Technical University of Denmark   *Page 1 of 26 pages.*

Written examination: (26. June 2024)

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

Duration: 4 hours

Aids and facilities allowed: All

The questions were answered by

_____   _____   _____

(student number)   (signature)   (table number)

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

5 points are given for a correct "multiple choice" answer, and -1 point is given for a wrong answer. ONLY the following 5 answer options are valid: 1, 2, 3, 4, or 5. If a question is left blank or an invalid answer is entered, 0 points are given for the question. Furthermore, if more than one answer option is selected for a single question, which is in fact technically possible in the online system, 0 points are given for the question. The number of points needed to obtain a specific mark or to pass the exam is ultimately determined during censoring.

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

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

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

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

The exam paper contains 26 pages.

Multiple choice questions: *Note that in each question, one and <u>only</u> 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 R.*

Exercise I

Two students are counting the number of cars passing by on different stretches of road. They assume that the number of cars passing by in specific time intervals follow Poisson distributions. On the first road (road 1) they assume that the expected number of cars passing by is $\lambda_1 = 10/\text{hour}$, while on the second road (road 2) they assume that the expected number of cars passing by is $\lambda_2 = 15/\text{hour}$.

Now they define two random variables:

- X_1: number of cars passing by on road 1 in 15 minutes

- X_2: number of cars passing by on road 2 in 10 minutes.

You can assume that X_1 and X_2 are independent.

Question I.1 (1)

What is the probability $P(X_1 = 10)$?

1 ☐   0.125

2 ☐   0.417

3 ☐   0.000216

4 ☐   0.875

5 ☐   0.583

Question I.2 (2)

Which of the following statements about the expected values of the two random variables is correct?

1 ☐ $\frac{E[X_1]}{E[X_2]} = 1.5$

2 ☐ $\frac{E[X_1]}{E[X_2]} = \frac{2}{3}$

3 ☐ $\frac{E[X_1]}{E[X_2]} = \frac{1}{3}$

4 ☐ $\frac{E[X_1]}{E[X_2]} = 3$

5 ☐ $\frac{E[X_1]}{E[X_2]} = 1$

Question I.3 (3)

What is the probability that the time between two cars passing by is greater than 2 minutes on road 2?

1 ☐   0.5

2 ☐   0.184

3 ☐   0.607

4 ☐   0.368

5 ☐   0.816

A farm made a study in which 225 chickens were randomly divided into 3 treatment groups of 75 animals each. Each group were fed with fodder from different feed producers during a period. For each chicken the weight change over the period of time was measured and the final data set consists of 225 observations of weight changes. The objective of the study is to determine if there is statistical evidence for difference in mean weight change for at least one of the groups. It may be assumed that the variance is the same in all treatment groups.

Question II.1 (4)

What kind of statistical analysis is most suitable for this?

1 □   Multiple linear regression analysis

2 □   Test for independence in a $r \times c$ frequency table (Contingency table)

3 □   Paired t-tests

4 □   One-way analysis of variance

5 □ t-tests

4

Exercise III

The engineers at an international airport have conducted a survey, in which they have timed 40 randomly selected security checks. The average duration of the security checks included in the survey was 34.66 seconds, and the sample standard deviation was 10.12 seconds, it is assumed that the times are normally distributed.

Question III.1 (5)

Based on the survey, what is the 99% confidence interval for the mean duration of the security checks?

1 ☐ $[7.26; 62.06]$

2 ☐ $[14.19; 55.13]$

3 ☐ $[30.33; 38.99]$

4 ☐ $[31.42; 37.90]$

5 ☐ $[33.06; 36.26]$

Question III.2 (6)

What is the p-value for the usual test of the null hypothesis $H_0 : \mu = 30$ against a two-sided alternative hypothesis?

1 ☐ 0.30%

2 ☐ 0.59%

3 ☐ 4.72%

4 ☐ 94.23%

5 ☐ 99.70%

5

Exercise IV

A company wants to estimate the cost of producing solar panels. Therefore, the engineers have designed an experiment to evaluate the costs of producing batches of different sizes and ensure that the observations are completely independent. The results are given in the table below.

