Technical University of Denmark *Page 1 of 37 pages.*

OBS: THIS FRONT PAGE WILL LOOK DIFFERENT AT THE REAL EXAM!

Written examination: TEST SET - a remake of the December 2024 set

Course name and number: **02402 Statistics (Polytechnical Foundation)**

Duration: 4 hours

Aids and facilities allowed: All written aids

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.

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 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 FORM WILL LOOK DIFFERENT AT THE REAL EXAM!

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

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

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

The exam paper contains 37 pages.

Continue on page 2

The use of Python code in this exam: *This exam set includes Python code. Note that we use the following libraries and abbreviations:*

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import scipy.stats as stats
import statsmodels.api as sm
import statsmodels.formula.api as smf
import statsmodels.stats.power as smp
import statsmodels.stats.proportion as smprop
```

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.*

Exercise I

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

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

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

	5 knots	10 knots	25 knots	50 knots	Average
100 kg	-33.72	-26.95	29.11	-38.87	-17.6075
200 kg	-5.75	-3.00	-15.41	20.56	-0.9000
300 kg	29.96	-24.77	-12.05	1.52	-1.3350
400 kg	-4.72	5.72	24.39	43.16	17.1375
500 kg	-22.36	23.99	-24.17	33.36	2.7050
Average	-7.318	-5.002	0.374	11.946	0.0000

Question I.1 (1)

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

1* ☐ -1.335

2 ☐ -0.900

3 ☐ 0.374

4 ☐ 2.705

5 ☐ 17.138

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

The researchers specify a two-way ANOVA model. The parameter estimates in such a model can be found using equations (8-38) through (8-40):

$$\hat{\alpha}_3 = \bar{y}_{3\cdot} - \bar{y} = -1.335 - 0 = -1.335$$

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

Question I.2 (2)

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

1 ☐ 415.3

2* ☐ 692.2

3 ☐ 2076.5

4 ☐ 2768.7

5 ☐ 8306.1

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

Theorem 8.20 shows that:

$$SS_{Total} = SS_{load} + SS_{knots} + SSE$$

$$SSE = SS_{Total} - SS_{load} - SS_{knots} = 8306.13$$

The formula for MSE comes by dividing SSE by the degrees of freedom in accordance with the ANOVA table on page 374. Hence, $d_f = (k-1)(l-1)$, from the problem $l = 4$ and $k = 5$, as such $d_f = 12$:

$$MSE = \frac{SSE}{d_f} = 692.18$$

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

Question I.3 (3)

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

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

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

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

1 ☐ 13.75

2 ☐ 35.83

3* ☐ 55.00

4 ☐ 220.00

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

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

Equation (8-41) shows how SS(load) can be derived as

$$\text{SS(load)} = l \sum_{i=1}^{k} \hat{\alpha}_i^2 = 4(1^2 + 2^2 + 3^2 + 4^2 + 5^2) = 220.$$

The load mean square is then given as

$$\text{MS(load)} = \frac{\text{SS(load)}}{k-1} = \frac{220}{5-1} = 55,$$

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

Continue on page 5

Exercise II

In a pass/fail course, a class of $n = 60$ students was evaluated, with the results presented:

	Passed	Failed
Results	42	18

Question II.1 (4)

What is the estimated probability of passing the course and its 95% confidence interval, assuming the usual assumptions are satisfied. You will need one of the following values to carry out the calculation, from standard normal distribution:

$P(Z < z)$	0.90	0.925	0.95	0.975	0.99
z	1.282	1.439	1.645	1.96	2.326

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

2* ☐ $\hat{p} = 0.70$ and $[0.58, 0.82]$

3 ☐ $\hat{p} = 0.30$ and $[0.18, 0.42]$

4 ☐ $\hat{p} = 0.70$ and $[0.45, 0.95]$

5 ☐ $\hat{p} = 0.30$ and $[0.16, 0.44]$

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

Since in the sample $x = 42$ of $n = 60$, we use can use the formulas from Method 7.3 in the book:

$$\hat{p} = \frac{Successes}{Total} = \frac{42}{60} = 0.70$$

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

From the table we choose the value of the z-distribution, since we have the 95% confidence interval, $\alpha = 0.05$, and thus we need the $1 - \alpha/2 = 0.975$ quantile:

$$CI = 0.7 \pm 1.96\sqrt{\frac{0.7 \cdot 0.3}{60}}$$

$$CI = [0.584, 0.816]$$

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

Question II.2 (5)

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

1 ☐ $\hat{\sigma}_{\hat{p}} = 0.0598$

2* ☐ $\hat{\sigma}_{\hat{p}} = 0.0579$

3 ☐ $\hat{\sigma}_{\hat{p}} = 0.0845$

4 ☐ $\hat{\sigma}_{\hat{p}} = 0.0573$

5 ☐ $\hat{\sigma}_{\hat{p}} = 0.0592$

