcel shop is the same across all 5 regions. The relevant test statistic has been calculated as 47.21. What is the p-value for the relevant test, and what is the corresponding conclusion (use a significance level of $\alpha = 0.05$)?

1 □ The p-value is 0.10, and the conclusion is that there is a difference in the proportion of packages delivered to the nearest parcel shop in the different regions.

2 □ The p-value is 0.10, and the conclusion is that there is no difference in the proportion of packages delivered to the nearest parcel shop in the different regions.

3 □ The p-value is 0.05, and the conclusion is that there is no difference in the proportion of packages delivered to the nearest parcel shop in the different regions.

4 □ The p-value is $1.4 \cdot 10^{-9}$, and the conclusion is that there is no difference in the proportion of packages delivered to the nearest parcel shop in the different regions.

5*□ The p-value is $1.4 \cdot 10^{-9}$, and the conclusion is that there is a difference in the proportion of packages delivered to the nearest parcel shop in the different regions.

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

```
1-stats.chi2.cdf(47.21,df=4)

1.3788027386496537e-09
```

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

Continue on page 34

Exercise XIII

A sensor measures the temperature of a machine that should not exceed 80°C. A sample of 40 temperature measurements gives a sample mean of 75.2°C and a sample standard deviation of 2.5°C. Assume that the temperature measurements are independent of each other and follow a normal distribution.

Question XIII.1 (26)

What is a 95% confidence interval for the true mean temperature?

1 □ $[72.7, 77.7]$ °C

2 □ $[70.1, 80.3]$ °C

3*□ $[74.4, 76.0]$ °C

4 □ $[75.1, 75.3]$ °C

5 □ $[71.0, 79.4]$ °C

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

Konfidensintervallet beregnes som:

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

Hvor $t_{\alpha/2,39}$ aflæses fra t-fordelingen med 39 frihedsgrader.

Beregning i Python:

```python
xbar, s, n = 75.2, 2.5, 40
t_value = stats.t.ppf(0.975, df=n-1)
margin_of_error = t_value * (s / (n ** 0.5))
conf_interval = (xbar - margin_of_error, xbar + margin_of_error)
print(conf_interval)
```

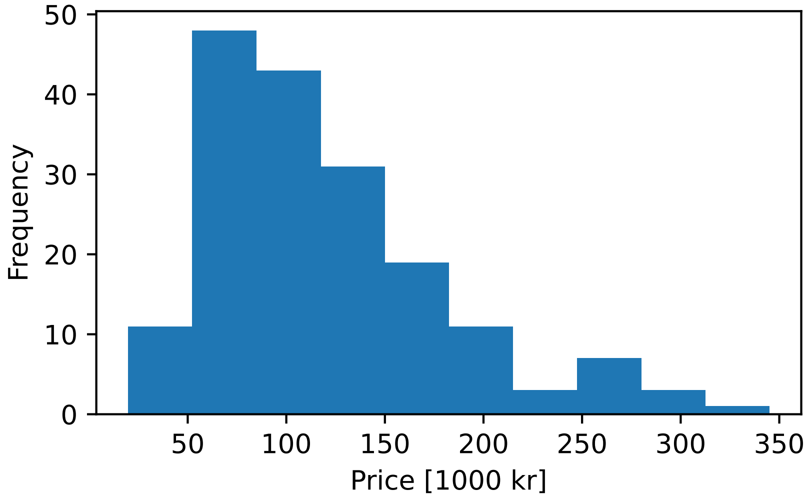
```
(74.4004612112678, 75.99953878873221)
```

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

Continue on page 35

Exercise XIV

A car owner wants to buy a new (used) car, to investigate what the price should be. She has collected prices for the car (make and model) she wants. The histogram below shows the distribution of prices for the car she wants.



In addition to the histogram, she has calculated the average and empirical variance for the observed prices (`price`) [1000 kr.].

```
np.mean(price)

np.float64(118.94622598870056)

np.var(price,ddof=1)

np.float64(3633.1131006077294)
```

Question XIV.1 (27)

Based on the above, which of the following assumptions about the distribution of the price (Y) is the most reasonable?

1* □ A log-normal distribution with parameters $\alpha = 4.66$ and $\beta^2 = 0.478$, i.e. $Y \sim LN(4.66, 0.478^2)$.

2 □ A normal distribution with parameters $\mu = 118.94$ and $\sigma^2 = 3633.1^2$, i.e. $Y \sim N(118.9, 3633.1^2)$.

3 □ An exponential distribution with parameter $\lambda = 118.9$, i.e. $Y \sim Exp(118.9)$.

4 □ A log-normal distribution with parameters $\alpha = 118.9$ and $\beta^2 = 60.28^2$, i.e. $Y \sim LN(118.9, 60.28^2)$.

5 □ A normal distribution with parameters $\mu = 118.94$ and $\sigma^2 = 60.28^2$, i.e. $Y \sim N(118.9, 60.28^2)$.