Batch size (units)	50	100	150	200	250	300	350	400	450	500
Costs (M DKK)	2.33	4.21	6.01	7.51	8.46	8.93	9.45	10.70	10.55	10.74

The data can be read in R using the below code:

```
Batch<-1:10 * 50
Costs<-c(2.33,4.21,6.01,7.51,8.46,8.93,9.45,10.70,10.55,10.74)
```

The engineers initially believe that the data can be described by a linear model on the form

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i \in \{1, ..., 10\},$$

where the errors, ϵ_i, are independent and identically distributed (i.i.d.) with a $N(0, \sigma^2)$ distribution. In the model, the response variable is the cost (in M DKK) and the explanatory variable is the batch size (in units). The engineers therefore fit a linear regression model using the least squares method.

Question IV.1 (7)

What proportion of the variation in the costs is explained by the regression model?

1 ☐ 89.4%

2 ☐ 90.6%
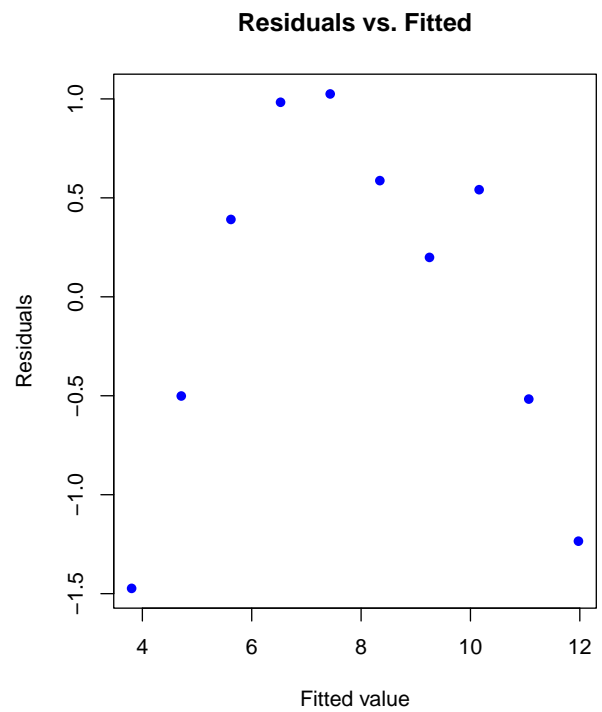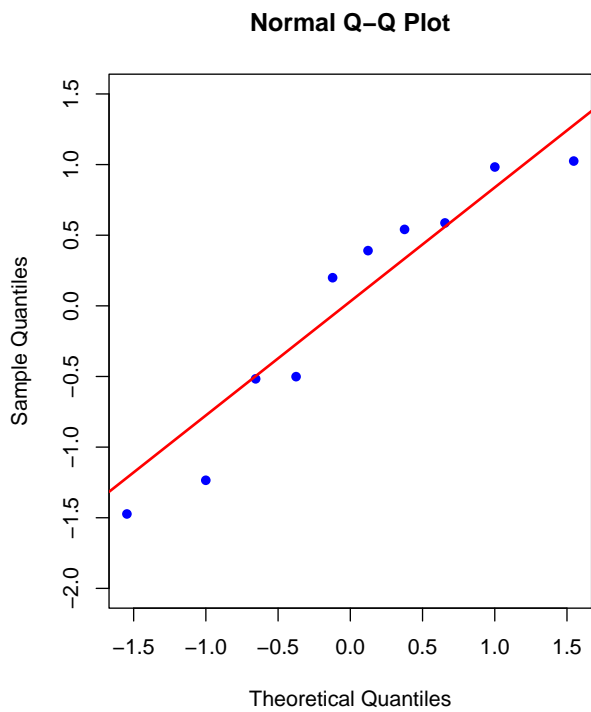
3 ☐ 93.9%

4 ☐ 95.2%

5 ☐ 96.9%

Question IV.2 (8)

What is the 99% confidence interval for the slope, β_1?

1 ☐ $[0.011, 0.025]$

2 ☐ $[0.013, 0.023]$

3 ☐ $[0.016, 0.020]$

4 ☐ $[0.742, 5.049]$

5 ☐ $[1.415, 4.375]$

To validate the model, the engineers make two diagnostic plots: A normal Q-Q plot and a plot of the residuals against the fitted values. The two plots are seen below:

Question IV.3 (9)

Which of the following statements regarding the validity of the model is most correct?