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

We use the formulas from remark 7.7, the "Plus 2"-approach:

$$\hat{p_2} = \frac{x+2}{n+4} = \frac{44}{64} = 0.6875$$

The standard error is the part before we multiply by the ppf of the z-distribution we are considering:

$$\hat{\sigma p} = \sqrt{\frac{\hat{p_2}(1-\hat{p_2})}{n+4}} = 0.0579$$

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

Continue on page 7

Exercise III

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

Question III.1 (6)

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

1 ☐ 1/4

2 ☐ 1/3

3 ☐ 1/2

4* ☐ 2/3

5 ☐ 3/4

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

The experiment is a case of drawing without replacement and therefore follows the hypergeometric distribution.
Method 1)
We will find the probability of getting two cups of regular coffee and two cups of decaffinated coffee, and based on that find the probability that we get one of each. Let p_c denote the probability of getting normal coffee and p_d the probability of getting decaffinated coffee. As such:

$$p_c = (\frac{1}{2}) \cdot (\frac{1}{3})$$

In a very similar way we find that:

$$p_d = (\frac{1}{2}) \cdot (\frac{1}{3})$$

The probability of getting one of each is $1 - p_c - p_d$:

$$p = 1 - p_c - p_d = 1 - 2 \cdot \frac{1}{6} = 1 - \frac{1}{3} = \frac{2}{3} = 0.6666667$$

Method 2)
We can also go with the idea that the first draw doesn't matter, but it only affects the probability of drawing the other type of coffee in the second draw, that being if we draw one type, we have a 2 in 3 chance of drawing the other cup. $p = \frac{2}{3} = 0.6666667$

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

Question III.2 (7)

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

1 ☐ $\mathrm{V}(Y) = 5.37$

2* ☐ $\mathrm{V}(Y) = 3.83$

3 ☐ $\mathrm{V}(Y) = 3.11$

4 ☐ $\mathrm{V}(Y) = 2.79$

5 ☐ $\mathrm{V}(Y) = 1.10$

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

In this setup it is "drawing" with replacement, hence X follows a binomial distribution

$$X \sim B(n = 30, p = 0.85)$$

and we can use Theorem 2.20 (Appendix page 440 and 450) to find the variance:

$$V[Y] = n \cdot p \cdot (1 - p) = 3.825$$

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

Continue on page 9

Exercise IV

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

Question IV.1 (8)

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

To answer this question you may need the following integral:

$$\int_0^a \lambda e^{-\lambda x}\,dx = [-\lambda\frac{1}{\lambda}e^{-\lambda x}]_0^a = -\lambda\frac{1}{\lambda}e^{-\lambda a} + \lambda\frac{1}{\lambda}e^0$$

1 ☐ $e^{-\frac{25}{50}}$

2 ☐ $1 - e^{-\frac{50}{25}}$

3 ☐ $e^{-\frac{50}{25}}$

4* ☐ $1 - e^{-\frac{25}{50}}$

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

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

To solve this problem, we'll use the exponential distribution formula $f(x;\lambda) = \lambda e^{-\lambda x}$ (Theorem 2.48, Appendix page 451) with $\lambda = \frac{1}{\mu} = \frac{1}{50}$ and $x = 25$:

We need to calculate $P(X \leq 25) = \int_0^{25} \lambda e^{-\lambda x}\,dx$ (since the exponential function is defined as equal to 0 for $x < 0$):

$$\int_0^{25} f(x;\lambda)\,dx = \lambda\int_0^{25} e^{-\lambda x}\,dx = \lambda[-\frac{1}{\lambda}e^{-\lambda x}]_0^{25} =$$

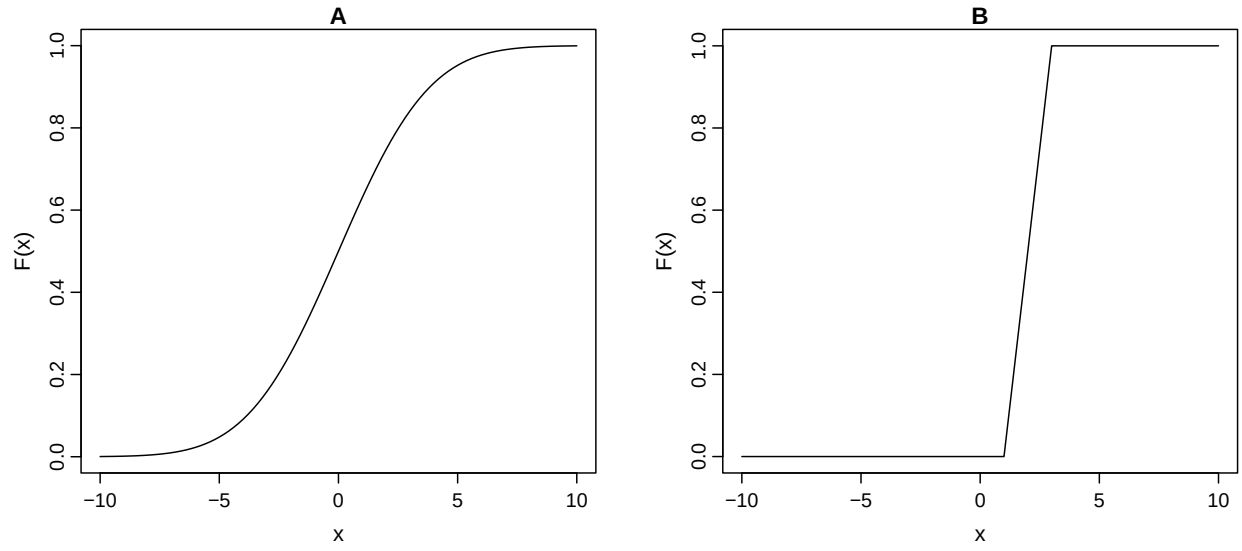
$$= \lambda\cdot(-\frac{1}{\lambda})[e^{-\lambda x}]_0^{25} = -[e^{-25\lambda} - e^0] = 1 - e^{-\lambda x} =$$

$$= 1 - e^{-25\frac{1}{50}} = 1 - e^{-0.5} \approx 0.3935$$

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

Question IV.2 (9)

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



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

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

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

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

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

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

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

Plot A is clearly the normal CDF, since it's smooth. Plot B reveals that $a = 1$ and $b = 3$, since that's the two points of the uniform CDF where a change in the slope occurs.

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

Continue on page 11

Exercise V

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

Yield data is recorded as follows:

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

Summary statistics:

$$\bar{x}_A = 45.5, \quad s_A = 3.03, \qquad \bar{x}_B = 52.0, \quad s_B = 2.58, \qquad n_A = n_B = 10$$

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

A hypothesis test for the mean difference in crop yield ($\mu_A - \mu_B$) was carried out in Python:

```
fertilizerA = [45, 48, 50, 42, 47, 49, 43, 44, 46, 41]
fertilizerB = [51, 53, 52, 50, 55, 48, 54, 49, 56, 52]
result = stats.ttest_ind(fertilizerA, fertilizerB, equal_var=False)
```

And the following results were found:

```
print(result.statistic)
-5.16567619255367
```

```
print(result.pvalue)
7.011228845863585e-05
```

```
print(result.df)
17.562162162162164
```

Question V.1 (10)

Which of the following statements is correct?

1 ☐ A paired two-sample t-test was performed. The conclusion of the test is that there is a significant difference in mean crop yield (at $\alpha = 0.05$), since the p-value is smaller than 0.05.

2 ☐ A pooled two-sample t-test was performed. The conclusion of the test is that there is NO significant difference in mean crop yield (at $\alpha = 0.05$), since the p-value is -5.2.

3 ☐ A pooled two-sample t-test was performed. The conclusion of the test is that there is a significant difference in mean crop yield (at $\alpha = 0.01$), since the p-value is smaller than 0.01.

4 ☐ A Welch two-sample t-test was performed. The conclusion of the test is that there is NO significant difference in mean crop yield (at $\alpha = 0.05$), since the p-value is smaller than 0.01.

5* ☐ A Welch two-sample t-test was performed. The conclusion of the test is that there is a significant difference in mean crop yield (at $\alpha = 0.01$), since the p-value is smaller than 0.01.

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

A Welch two-sample t-test was performed (no assumption of equal variances in the two samples). Since the p-value is smaller than 0.01, the conclusion of the test is that there is a significant difference in mean crop yield (both for $\alpha = 0.01$ and $\alpha = 0.05$).

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

Question V.2 (11)

How should the researchers estimate a 99% confidence interval for the mean difference in crop yields?

1 ☐ $45.5 \pm t_{0.975} \cdot 3.03/\sqrt{10}$, where $t_{0.975}$ is the 97.5% quantile in a t-distribution with 9 degrees of freedom.

2 ☐ $-6.70 \pm t_{0.975} \cdot \sqrt{\frac{3.03^2}{10} - \frac{2.58^2}{10}}$, where $t_{0.975}$ is the 97.5% quantile in a t-distribution with 17.56 degrees of freedom.

3 ☐ $52.0 \pm t_{0.975} \cdot 2.58/\sqrt{10}$, where $t_{0.975}$ is the 97.5% quantile in a t-distribution with 9 degrees of freedom.

4 ☐ $-5.16 \pm t_{0.975} \cdot \sqrt{\frac{3.03^2}{10} + \frac{2.58^2}{10}}$, where $t_{0.975}$ is the 97.5% quantile in a t-distribution with 17.56 degrees of freedom.

5* ☐ $-6.70 \pm t_{0.975} \cdot \sqrt{\frac{3.03^2}{10} + \frac{2.58^2}{10}}$, where $t_{0.975}$ is the 97.5% quantile in a t-distribution with 17.56 degrees of freedom.

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

Answer 5 is correct (see Method 3.47)

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

Continue on page 14

Exercise VI

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

The expression relating weight to diameter is

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

and therefore the expression relating diameter to weight is

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

Here $\rho = 2.6$ g/cm^3 is the density (equal to the density of glass).

The average weight of the marbles is $\mu_W = 1.36$ g.

You can use $\pi = 3.14$.

Question VI.1 (12)

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

1 ☐ $\sigma_D = 0.006$ cm

2 ☐ $\sigma_D = 0.086$ cm

3 ☐ $\sigma_D = 0.015$ cm

4* ☐ $\sigma_D = 0.007$ cm

5 ☐ $\sigma_D = 0.04$ cm

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

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

$$\sigma_D \approx \left|\frac{dD}{dW}\right| \sigma_W, \qquad \frac{dD}{dW} = \frac{2}{3}\left(\frac{3}{4\pi\rho}\right)^{1/3} W^{-2/3}.$$

We evaluate $\frac{dD}{dW}$ at $W = 1.36$ g, thus $\left|\frac{dD}{dW}\right|_{W=1.36} \approx 0.245$.

Inserting this in the expression for σ_D gives:

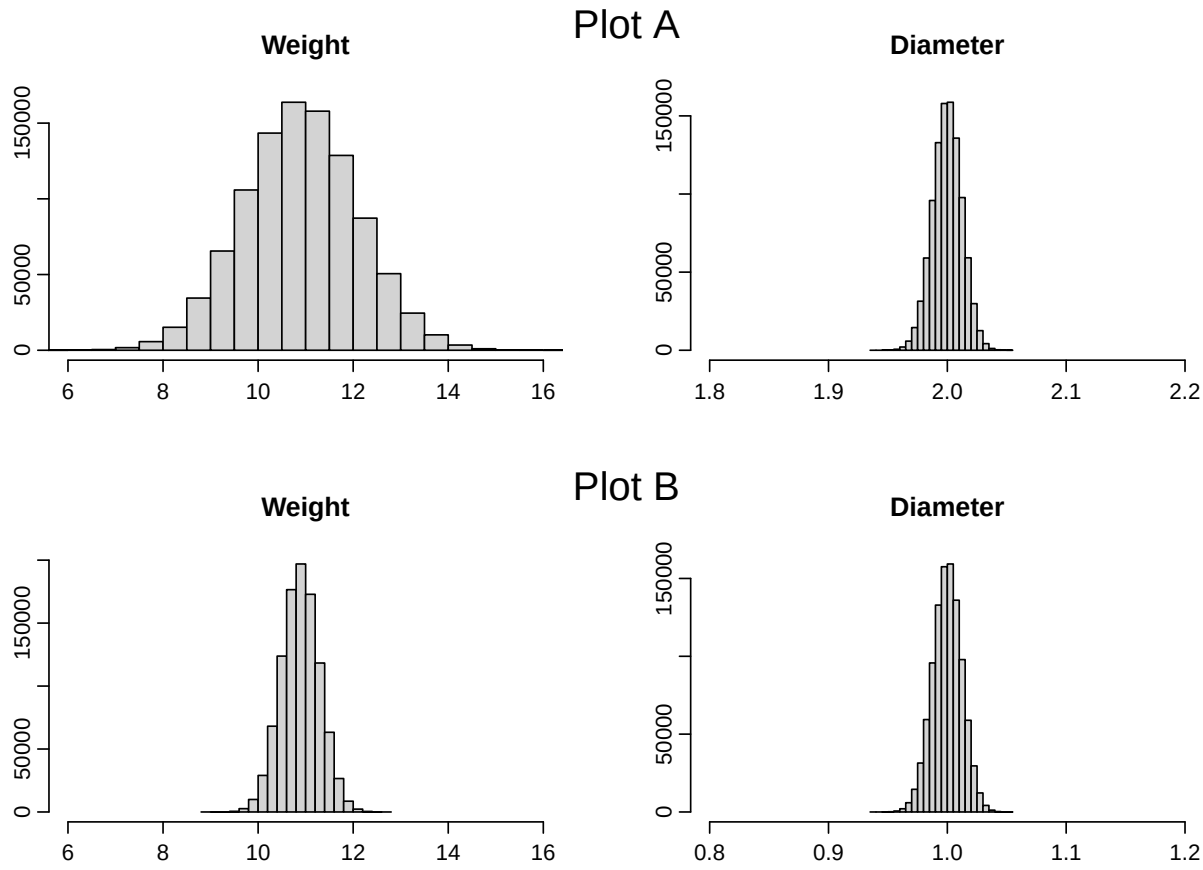
$$\sigma_D \approx 0.245 \cdot 0.03 \approx 0.007 \text{ cm}.$$

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

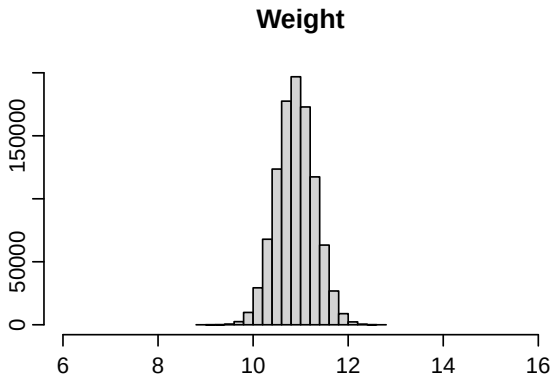
Question VI.2 (13)

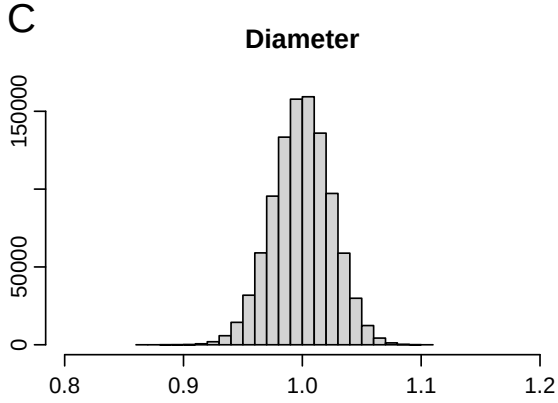
Another brand of marbles has mean diameter of 2 cm, mean weight 10.9 g and $\sigma_W^2 = 0.4^2$ g^2. These marbles also have density $\rho = 2.6$ g/cm^3.

Which set of histograms matches the marbles from this other brand?

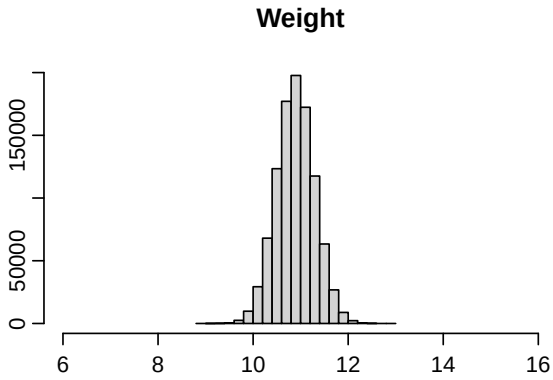


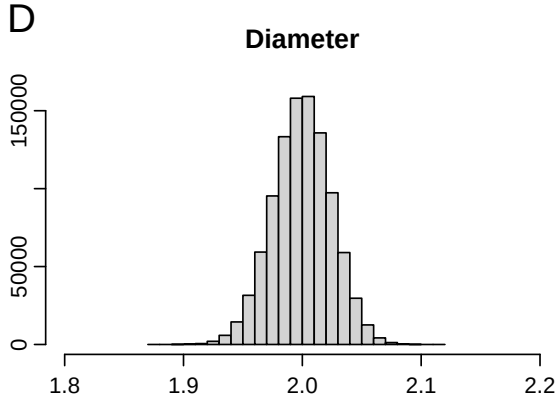
Plot C



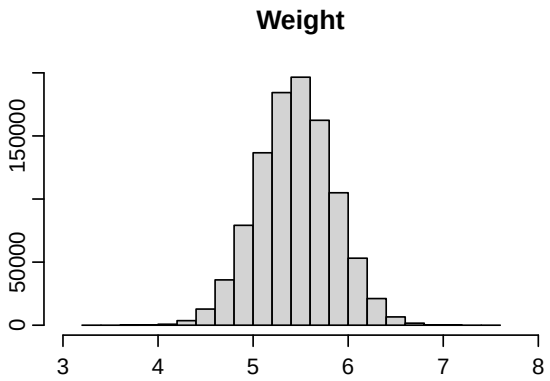


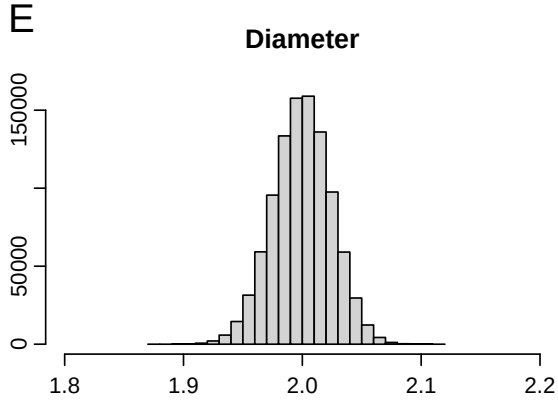
Plot D





Plot E





1 ☐ Plot A

2 ☐ Plot B

3 ☐ Plot C

4*☐ Plot D

5 ☐ Plot E

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

The histogram for D should be centered around 2 and the histogram of W should be centered around 10.9. This leaves Plot A and D.

