Technical University of Denmark *Page 1 of 41 pages.*

Written examination: 26.06.2025

Course name and number: **Statistics (02403)**

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

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

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

The exam paper contains 41 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

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 I.1 (1)

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 ----------

We use the two sided t-test for the mean value with the test statistic

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

The p-value is calculated in Python with

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

2.7386127875258306

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

0.01043738949886733
```

---------- 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 I.2 (2)

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 ----------

The sample size for a wanted margin or error is

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

in this case we have $ME = 2$ and we use $\sigma = 2$, and it can be calculated in Python by

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

np.float64(384.14588206941244)
```

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

Continue on page 4

Exercise II

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 II.1 (3)

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 ----------

This is a 1-way anova set up, and the model can be formulated as given in answer 1), 3) and 5). In answer 1) and 3) we are not testing the hypothesis that the group means are equal. In answer 1) the model is over parametrized and setting $\mu = 0$ just give $\hat{\alpha}_i = \bar{y}_i$ (i.e. the group averages). In answer 3 we test if all group means are equal zero (which is not the null-hypothesis). In answer 5 setting $\alpha_i = 0$ correspond to having the same mean value in all groups, hence this is the correct answer.

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

Exercise III

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

Question III.1 (4)

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

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

where $F(y)$ is the cdf of a exponential distribution with parmater 1/3, and in Python it is calculated by

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

np.float64(0.2498199809168653)
```

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

Continue on page 6

Exercise IV

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 IV.1 (5)

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 ----------

The estimated probability is obviously

$$\hat{p} = \frac{320}{1000} = 0.32$$

the standard erro of that estimate is given by

$$se_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{1000}}$$

which is calculated in Python by

```
ph = 320/1000
ph

0.32

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

np.float64(0.014751271131668619)
```

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

Question IV.2 (6)

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 ----------

The test statistics for the null-hypothesis $H_0 : p = p_0$, is given by

$$z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

and the p-value is calculated from the standard normal, the result is

```
ph = 320/1000
p0 = 0.2
n = 1000
se0 = np.sqrt(p0*(1-p0)/n)
```

```
z_obs = (ph-p0)/se0
print(z_obs)

9.486832980505136

2*(1 - stats.norm.cdf(z_obs))

np.float64(0.0)
```

The p-value is rounded to zero, but with $z_{obs} = 9.5$ is is extremely small. Hence the p-values given in answer 2 and 3 are wrong. The precise p-value can be calculated by

```
z_obs,p_value = smprop.proportions_ztest(320, 1000, value=0.2,
                                         prop_var=0.2)
p_value

np.float64(2.3816001643963165e-21)

z_obs

np.float64(9.486832980505136)
```

which are the p-values given in answer 4 and 5. The very low p-value also exclude answer 1.

Answer 4 state that the sample is clearly representative, this is a very doubtfull statement as this is a sample from customers in a pet-store (presumably they all have pets), and therfore the statement in answer 5 is the correct one.

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

Continue on page 9

Exercise V

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 V.1 (7)

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 ----------

With the given data we can directly calculate the ANOVA table as

```
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
```

from which the answer is given directly as no. 2.

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

Continue on page 11

Question V.2 (8)

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 ----------

We will simply go through each answer one by one

1. This is part of the defenition of the ANOVA model hence correct.
2. We know from the separation of variation that $SST > SSE$ and therefore $\frac{SST}{N-1}$ will also usually be grater than $\hat{\sigma}^2 = \frac{SSE}{N-k}$, so the statement is not correct.
3. This is exactly the vaariance estimate.
4. The is in essence how the F-test statistics is calculated.
5. Here we need a concrete calculation, see below

for answer 5 we calculate

```
np.mean(np.array([89, 102, 94, 90, 100]))-np.mean(y)