1 ☐   The diagnostic plots do not indicate any violations of the model assumptions

2 ☐   The normal Q-Q plot indicates that the residuals are not normally distributed

3 ☐   The residuals vs. fitted values plot indicates that the residuals are not normally distributed

4 ☐   The residuals vs. fitted values plot indicates that the residuals are not independent of the fitted values

5 ☐   The model assumptions must be satisfied as the R^2-value is high.

The number of persons living in Danish dorms in 2023 is provided by Statistics Denmark. Here we focus only on a few age-categories:

	males	females	Total
18-24	14048	14128	28176
25-29	8215	6028	14243
30-39	2735	1397	4132
Total	24998	21553	46551

Question V.1 (10)

What proportion of the residents of Danish dormitories are males (among the 18-39 year olds)?

1 ☐   0.463

2 ☐   0.500

3 ☐   0.521

4 ☐   0.537

5 ☐   0.409

Question V.2 (11)

We would like to know whether the proportions of males in different age-groups is significantly different (using a 5% significance level). Which of the following statements is true?

1 ☐   We should use a paired t-test with $\alpha = 0.05$ to test if there is a significant difference between age-groups. The result is that we do observe a significant difference.

2 ☐   We should use an unpaired t-test with $\alpha = 0.025$ to test if there is a significant difference between age-groups. The result is that we do observe a significant difference.

3 ☐   We should use an unpaired t-test with $\alpha = 0.05$ to test if there is a significant difference between age-groups. The result is that we do NOT observe a significant difference.

4 ☐   We should use a χ^2 test with 6-degrees of freedom to test if there is a significant difference between age-groups. The result is that we do observe a significant difference.

5 ☐   We should use a χ^2 test with 2-degrees of freedom to test if there is a significant difference between age-groups. The result is that we do observe a significant difference.

Question V.3 (12)

Under the hypothesis that the distribution between for male and females is the same for all age groups, what is the expected number of males in the age group 18-24 living in dorms (to be compared with the table above and used for calculating the appropriate test-statistic)?

1 □   14088

2 □   15131

3 □   15517

4 □   14048

5 □   7759

Question V.4 (13)

We now consider only the 18-24 year-olds. Which of the following statements is true (if relevant, in the answer options, significance level $\alpha = 0.05$ is used)?

1 □   Among the 18-24 year-olds the proportion of males is NOT significantly different from 0.5, as the estimated 95% confidence interval for the proportion of males in this age-group is [0.493, 0.504]

2 □   Among the 18-24 year-olds the proportion of males is exactly 0.5.

3 □   Among the 18-24 year-olds the proportion of males is significantly different from 0.5, as the estimated 95% confidence interval for the proportion of males in this age-group is [0.501, 0.533]

4 □   Among the 18-24 year-olds the proportion of males is NOT significantly different from 0.5, as the estimated 95% confidence interval for the proportion of males in this age-group is [0.501, 0.533]

5 □   Among the 18-24 year-olds the proportion of males is significantly different from 0.5, as the estimated 95% confidence interval for the proportion of males in this age-group is [0.532, 0.542]

Question V.5 (14)

Assuming independence between individuals. What is the probability, that 100 or more females live in a dorm with 190 people, if the probability of an individual being a female is assumed to be 0.45?

1 ☐ The probability is 0.021

2 ☐ The probability is 0.015

3 ☐ The probability is 0.50

4 ☐ The probability is 0.45

5 ☐ The probability is 0.985

Exercise VI

A simple predator-prey model is the Lotka-Volterra model

$$\frac{dx}{dt} = \alpha x - \beta xy$$
$$\frac{dy}{dt} = \delta xy - \gamma y,$$

where x is the size of the prey population and y is size of the predator population. The equation allows for a constant of motion (i.e. the quantity will stay constant through time for given initial conditions) given by

$$K = y^\alpha e^{-\beta y} x^\gamma e^{-\delta x}.$$

Assume that $\alpha = 2/3$, $\beta = 4/3$, $\gamma = \delta = 1$, and that the predator and prey population sizes have been observed at $y = 1/2$ and $x = 1$ respectively. The uncertainties of the observations are assumed to be $\sigma_y^2 = 1/16^2$ and $\sigma_x^2 = 1/8^2$, and further the observations are assumed independent.