Since the histograms for weight are approximately normal we can visually deduce that the standard deviation in Plot A is too large to agree with $\sigma_W = 0.4$ g.

This leaves only **Plot D**.

Moreover, we can estimate σ_D to check:

$\frac{dD}{dW}$ at $W = 10.9$ g, gives $\left|\frac{dD}{dW}\right|_{W=10.9} \approx 0.0612$.

Inserting this in the expression for σ_D gives:

$$\sigma_D \approx 0.0612 \cdot 0.04 \approx 0.024 \text{ cm}.$$

This also matches the histogram in Plot D (while the histogram in plot A seems to have even smaller σ_D).

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

Question VI.3 (14)

A third brand of marbles has mean diameter of 2 cm, mean weight 10.9 g and $\sigma_W^2 = 0.6^2$ g^2 (and density $\rho = 2.6$ g/cm^3). The weights follow a normal distribution.

We would like to simulate a sample of 10000 marbles, in order to estimate the standard deviation of their diameters.

Which code simulates the marbles and estimates the standard deviation of marble diameters correctly?

1 ☐
```
weights = stats.norm.rvs(loc=10.9, scale=0.4, size=10000)
diameters = 2*(3*weights/(4*3.14*2.6))**(1/3)
print(diameters.std(ddof=1))
0.024374313384913495
```

2 ☐
```
diameters = stats.norm.rvs(loc=2, scale=0.6, size=10000)
weights = 2.6*4/3*3.14*(diameters/2)**2
print(diameters.std(ddof=1))
0.6045082160581173
```

3* ☐
```
weights = stats.norm.rvs(loc=10.9, scale=0.6, size=10000)
diameters = 2*(3*weights/(4*3.14*2.6))**(1/3)
print(diameters.std(ddof=1))
0.03672201484743135
```

4 ☐
```
diameters = stats.norm.rvs(loc=10.9, scale=0.6, size=10000)
weights = 2.6*4/3*3.14*(diameters/2)**2
print(diameters.std(ddof=1))
0.5963686159324498
```

5 ☐
```
weights = stats.uniform.rvs(loc=10.9, scale=0.4, size=10000)
diameters = 2*(3*weights/(4*3.14*2.6))**(1/3)
print(diameters.std(ddof=1))
0.006946179337405633
```

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

The simulation need to simulate weights from a normal distribution with $\mu = 10.9$ and $\sigma = 0.6$. Hence answer 3 is correct. Only after the weights are simulated, can the diameters be calculated and in the end the standard deviation of diameters are estimated using the sample standard deviation of the simulated diameters.

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

Continue on page 19

Exercise VII

Suppose we have collected exam scores from two groups, the data and summary statistics are as follows:

$$\text{Group 1: } 82, 91, 85, 89, 88 \quad (\bar{x}_1 = 87.0,\ s_1 = 3.54)$$
$$\text{Group 2: } 76, 84, 80, 82, 83 \quad (\bar{x}_2 = 81.0,\ s_2 = 3.16)$$

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

Question VII.1 (15)

What is the estimate of the pooled variance?

1 ☐ 9.00

2 ☐ 27.10

3* ☐ 11.25

4 ☐ 10.00

5 ☐ 8.00

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

By Method 3.52 (pooled two-sample estimate of variance),

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Substituting $n_1 = n_2 = 5,\ s_1 = 3.54,\ s_2 = 3.16$,

$$s_p^2 = \frac{4 \cdot (3.54)^2 + 4 \cdot (3.16)^2}{8} = 11.25.$$

.

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

Question VII.2 (16)

Suppose we want to test whether the mean exam score in Group 1 differs from a reference value of 85. Assume that the population variance remains $\sigma^2 = 3.54$, the significance level is $\alpha = 0.05$, and we want a power of 80% to detect a true mean difference of at least 2 points.

What is the minimum number of students required in such a study?

You may need some of the following quantiles, from the standard normal distribution, to carry out the calculation:

$P(Z < z)$	0.800	0.825	0.850	0.875	0.900	0.925	0.950	0.975	0.990
z	0.842	0.935	1.036	1.150	1.282	1.439	1.645	1.960	2.326

1 ☐ 31

2 ☐ 28

3* ☐ 25

4 ☐ 35

5 ☐ 39

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

For a one-sample z-test with known variance, the required sample size is

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

Here $\sigma = \sqrt{3.54}$, $\delta = 2$, $\alpha = 0.05$ ($z_{1-\alpha/2} = 1.960$), and power 0.80 ($z_{1-\beta} = 0.842$). Hence,

$$n \approx \left(\frac{\sqrt{3.54}}{2}\right)^2 (1.960 + 0.842)^2 \approx 24.6$$

Thus, at least **25 students** are required.

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

Continue on page 21

Exercise VIII

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

Question VIII.1 (17)

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

1* ☐ $1 - \sum_{k=1}^{240} \frac{200^k}{k} e^{-200} = 0.0027$

2 ☐ $\frac{200^{240}}{240} e^{-200} = 0.0006$

3 ☐ $e^{-240/200} = 0.301$