np.float64(9.75)
```

hence answer 5 is also correct. And so the correct answer is answer 2.

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

If we denote the observations from each group $\boldsymbol{y}_i$ (e.g. $\boldsymbol{y}_1 = [89 \quad 102 \quad 94 \quad 90 \quad 100]^T$), and the collection of all observations by $\boldsymbol{Y}$, the model can be written as

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \boldsymbol{y}_3 \\ \boldsymbol{y}_4 \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

where it is assumed that $\boldsymbol{\beta}$ is identifiable.

The corresponding fitted values can be written as

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y},$$

where $\boldsymbol{H}$ is calculated from $\boldsymbol{X}$.

Question V.3 (9)

Which of the following statements about $\boldsymbol{X}$ and $\boldsymbol{H}$ is correct?

1 ☐ $\boldsymbol{X} \in \mathbb{R}^{20\times5}$, and $Trace(\boldsymbol{H}) = 5$.

2* ☐ $\boldsymbol{X} \in \mathbb{R}^{20\times4}$, and $Trace(\boldsymbol{H}) = 4$.

3 ☐ $\boldsymbol{X} \in \mathbb{R}^{20\times4}$, and $Trace(\boldsymbol{H}) = 20$.

4 ☐ $\boldsymbol{X} \in \mathbb{R}^{20\times20}$, and $Trace(\boldsymbol{H}) = 20$.

5 ☐ $\boldsymbol{X} \in \mathbb{R}^{20\times5}$, and $Trace(\boldsymbol{H}) = 20$.

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

$\boldsymbol{X}$ can be written in several different ways, but the number of columns will be equal to 4 (the number of groups), when $\boldsymbol{\beta}$ is identifiable, the number of rows is equal the number of observations (i.e. 20). Also $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, which is unique and have rank (i.e. trace) equal the number of columns in $\boldsymbol{X}$ (i.e. 4).

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

Continue on page 13

Exercise VI

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 VI.1 (10)

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 ----------

The correct conclusion is that there is a significant difference between the means of sample A and sample B, at a significance level of 5% (but not at 1%, since the relevant p-value will be 0.013).

```
# Average of the two meterials
mean_A, mean_B = (15.2 + 17.8)/2, (13.0 + 15.5)/2

# Margin of erro for the two samples
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 is calculated from the 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

# Confidence interval for the difference
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 15

Exercise VII

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 VII.1 (11)

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 VII.2 (12)

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)

np.float64(0.02275013194817921)
```

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

Continue on page 17

Exercise VIII

A company wants to investigate the effect of lighting and music conditions on employee productivity (measured in the number of units produced per hour).

Two factors are tested:

1) Lighting (Factor A) with two levels: Low and High

2) Music (Factor B) with three levels: None, Calm, and Energetic

All combinations are tested, and the employees' average productivity is measured:

	No Music	Calm Music	Energetic Music
Low Lighting	20	23	19
High Lighting	25	27	22

It is now assumed that the model is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}; \quad \epsilon_{ij} \sim N(0, \sigma^2),$$

where ϵ_{ij} are iid., Y_{ij} (`y`) is productivity, and α_i and β_j represent the effects of Music (`Music`) and Lighting (`Light`). To investigate the effects, the following result has been obtained from Python (data from the table above is stored in `D`).

```
fit = smf.ols('y ~ Light + Music', data = D).fit()
sm.stats.anova_lm(fit)

            df     sum_sq    mean_sq          F    PR(>F)
Light      1.0  24.000000  24.000000  48.000000  0.020204
Music      2.0  20.333333  10.166667  20.333333  0.046875
Residual   2.0   1.000000   0.500000        NaN       NaN
```

Continue on page 18

Question VIII.1 (13)

What is the conclusion from the relevant statistical tests, using a significance level of $\alpha = 0.05$?

1*☐ There is a significant difference in productivity both with respect to different lighting and different music.

2 ☐ There is a significant difference in productivity with respect to different lighting, but not with respect to different music.

3 ☐ There is a significant difference in productivity with respect to different music, but not with respect to different lighting.

4 ☐ There is no significant difference in productivity, either with respect to different music or different lighting.

5 ☐ There is a significant difference in productivity with respect to different music, but one cannot conclude whether there is an effect of different lighting, since there are only two levels.

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

The p-value show a significant effect of both `Light` and `Music`, this is answer no. 1.

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

Question VIII.2 (14)

To perform further investigations on the effect of music, pairwise comparisons of its effects need to be made. While correcting for multiple tests, what is the Least Significant Distance (LSD) for the pairwise comparisons of the effect of music (using a significance level of $\alpha = 0.05$)?

1☐ 2.5

2 ☐ 2.0

3 ☐ 1.5

4* ☐ 5.4

5 ☐ 6.2

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

The confinece interval for the difference between two groups is given by

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2}(df)\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

and hence

$$LSD = t_{1-\alpha/2}(df)\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

In our case we need the Bonferoni corrected confidence level, and with 3 levels of music we will have 3 pairwise comparisons, and 2 observations in each group, and finalle we have $MSE = 0$. so

$$\begin{aligned} LSD =& t_{1-0.05/2\cdot 3}(2)\sqrt{0.5\left(\frac{1}{2}+\frac{1}{2}\right)} \\ =& t_{1-0.05/2\cdot 3}(2)\sqrt{0.5} \end{aligned}$$

which can be calculated in Python by