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

The distribution is right skrewed (excluding the normal distribution), the mode (max of the frequency) is not at zero (excluding the exponential distribuution). The log-normal distribution can take the approximate form from the histogram (leaving us with answer or 1 and 4 as possible correct) and further if $Y \sim LN(\alpha, \beta^2)$, then

$$E[Y] = e^{\alpha+\beta^2/2} \tag{3}$$

$$V[Y] = e^{2\alpha+\beta^2}\left(e^{\beta^2} - 1\right). \tag{4}$$

inserting the values for answer 1 gives

```
np.exp(4.65+0.488**2/2)

117.80986366351067

np.exp(2*4.65+0.488**2)*(np.exp(0.488**2)-1)

3731.9947734632988
```

which is very close to the observed mean and variance, using the other option (answer 4) will result in values that are completely off, and hence trhe right answer i no. 1.

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

The car owner has also collected data on the age and mileage of the cars. To investigate the relationship between price, age and mileage of the car, the car owner has run the following code (`price` [1000 kr.] is the price, `age` [years] is the age of the car, `dist` [1000 km.] is the mileage of the car, and `cars` is the dataset with the collected numbers),

```
fit = smf.ols("price ~ age + dist", data = cars).fit()
```

corresponding to the model

$$Y_i = \mu_i + \epsilon_i,$$

where the formula for μ_i is determined from the input to `smf.ols` and $E(\epsilon_i) = 0$.

Question XIV.2 (28)

Which of the following statements about the model or model assumptions is correct?

1* □ The ϵ_i's are normally distributed and iid. (independent and identically distributed).

2 □ It is assumed that the observed correlation between `age` and `dist` is equal zero.

3 □ The Y_i's are normally distributed and iid. (independent and identically distributed).

4 □ The μ_i's follows a normal distribution and are iid.

5 □ $\mu_i = \beta_1 x_{1i} + \beta_2 x_{2i}$, where x_{1i} and x_{2i} are the age and mileage of car i, respectively.

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

We will simply go through each of the options. 1) this is the assumption in the general linear model (hence 1 is correct). 2) There is no assumptions on the correlations (strong correlation are called multicolinarity), hence the satement is wrong. 3) the Y_i's are normally distribted but with different mean values (hence the satement is wrong). 4) μ_i is not stochastic (hence the satement is wrong), 5) $\mu_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$ hence the intercept is missing and hence the satement is wrong.

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

The results of the estimation above are given below (some numbers have been replaced with symbols)

```
fit.summary(slim = True)

                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.793
Model:                            OLS   Adj. R-squared:                  0.791
No. Observations:                 177   F-statistic:                     334.0
Covariance Type:            nonrobust   Prob (F-statistic):           2.65e-60
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    255.8098      5.707         T1      P1          Ql_1        Qu_1
age           -9.6077      0.963         T2      P2          Ql_2        Qu_2
dist          -0.4756      0.055         T3      P3          Ql_3        Qu_3
==============================================================================
```

Question XIV.3 (29)

Which of the following statements is correct when using a significance level of $\alpha = 0.05$?

1* □ Both effects (age and mileage) are significantly different from zero and the expected price decreases with age and mileage.

2 □ None of the effects (age and mileage) are significantly different from zero.

3 □ The age of the car has a significant effect on the price, while an effect of the mileage cannot be demonstrated. The price increases as the age increases.

4 □ Mileage has a significant effect on price, while and effect of age cannot be demonstrated. The price decreases as mileage increases.

5 □ The age of the car has a significant effect on the price, while an effect of the mileage cannot be demonstrated. The price decreases as the age increases.

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

Both test statistics are about -10, and hence significant on any relavant (e.g. 0.05) level, further both parameter estimates are negative (as expected), hence answer 1 is correct.

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

Continue on page 39

Exercise XV

A consultant has received data on arrival and departure times for 35 employees at a given workplace. Arrival and departure times are recorded for the same 35 employees on two different days - one day in the summer and one day in the winter. The consultant now wishes to assess whether the average working hours are the same on both days.

Question XV.1 (30)

Which analysis is relevant to perform?

1 □ For each arrival and departure time, the working hours are calculated. Now, you have two independent samples (one for the summer day and one for the winter day) with 35 measurements in each. The means of these samples are compared using a t-test with the null hypothesis H_0: $\mu_1 = \mu_2$.

2 □ For each of the two days, the average arrival time and the average departure time are calculated. Then, two t-tests are performed: one t-test tests for a significant difference in arrival times, and the other tests for a significant difference in departure times.

3*□ For each arrival and departure time, the working hours are calculated. Now, you have two paired samples (one for the summer day and one for the winter day) with 35 measurements in each. A paired t-test is used to examine whether the average difference in working hours is significantly different from zero.

4 □ For each of the two days, a 95% confidence interval is calculated for the average arrival time $CI_{\bar{x}_{arrive}}$ and for the average departure time $CI_{\bar{x}_{leave}}$. If the two confidence intervals do not overlap, there is a significant difference in the average working hours between the two days.

5 □ For each arrival and departure time, a total working time is calculated. Now, you have two samples with 35 measurements. A one-way ANOVA model is used to test whether there is a difference in the average working hours between the two days.

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

Der skal laves en parret test.

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

The exam is finished. Enjoy the vacation!