4 ☐ $\frac{240^{200}}{200} e^{-240} = 0.0008$

5 ☐ $1/200 = 0.05$

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

Let X follows a Poisson distribution: $X \sim \text{Poisson}(\lambda = 200)$.

$P(X = x) = f(x) = \frac{\lambda^x}{x} e^{-\lambda}$.

And since $P(X \leq x) = F(x) = 1 - \sum_{k=1}^{x} \frac{\lambda^k}{k} e^{-\lambda}$

We need

$$P(X > 240) = 1 - P(X \leq 240) = 1 - \sum_{k=1}^{240} \frac{\lambda^k}{k} e^{-\lambda} = 0.0027.$$

Hence, the probability that demand exceeds capacity is **0.0027**.

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

Continue on page 22

Exercise IX

In a certain production company, the productivity of its employees follows a normal distribution. Monthly, 50% of the employees can produce 170 units or more (and thus 50% can produce under 170 units).

68.3% of the employees produce within the interval 160-180 units (and thus 84.1% of the employees produce below 180 units).

Question IX.1 (18)

How many units does the top 2.5% (most productive) employees produce monthly?

1* ☐ The top 2.5% most productive employees all produce 189.6 units or more

2 ☐ The top 2.5% most productive employees all produce 190 units or more

3 ☐ The top 2.5% most productive employees all produce 180 units or more

4 ☐ The top 2.5% most productive employees all produce 184.6 units or more

5 ☐ The top 2.5% most productive employees all produce 181.9 units or more

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

First we can see that since the median equals the mean, the mean is 170 units

Then we determine the standard deviation, based on the concept that 68.3% of the data fall withing one standard deviation from the mean, i.e. 160-180, which are 10 units from the mean 170. So, $\sigma = 10$.

The top 2.5% values lie above $\mu + 1.96 \cdot \sigma = 170 + 1.96 \cdot 10 = 189.6$.

(see Example 2.44 in the book)

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

Continue on page 23

Exercise X

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

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

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

Question X.1 (19)

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

1 ☐ 25.50

2 ☐ 162.56

3 ☐ 407.70

4* ☐ 655.48

5 ☐ 1966.43

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

The engineers apply Theorem 8.4 to calculate the residual mean square. In this question, there are $n = 36$ observations (months) equally divided into $k = 3$ groups (years), and Equation (8-14) thus becomes:

$$\text{MSE} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{n - k} = \frac{11 \cdot 22.3^2 + 11 \cdot 27.5^2 + 11 \cdot 26.7^2}{36 - 3} = 655.48.$$

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

Question X.2 (20)

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

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

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

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

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

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

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

The engineers use Method 8.9 to calculate the pairwise confidence intervals. Since two confidence intervals are calculated, the Bonferroni corrected significance level is given as

$$\alpha_{\text{Bonferroni}} = \alpha/M = 0.10/2 = 0.05,$$

where M refers to the number of confidence intervals. Therefore, the engineers must use the $1 - \alpha_{\text{Bonferroni}}/2 = 0.975$ quantile of the t-distribution with $n - k = 36 - 3 = 33$ degrees of freedom.

OBS: For the exam held in December 2024 we decided to accept both answer 2 and 4 ask correct answers for this question. The reason for this being that one would not always do a Bonferoni correction for only two tests and so the decision to do this here is a little specific to this course and the fact that we only include Bonferoni corrections in the chapter about ANOVA.

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

Continue on page 25

Exercise XI

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

Year	Young	Mid-age	Old	Total
2015	2048	1072	821	3941
2016	2208	998	836	4042
2017	2359	1092	853	4304
2018	2138	1093	880	4111
2019	1984	935	799	3718
2020	1777	872	860	3509
2021	1604	818	729	3151
2022	1564	943	753	3259
Total	**15682**	**7823**	**6531**	**30036**

Question XI.1 (21)

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

1 ☐ The p-value is $6.65 \cdot 10^{-6}$ and the conclusion is that there is no significant change in distribution across the years.