```
np.sqrt(0.5) * stats.t.ppf(1-0.05/(2*3), 2)

np.float64(5.408521132466451)
```

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

Continue on page 20

Exercise IX

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 IX.1 (15)

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 ----------

Under the null hypethesis the probability two probabilities (nearest and other) are given by

$$\hat{p}_{near} = \frac{370}{750}$$
$$\hat{p}_{othe} = \frac{380}{750}$$

and the expected values are given by (same number of total in all regions)

$$\hat{e}_{near} = 150\frac{370}{750}$$
$$\hat{e}_{othe} = 150\frac{380}{750}$$

which is calculated by

```
370 * 150 / 750
```

```
74.0
```

```
380 * 150 / 750
```

```
76.0
```

this is answer 4.

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

Question IX.2 (16)

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 ----------

since we have 2 columns (c) and five rows, the teststatistic should be compared to a χ^2-distributionwith $(c-1)(r-1) = 4$ degrees of freedom, and the p-value can be calculated by

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

np.float64(1.3788027386496537e-09)
```

which is the p-value from answer 4 and 5, and since the p-value is low the conclusion is that there is a difference (answer 5).

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

Continue on page 23

Exercise X

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 X.1 (17)

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 ----------

The confidence interval is calculated by:

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

Where $t_{1-\alpha/2}$ is the $1 - \alpha/2$ quantiule in a t-distribution with $n - 1 = 39$ degrees of freedom.

In Python this is calculated by:

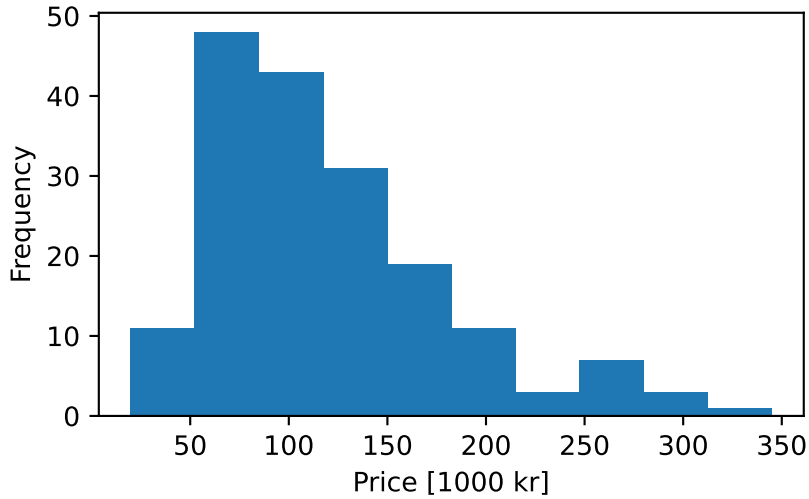
```
xbar, s, n = 75.2, 2.5, 40
t_value = stats.t.ppf(0.975, df=n-1)
margin_of_error = t_value * (s / (n ** 0.5))
conf_interval = (xbar - margin_of_error, xbar + margin_of_error)
print(conf_interval)

(np.float64(74.4004612112678), np.float64(75.99953878873221))
```

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

Continue on page 24

Exercise XI

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 XI.1 (18)

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)

np.float64(117.80986366351067)

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

np.float64(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 XI.2 (19)

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 ☐ $\mu_i = \beta_1 \texttt{age}_i + \beta_2 \texttt{dist}_i$.

---------- 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 XI.3 (20)

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.