```
t_stat, p_val = stats.ttest_ind(tbl['Young'], tbl['Old'], equal_var=False)
print(p_val)
6.654415781263043e-06
```

2 ☐ The p-value is $6.65 \cdot 10^{-6}$ and the conclusion is that there is a significant change in distribution across the years.

```
t_stat, p_val = stats.ttest_ind(tbl['Young'], tbl['Old'], equal_var=False)
print(p_val)
6.654415781263043e-06
```

3 ☐ The p-value is $2.60 \cdot 10^{-8}$ and the conclusion is that there is a significant change in distribution across the years.

```
t_stat, p_val = stats.ttest_ind(tbl['Young'], tbl['Old'], equal_var=True)
print(p_val)
2.6007787343112833e-08
```

4 ☐ The p-value is $4.1 \cdot 10^{-10}$ and the conclusion is that there is no significant change in distribution across the years.

```
chi2, p_val, dof, expected = stats.chi2_contingency(tbl, correction=True)
print(p_val)
4.10013664044193e-10
```

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

```
chi2, p_val, dof, expected = stats.chi2_contingency(tbl, correction=True)
print(p_val)
4.10013664044193e-10
```

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

To test the hypothesis we need to perform a χ^2-test. Answer 5 is correct.

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

Question XI.2 (22)

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

1 ☐ 978

2 ☐ 1364

3* ☐ 1566

4 ☐ 1960

5 ☐ 2048

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

Under the null hypothesis the estimate of the proportion in the Young category is found by summing over all the years, which is then multiplied with the 3000 for that year:

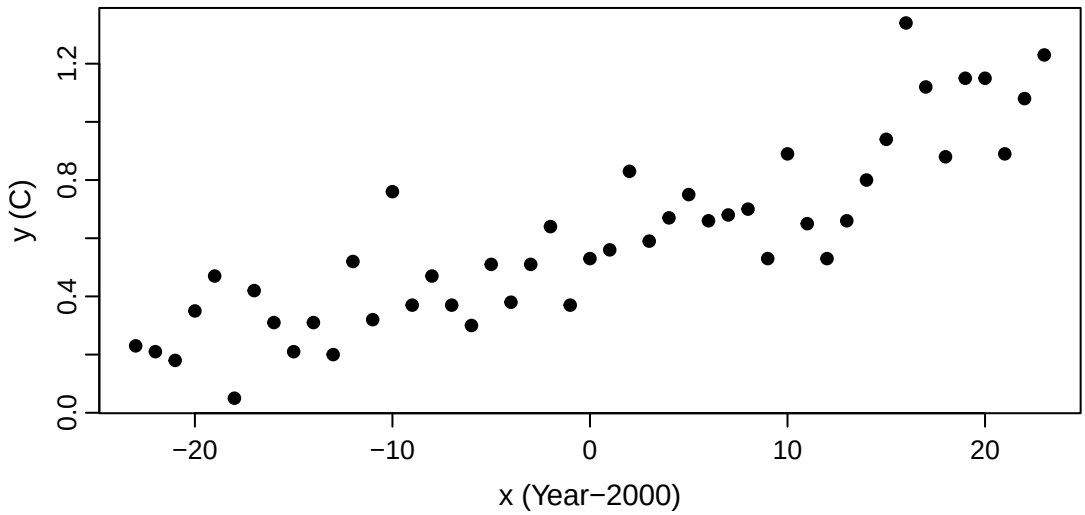
$$E_{\text{Young}} = \frac{15682}{30036} \times 3000 = 1566.3$$

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

Continue on page 28

Exercise XII

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



As a first approach a simple linear regression model

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

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

```
fit = smf.ols(formula = 'y ~ x', data = dat).fit()
print(fit.summary(slim=True))
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.754
Model:                            OLS   Adj. R-squared:                  0.748
No. Observations:                  47   F-statistic:                     137.7
Covariance Type:            nonrobust   Prob (F-statistic):           2.76e-15
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.6015      0.023     26.639      0.000       0.556       0.647
x              0.0195      0.002     11.734      0.000       0.016       0.023
==============================================================================
```

```
print(round(np.sqrt(fit.scale),4))
0.1548
print(fit.pvalues)
Intercept    3.420248e-29
x            2.758270e-15
dtype: float64
```

Question XII.1 (23)

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

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

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

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

4* ☐ Y_i and ε_i are independent.

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

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

The assumption is $\varepsilon_i \sim N(0, \sigma^2)$ and i.i.d., hence Answer 1 is correct, Answer 2 is the first "i" in i.i.d, Answer 3 just state that the variance is the same for all i, hence also correct.

For answer 4 consider

$$Cov(Y_i, \varepsilon_i) = Cov(\beta_0 + \beta_1 x_i + \varepsilon_i, \varepsilon_i) = Cov(\varepsilon_i, \varepsilon_i) = V(\varepsilon_i) > 0 \quad (1)$$

hence Y_i and ε_i are not independent.

For Answer 5, note that if $\varepsilon_i \sim N(0, \sigma^2)$ then $\varepsilon_i + a \sim N(a, \sigma^2)$ and hence 5 is also true.

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

Question XII.2 (24)

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

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

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

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

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

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

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

The p-value for the slope β_1 is far below $\alpha = 0.05$, as such we must reject the null hypothesis $H_0 : \beta 1 = 0$. As such time has a statistically significant effect on temperature.

The answer where the p-value is compared to the significance level is the correct argument, and since it's lower the β_1 is significant different from zero, hence relationship is significant.

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

Continue on page 31

Question XII.3 (25)

According to the model, in what year will the expected value of the temperature be 1 degree higher than the temperature in 2000 estimated by the model?

1* ☐ 2051

2 ☐ 2065

3 ☐ 2075

4 ☐ 2102

5 ☐ 2215

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

y increase with $\hat{\beta} = 0.0195$ per year according to the model, hence

$$1 = \hat{\beta}_1 x_{1\text{degree}}$$
$$1/\hat{\beta}_1 = x_{1\text{degree}}$$

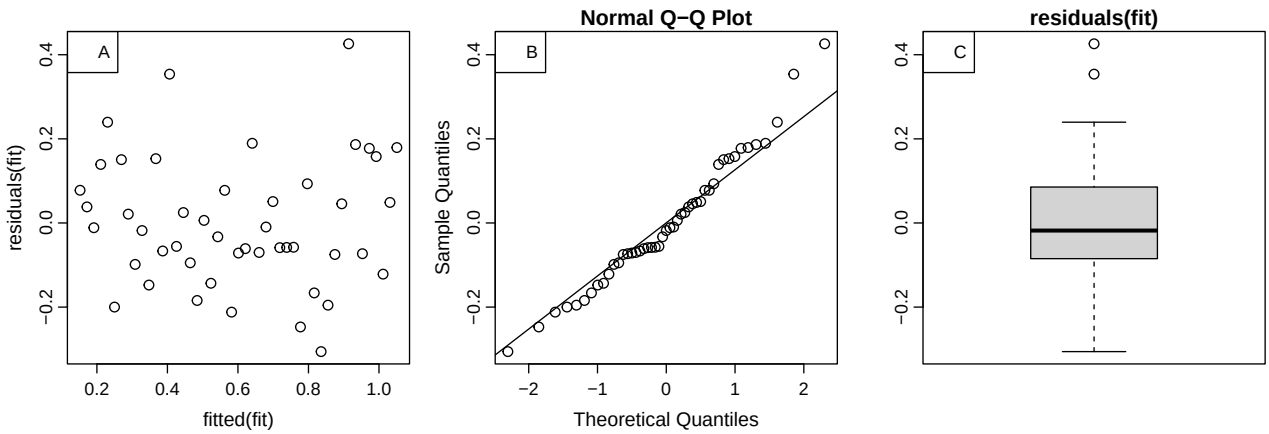
Therefore,

$$x_{1\text{degree}} = \frac{1}{\hat{\beta}_1} + 2000 = \frac{1}{0.0195} + 2000 = 2051.3$$

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

Question XII.4 (26)

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



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

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

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

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

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

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

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

Plot B cannot be used for assessing independence or variance homogeneity hence 1 and 5 are not correct. Plot C is a summary of all residuals hence it cannot be used for assessing if residuals are identically distributed or for systematic patterns, so 2 and 4 are not correct. Plot A can be used for identifying systematic patterns in the residuals, and there does not appear to be any, so answer 3 is correct.

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

Question XII.5 (27)

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

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

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

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

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