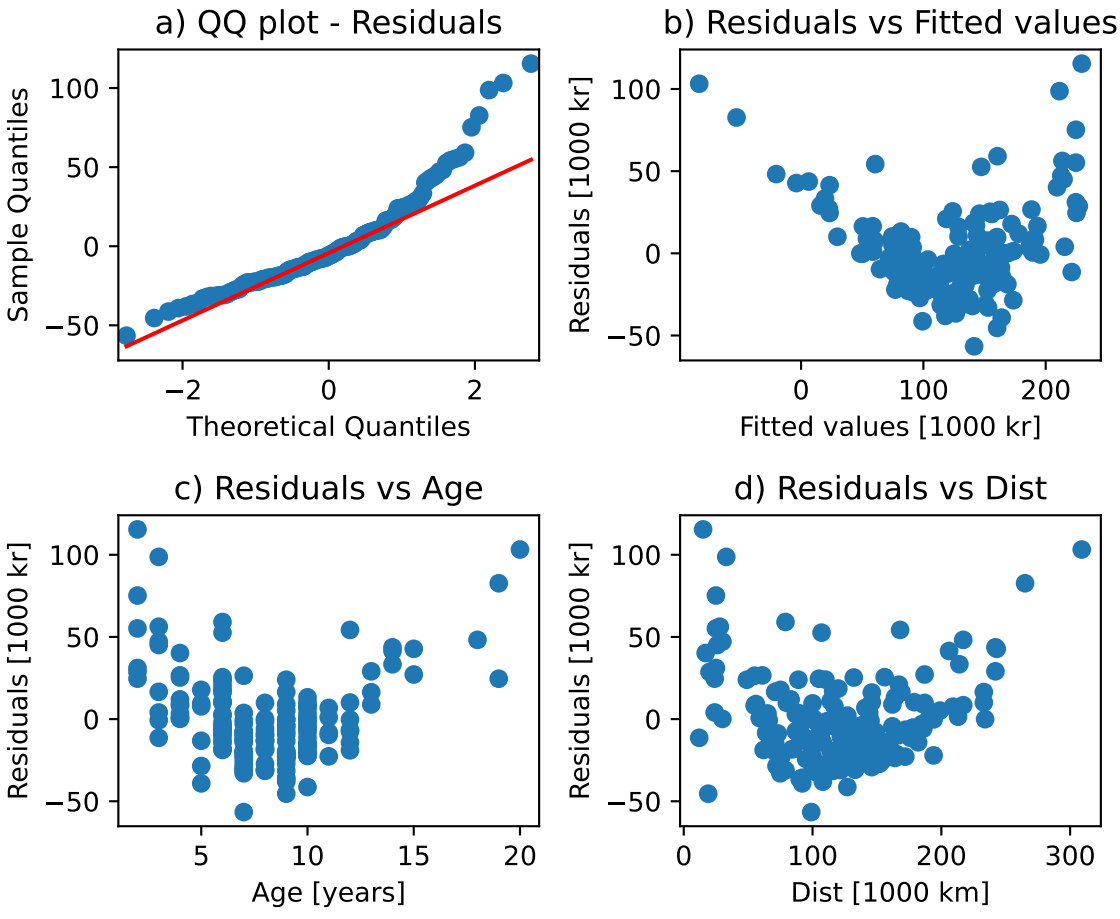
5 ☐ The age of the car has a significant effect on the price, while an effect of the mileage cannot be demonstrated. The price decreases as the age increases.

---------- 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 ----------

To assess the validity of the model, the car owner has created a series of residual plots, as seen below.



Continue on page 28

Question XI.4 (21)

Which of the following statements is NOT correct (both statement and figure reference must be correct)?

1* ☐ The independence assumption is clearly not met (plot a).

2 ☐ The variance homogeneity assumption does not appear to be met (plot b).

3 ☐ A term of the form $\texttt{age}^2$ will probably improve the model (plot c).

4 ☐ There are clear systematic effects in residuals vs. fitted (plot b).

5 ☐ The normal distribution assumption is clearly not met (plot a).

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

We go through the claims. 1) the independence assumption cannot be assessed by the qq-plot, hence the claim is incorrect. 2) The variance seems to be increasing as a function of the fitted values hence the calim is correct. 3) There does indeed seem to a quadratic relation between the age and the the residuals, and hence the model would probabliy also benefit from that, hence the claim is correct. 4) There is indeed a systematic relation ship between the fitted values and the residuals (implying that the model should include more terms). 5) The qq-plot clearly indicate that the normal assumption is not fulfilled, hence the claim is correct.

In conclusion the correct answer is 1).

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

Regardless of the conclusion above, the car owner decides to continue analyzing the model. The model corresponding to the Python summary above (just below question 19) can be written in matrix-vector notation as

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

where $\boldsymbol{X}$ is the design matrix corresponding to the Python summary. To help with the further calculations, the Python code below is provided

```
C = np.array([[177,1413,22393], [1413,13099,202256],
              [22393, 202256, 3384553]])
```

where $\texttt{C}=\boldsymbol{X}^T\boldsymbol{X}$.

Question XI.5 (22)

If $\hat{\sigma}$ denotes the usual standard deviation, what is then the 95% confidence interval for the price [1000 kr.] of a 10-year-old car that has been driven 100,000 km?

1* ☐ $112.2 \pm 0.272 \cdot \hat{\sigma}$

2 ☐ $112.2 \pm 0.0112 \cdot \hat{\sigma}$

3 ☐ $112.2 \pm 0.148 \cdot \hat{\sigma}$

4 ☐ $112.2 \pm 0.0375 \cdot \hat{\sigma}$

5 ☐ $112.2 \pm 1.97 \cdot \hat{\sigma}$

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

The confidence interval at a new point $\boldsymbol{x}$ is given by

$$\boldsymbol{x}^T\hat{\boldsymbol{\beta}} \pm \hat{\sigma} t_{0.975}(n-3)\sqrt{\boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}} \tag{5}$$

which can be calculated in Python by.

```
CI = np.linalg.inv(C)
x = np.array([1,10,100])
np.sqrt(x @ CI @ x) * stats.t.ppf(0.975,df=174)

np.float64(0.27220378916149096)

255.8098 -9.6077 *10 -0.4756 * 100

112.1728
```

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

The car owner now wants to test whether there is an effect of including `age` and `dist` in the model corresponding to the null hypothesis

$$\boldsymbol{Y} = \mathbf{1}\mu + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}),$$

against the alternative hypothesis (as stated above)

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

The projection matrices corresponding to the null hypothesis and the alternative hypothesis are denoted as $\boldsymbol{H}_0$, and $\boldsymbol{H}$ and she calculates the usual test statistic

$$Q = \frac{\frac{1}{df_0}SS_0}{\frac{1}{df_{SSE}}SSE},$$

where

$$SS_0 = \boldsymbol{y}^T(\boldsymbol{H}_0 - \boldsymbol{H})\boldsymbol{y}$$
$$SSE = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}.$$

Question XI.6 (23)

What is Q in the example above and which distribution should Q be compared to?

1 ☐ $Q = 0.793$ which should be compared to an F distribution with 1 and 3 degrees of freedom.

2 ☐ $Q = 0.793$ which should be compared to an F distribution with 1 and 2 degrees of freedom.

3 ☐ $Q = 2.65 \cdot 10^{-60}$ which should be compared to an F-distribution with 3 and 177 degrees of freedom.

4* ☐ $Q = 334.0$ which should be compared to an F distribution with 2 and 174 degrees of freedom.

5 ☐ $Q =$`T1`^{2}`+T2`^{2}`+T3`^{2} which is to be compared to an F-distribution with 3 and 177 degrees of freedom.

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

This is the test for total homogenity and Q is given directly in the model summary (`F-statistics`) as 334, which should be compared to an F-distribution with $df_1 - df_0 = 3 - 1 = 2$ and $df_{SSE} - df_1 = 177 - 3 = 174$ degrees of freedom.

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

Continue on page 32

Exercise XII

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 XII.1 (24)

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 ----------

This is clearly a paired situation, where we compare the difference in working time for each individual at the two days, this is answer no. 4.

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

Continue on page 33

Exercise XIII

In an experiment, it has been assumed that the relationship between an input (x) and an output (y) is given by

$$Y_i = f(x_i) + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2),$$

and that ϵ_i is iid. Since the function f is unknown, it has been decided to fit the model

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2).$$

The parameters of the model are estimated, i.e. we have parameter estimates ($\hat{\boldsymbol{\beta}} = [\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3]^T$) and a corresponding covariance matrix $\boldsymbol{\Sigma}_\beta$ (with $(\boldsymbol{\Sigma}_\beta)_{kl} = \sigma_{kl}$ equal to the covariance between $\hat{\beta}_k$ and $\hat{\beta}_l$). We now want to investigate where the extremum of the second-degree polynomial lies, corresponding to solving

$$f'(x) = \beta_2 + 2\beta_3 x = 0,$$

for x. The solution is called $x^*(\boldsymbol{\beta})$, and it is assumed that $\boldsymbol{\beta} \sim N(\hat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_\beta)$.

Question XIII.1 (25)

Using error propagation, what will the approximation of the variance of x^* be?

1 ☐ $V[x^*] \approx \frac{\hat{\beta}_1^2}{\hat{\beta}_3^2}\sigma_{11} + \frac{1}{4\hat{\beta}_3^2}\sigma_{22} + \frac{\hat{\beta}_2^2}{4\hat{\beta}_3^4}\sigma_{33}$

2 ☐ $V[x^*] \approx \frac{1}{2\hat{\beta}_3^2}\left(\frac{1}{2}\sigma_{22} + \frac{\hat{\beta}_2^2}{2\hat{\beta}_3^2}\sigma_{33}\right)$

3 ☐ $V[x^*] \approx \frac{\hat{\beta}_1^2}{\hat{\beta}_3^2}\sigma_{11} + \frac{1}{4\hat{\beta}_3^2}\sigma_{22} + \frac{\hat{\beta}_2^2}{4\hat{\beta}_3^4}\sigma_{33} - \frac{\hat{\beta}_1\hat{\beta}_2}{\hat{\beta}_3^3}\sigma_{12} + \frac{\hat{\beta}_1}{\hat{\beta}_3^2}\sigma_{13} - \frac{\hat{\beta}_2}{\hat{\beta}_3^3}\sigma_{23}$