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

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

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

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

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

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

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

It appear that we will need the test statistics

$$A = \frac{0.5472833}{0.0324647} = 16.86$$

$$B = \frac{0.0195317}{0.0015949} = 12.25$$

$$C = \frac{0.0002946}{0.0001315} = 2.24$$

hence only answer 2 and 5 could be correct. `C` should be compared to the critical value and since `C` is greater than the critical value then $\hat{\beta}_2$ is significantly different from 0 (answer 2).

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

Question XII.6 (28)

For the model:

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

Which of the following, could be the top 10 rows in the corresponding Design Matrix?

1 □

$$\begin{bmatrix} 1 & -23.00 \\ 1 & -22.00 \\ 1 & -21.00 \\ 1 & -20.00 \\ 1 & -19.00 \\ 1 & -18.00 \\ 1 & -17.00 \\ 1 & -16.00 \\ 1 & -15.00 \\ 1 & -14.00 \\ \vdots & \vdots \end{bmatrix}$$

2* □

$$\begin{bmatrix} 1.00 & -23.00 & 529.00 \\ 1.00 & -22.00 & 484.00 \\ 1.00 & -21.00 & 441.00 \\ 1.00 & -20.00 & 400.00 \\ 1.00 & -19.00 & 361.00 \\ 1.00 & -18.00 & 324.00 \\ 1.00 & -17.00 & 289.00 \\ 1.00 & -16.00 & 256.00 \\ 1.00 & -15.00 & 225.00 \\ 1.00 & -14.00 & 196.00 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

3 □

$$\begin{bmatrix} 0.23 & -23.00 & 529.00 \\ 0.21 & -22.00 & 484.00 \\ 0.18 & -21.00 & 441.00 \\ 0.35 & -20.00 & 400.00 \\ 0.47 & -19.00 & 361.00 \\ 0.05 & -18.00 & 324.00 \\ 0.42 & -17.00 & 289.00 \\ 0.31 & -16.00 & 256.00 \\ 0.21 & -15.00 & 225.00 \\ 0.31 & -14.00 & 196.00 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

4 ☐

$$\begin{bmatrix} 1.00 & 0.23 \\ 2.00 & 0.21 \\ 3.00 & 0.18 \\ 4.00 & 0.35 \\ 5.00 & 0.47 \\ 6.00 & 0.05 \\ 7.00 & 0.42 \\ 8.00 & 0.31 \\ 9.00 & 0.21 \\ 10.00 & 0.31 \\ \vdots & \vdots \end{bmatrix}$$

5 ☐

$$\begin{bmatrix} 1 & -23 & 529 \\ 2 & -22 & 484 \\ 3 & -21 & 441 \\ 4 & -20 & 400 \\ 5 & -19 & 361 \\ 6 & -18 & 324 \\ 7 & -17 & 289 \\ 8 & -16 & 256 \\ 9 & -15 & 225 \\ 10 & -14 & 196 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

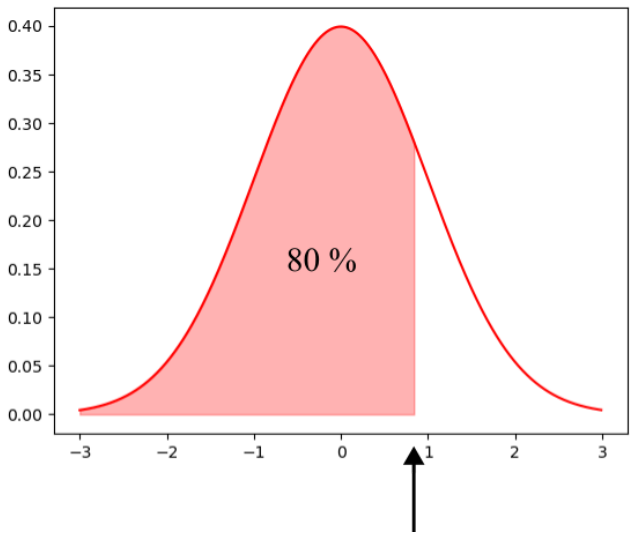
---------- FACIT-BEGIN ----------

The first column is set to 1 (these values are multiplied by the intercept β_0), while the other two columns should follow x_i (these values are multiplied by β_1) and x_i^2 (these values are multiplied by β_2) respectively. As such option 2 is the best fit (the values in the third column equals the square of the values in the middle column).

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

Exercise XIII

We want to find the 0.80-quantile (80% percentile) of a standard normal distribution.



Question XIII.1 (29)

Which of the following Python codes computes the 0.80-quantile of a standard normal distribution?

1 ☐ `stats.uniform.ppf(0.80, loc = 0, scale = 1)`

2 ☐ `stats.t.cdf(0.80, loc = 0, scale = 1)`

3 ☐ `stats.norm.cdf(0.80, loc = 0, scale = 1)`

4* ☐ `stats.norm.ppf(0.80, loc = 0, scale = 1)`

5 ☐ `stats.norm.pdf(0.80, loc = 0, scale = 1)`

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

Answer 4 is correct: To compute the 0.80-quantile we need the inverse cdf of the *standard normal distribution* (stats.norm.ppf) with mean 0 (loc = 0) and standard deviation 1 (scale = 1). The first input to the function (0.80) is the integral (indicated by the red area in the plot).

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

Continue on page 37

Exercise XIV

10 individuals recorded their systolic blood pressure levels in the morning on January 1 and July 1. From these measurements, we aim to explore whether there's a significant difference in systolic blood pressure between winter and summer. It can be assumed that the systolic blood pressure measurements in winter and summer follow a normal distribution.

Question XIV.1 (30)

Which is the most appropriate analysis?

1 ☐ Test of difference between two proportions

2 ☐ t-test using a pooled variance

3 ☐ Welch t-test

4* ☐ Paired t-test

5 ☐ Test using a binomial distribution

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

Since it's the same 10 individuals on which the measurements are done, then the two samples must be paired.

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

The exam is finished.