4* ☐ $V[x^*] \approx \frac{1}{2\hat{\beta}_3^2}\left(\frac{1}{2}\sigma_{22} + \frac{\hat{\beta}_2^2}{2\hat{\beta}_3^2}\sigma_{33} - \frac{\hat{\beta}_2}{\hat{\beta}_3}\sigma_{23}\right)$

5 ☐ $V[x^*] \approx \frac{\hat{\beta}_1^2}{\hat{\beta}_3^2}\sigma_{11} + \frac{1}{4\hat{\beta}_3^2}\sigma_{22} + \frac{\hat{\beta}_2^2}{4\hat{\beta}_3^4}\sigma_{33} + \frac{\hat{\beta}_1\hat{\beta}_2}{\hat{\beta}_3^3}\sigma_{12} + \frac{\hat{\beta}_2}{\hat{\beta}_3^3}\sigma_{23}$

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

With

$$x^*(\boldsymbol{\beta}) = -\frac{\beta_2}{2\beta_3} \tag{6}$$

we have

$$\frac{\partial x^*(\boldsymbol{\beta})}{\partial \beta_1} = 0 \tag{7}$$

$$\frac{\partial x^*(\boldsymbol{\beta})}{\partial \beta_2} = -\frac{1}{2\beta_3} \tag{8}$$

$$\frac{\partial x^*(\boldsymbol{\beta})}{\partial \beta_3} = \frac{\beta_2}{2\beta_3^2} \tag{9}$$

and the variance of $x^*(\boldsymbol{\beta})$ is given by

$$V[x^*(\boldsymbol{\beta})] = \sum_{i=1}^{3} \left(\frac{\partial x^*(\boldsymbol{\beta})}{\partial \beta_i} \right)^2 \sigma_{ii} + 2 \sum_{i=1}^{3} \sum_{j=i+1}^{3} \frac{\partial x^*(\boldsymbol{\beta})}{\partial \beta_i} \frac{\partial x^*(\boldsymbol{\beta})}{\partial \beta_j} \sigma_{ij} \tag{10}$$

$$= \sum_{i=2}^{3} \left(\frac{\partial x^*(\boldsymbol{\beta})}{\partial \beta_i} \right)^2 \sigma_{ii} + 2 \frac{\partial x^*(\boldsymbol{\beta})}{\partial \beta_2} \frac{\partial x^*(\boldsymbol{\beta})}{\partial \beta_3} \sigma_{23} \tag{11}$$

$$= \left(\frac{1}{2\beta_3} \right)^2 \sigma_{22} + \left(\frac{\beta_2}{2\beta_3^2} \right)^2 \sigma_{33} - 2 \frac{1}{2\beta_3} \frac{\beta_2}{2\beta_3^2} \sigma_{23} \tag{12}$$

$$= \frac{1}{2\beta_3^2} \left(\frac{1}{2} \sigma_{22} + \frac{\beta_2^2}{2\beta_3^2} \sigma_{33} - \frac{\beta_2}{\beta_3} \sigma_{23} \right) \tag{13}$$

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

Continue on page 35

Now assume that in a specific study it is observed that

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}; \quad \boldsymbol{\Sigma}_{\beta} = \begin{bmatrix} 0.2 & 0 & -0.01 \\ 0 & 0.01 & 0 \\ -0.01 & 0 & 0.001 \end{bmatrix}$$

Question XIII.2 (26)

Using simulation from the assumed distribution, in what interval does the standard deviation of x^* lie (use at least 1000 repetitions)?

1 ☐ [0.09;0.14]

2* ☐ [0.04;0.08]

3 ☐ [0.01;0.03]

4 ☐ [0.001;0.005]

5 ☐ [0.25;1.25]

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

Since $\sigma_{23} = 0$ the simulation can be based in independent Gaussian's, and hence the answer is calculated by

```
k = 1000
b2 = stats.norm.rvs(loc = 2, scale = np.sqrt(0.01), size = k)
b3 = stats.norm.rvs(loc = -1, scale = np.sqrt(0.001), size = k)
xs = - b2 / (2 * b3)
np.std(xs,ddof=1)

np.float64(0.05832389205476338)
```

which is answer no. 2.

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

The parameterization of the second degree polynomial above is not unique and can be formulated as

$$Y_i = \beta_1 p_0(x_i) + \beta_2 p_1(x_i) + \beta_3 p_2(x_i) + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2),$$

where $p_j(x)$ is a polynomial of degree j.

Question XIII.3 (27)

If the number of observations is $n = 11$ and $[x_1, x_2, ..., x_{11}] = [-5, -4, ..., 4, 5]$, what set of polynomials then gives an orthogonal design (i.e. $(\boldsymbol{X}^T\boldsymbol{X})_{ji} = 0$ for $i \neq j$)?

1 ☐ $p_0(x_i) = \frac{1}{11}$, $p_1(x_i) = x_i - 5$, $p_2(x_i) = x_i^2 - x_i - 25$.

2 ☐ $p_0(x_i) = 1$, $p_1(x_i) = x_i - 2.5$, $p_2(x_i) = x_i^2 - 10$.

3* ☐ $p_0(x_i) = 1$, $p_1(x_i) = x_i$, $p_2(x_i) = x_i^2 - 10$.

4 ☐ $p_0(x_i) = 1$, $p_1(x_i) = x_i$, $p_2(x_i) = x_i^2$.

5 ☐ $p_0(x_i) = \frac{1}{11}$, $p_2(x_i) = x_i + 5$, $p_1(x_i) = x_i^2 - 25$.

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

Orthogonality imply that

$$\sum_{i=1}^{11} p_0(x_i)p_1(x_i) = 0 \tag{14}$$

$$\sum_{i=1}^{11} p_0(x_i)p_2(x_i) = 0 \tag{15}$$

$$\sum_{i=1}^{11} p_1(x_i)p_2(x_i) = 0 \tag{16}$$

with $p_0(x_i) = a_0$, $p_1(x_i) = a_1 + b_1x_i$, and $p_2(x_i) = a_2 + b_2x_i + c_2x_i^2$. Implying that

$$\sum_{i=1}^{11} p_0(x_i)p_1(x_i) = a_0\sum_{i=1}^{11}(a_1 + b_1x_i) = 11a_0a_1 + a_0b_1\sum_{i=1}^{11} x_i \tag{17}$$

Now $\sum_{i=1}^{11} x_i = 0$ and hence if $a_0 \neq 0$, the we must have $a_1 = 0$, this is answer 3 and 4 (wich mean that we can set $a_0 = 1$). Now consider

$$\sum_{i=1}^{11} p_0(x_i)p_2(x_i) = \sum_{i=1}^{11}(a_2 + b_2x_i + c_2x_i^2) = (11a_2 + c_2\sum_{i=1}^{11} x_i^2) \tag{18}$$

now $\sum_{i=1}^{11} x_i^2 = 2(25+16+9+4+1) = 110$, hence with $c_2 = 1$ and $a_2 = -10$ we get $\sum p_0p_2 = 0$, and this is answer 3 (with $b_2 = 0$). Now we just need to check that $\sum p_1p_2 = 0$

$$\sum_{i=1}^{11} p_1(x_i)p_2(x_i) = \sum_{i=1}^{11} x_i(x_i^2 - 10) = \sum_{i=1}^{11} x_i^3 - 10x_i = \sum_{i=1}^{11} x_i^3 \tag{19}$$

since $\sum_{i=1}^{11} x_i^3 = 0$ (e.g. $(-5)^3 + 5^3 = 0$), the polynomials in answer 3 are orthogonal (for the given $\boldsymbol{x}$).

Another solution is simply to check each of the 5 cases

```
x = np.array([-5,-4,-3,-2,-1,0,1,2,3,4,5])
## Ans 1
np.sum(1 / 11 * (x-5)) ## sum p_0p_1

np.float64(-5.0)

## hence not orthogonal

## Ans 2
np.sum( (x-2.5))

np.float64(-27.5)

## hence not orthogonal

## Ans 3
np.sum(x)  ## sum p_0p_1

np.int64(0)

np.sum(x**2-10)  ## sum p_0p_2

np.int64(0)

np.sum(x*(x**2-10))  ## sum p_0p_1

np.int64(0)

## hence orthogonal

## Ans 4
np.sum(x)  ## sum p_0p_1

np.int64(0)

np.sum(x**2)  ## sum p_0p_2

np.int64(110)

## hence not orthogonal

## Ans 5
np.sum(1/11*(x+5))  ## sum p_0p_1

np.float64(5.0)

## hence not orthogonal
```

Again the correct answer is 3.

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

Continue on page 39

Exercise XIV

In order to determine the weight of a particular object, two people weigh the object a number of times on a scale. It is assumed that the measurement error on the scale is normally distributed with mean 0 and variance σ^2, and that all weighings are independent. The two people weigh the object 2 and 3 times, respectively, and report the average ($\overline{Y}_1$ and $\overline{Y}_2$).

Question XIV.1 (28)

If the weight of the object is equal to μ, what are the mean and variance of $\overline{\overline{Y}} = \frac{1}{2}\overline{Y}_1 + \frac{1}{2}\overline{Y}_2$ then?

1 ☐ $E(\overline{\overline{Y}}) = \mu$, $V(\overline{\overline{Y}}) = \sigma^2$

2 ☐ $E(\overline{\overline{Y}}) = \mu$, $V(\overline{\overline{Y}}) = \frac{\sigma^2}{2}$

3 ☐ $E(\overline{\overline{Y}}) = \mu$, $V(\overline{\overline{Y}}) = \frac{5\sigma^2}{6}$

4* ☐ $E(\overline{\overline{Y}}) = \mu$, $V(\overline{\overline{Y}}) = \frac{5\sigma^2}{24}$

5 ☐ $E(\overline{\overline{Y}}) = \mu$, $V(\overline{\overline{Y}}) = \frac{\sigma^2}{5}$

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

We have $\overline{Y}_1 \sim N(\mu, \frac{\sigma^2}{2})$, $\overline{Y}_1 \sim N(\mu, \frac{\sigma^2}{3})$, and hence

$$V[\overline{\overline{Y}}] = \frac{1}{4}\left(\frac{\sigma^2}{2} + \frac{\sigma^2}{3}\right) \tag{20}$$

$$= \frac{\sigma^2}{4}\left(\frac{3}{6} + \frac{2}{6}\right) \tag{21}$$

$$= \frac{5\sigma^2}{24} \tag{22}$$

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

In order to construct a test statistic, a weighted average ($w \in (0, 1)$) is calculated

$$\widetilde{\overline{Y}} = w\overline{Y}_1 + (1 - w)\overline{Y}_2,$$

such that $\widetilde{\overline{Y}}$ and $\overline{Y}_i - \widetilde{\overline{Y}}$ are independent.

Question XIV.2 (29)

For which w is it true that $Cov(\overline{Y}_1 - \widetilde{\overline{Y}}, \widetilde{\overline{Y}}) = 0$?

1 ☐ $w = \frac{1}{2}$

2 ☐ $w = \frac{1}{3}$

3* ☐ $w = \frac{2}{5}$

4 ☐ $w = \frac{5}{12}$

5 ☐ $w = \frac{5}{24}$

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

$$
\begin{aligned}
Cov(\overline{Y}_1 - \widetilde{\overline{Y}}, \widetilde{\overline{Y}}) =& Cov\left[\overline{Y}_1 - (w\overline{Y}_1 + (1-w)\overline{Y}_2), w\overline{Y}_1 + (1-w)\overline{Y}_2\right] && (23)\\
=& Cov\left[(1-w)\overline{Y}_1 - (1-w)\overline{Y}_2, w\overline{Y}_1 + (1-w)\overline{Y}_2\right] && (24)\\
=& (1-w)Cov\left[\overline{Y}_1 - \overline{Y}_2, w\overline{Y}_1 + (1-w)\overline{Y}_2\right] && (25)\\
=& (1-w)\left[wV(\overline{Y}_1) - (1-w)V(\overline{Y}_2)\right] && (26)\\
=& (1-w)\sigma^2\left[\frac{w}{2} - (1-w)\frac{1}{3}\right] && (27)\\
& && (28)
\end{aligned}
$$

and hence we need to solve

$$\frac{1}{3} = \frac{w}{2} + w\frac{1}{3} = w\frac{5}{6} \tag{29}$$

and hence $w = \frac{2}{5}$.

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

Using the results above, two independent computational quantities, Q_1 and Q_2, have been constructed, such that $\frac{k_1}{\sigma^2}Q_1 \sim \chi^2(1)$ and $\frac{k_2}{\sigma^2}Q_2 \sim \chi^2(1)$, where k_1 and k_2 are constants.

Question XIV.3 (30)

Which of the following statements is correct?

1 ☐ $\frac{1}{2\sigma^2}\frac{k_1Q_1}{k_2Q_2} \sim F(2,1)$.

2* ☐ $\frac{k_1Q_1}{k_2Q_2} \sim F(1,1)$.

3 ☐ $\frac{1}{2}\frac{Q_1/k_1}{Q_2/k_2} \sim F(1,2)$.

4 ☐ $\frac{Q_1}{Q_2} \sim F(1,1)$.

5 ☐ $\frac{Q_1/k_1}{Q_2/k_2} \sim F(1,1)$.

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

Since that ratio between two χ^2-distributions, each divided by their respective degrees of freedom we have

$$\frac{\frac{k_1}{\sigma^2}Q_1}{\frac{k_2}{\sigma^2}Q_2} = \frac{k_1 Q_1}{k_2 Q_2} \sim F(1,1).$$

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

The exam is finished. Enjoy the vacation!