Question VI.1 (15)

Using the non-linear error propagation rule what is the approximation of the variance of K?

1 ☐   0.312

2 ☐   0

3 ☐   0.559

4 ☐   0.75

5 ☐   0.889

Question VI.2 (16)

Assume now that the predator population is observed without error ($\sigma_y^2 = 0$), and hence that the only source of uncertainty is the prey population. Using the variance from above ($\sigma_x^2 = 1/8^2$) and assmuming normality of the error in x, i.e. $X = x + \epsilon$ with $\epsilon \sim N(0, \sigma_x^2)$. In what interval does the standard deviation of K fall (the answer should not rely on the non-linear approximations of the error propagation rule and should rather be base on simulation)?

1 ☐ $(0.07, 0.1)$

2 ☐ $(0.12, 0.2)$

3 ☐ $(10^{-4}, 0.01)$

4 ☐ $(0.03, 0.05)$

5 ☐ $(0.3, 0.5)$

Exercise VII

Let the function $f(x)$ be defined by

$$f(x) = \alpha\phi_1(x) + \beta\phi_2(x),$$

where $\phi_i(x)$ is the probability density function of a normal random variable with mean μ_i and variance σ_i^2.

Question VII.1 (17)

Under what conditions is $f(x)$ a probability density function (the answer should apply for any value of $\sigma_i > 0$ and $\mu_i \in \mathbb{R}$)?

1 □ $\alpha = \beta = 1$

2 □ $\alpha \in [0, 2]$ and $\beta = 2 - \alpha$

3 □ $\alpha = \frac{\sigma_1^2}{\sigma_2^2}$ and $\beta = \frac{\sigma_2^2}{\sigma_1^2}$

4 □ $\alpha \in [0, 1]$ and $\beta = 1 - \alpha$

5 □ $\alpha = \frac{\mu_1^2}{\sigma_1^2}$ and $\beta = \frac{\mu_2^2}{\sigma_2^2}$

Exercise VIII

An aircraft manufacturer uses an expensive type of screws in the production of a certain model. To reduce production costs, the manufacturer considers replacing the expensive screws with a cheaper type of screws. Therefore, the manufacturer tests the tensile strength (MPa) of the two types of screws, and the results are shown in the below table.

Tensile strengt	Cheap	Expensive
Sample mean (MPa)	1250	1300
Sample standard deviation (MPa)	54.24	28.54
Sample size	25000	15000

Question VIII.1 (18)

Assuming the samples were completely random, what is the 95% confidence interval for the difference in mean tensile strengths (mean of the cheap type minus mean of the expensive type) based on the test results?

1 □ $[-50.07; -49.93]$

2 □ $[-50.13; -49.87]$

3 □ $[-50.34; -49.66]$

4 □ $[-50.68; -49.32]$

5 □ $[-50.81; -49.19]$

Question VIII.2 (19)

Under the null hypothesis $H_0 : \mu_{\text{cheap}} - \mu_{\text{expensive}} = -50$, what is the observed test statistic based on the test results?

1 □ -241.14

2 □ -120.57

3 □ -2.31

4 □ 0.00

5 □ 241.14

The manufacturer also considers buying a new and more fuel-efficient aircraft model, and therefore they have measured the fuel-consumption (in kg) of the two models on 10 popular routes under similar weather and weight conditions. The manufacturer is only interested in comparing the logarithm of the fuel consumption as given in the below table:

Log of fuel consumption	Current model	New model
Sapporo - Tokyo	7.964	7.932
Sydney - Melbourne	7.813	7.762
Mumbai - Delhi	8.299	8.243
Beijing - Shanghai	8.219	8.174
Paris - Montreal	9.832	9.782
Dubai - London	9.829	9.775
London - New York	9.842	9.794
New York - Los Angeles	9.498	9.445
Kuala Lumpur - Singapore	7.023	6.942
Cancun - Mexico City	8.408	8.347

The data can be read in R using the below code:

```
log_c <- c(7.964,7.813,8.299,8.219,9.832,9.829,9.842,9.498,7.023,8.408)
log_n <- c(7.932,7.762,8.243,8.174,9.782,9.775,9.794,9.445,6.942,8.347)
```

Question VIII.3 (20)

What is the p-value for the appropriate test of the null hypothesis $H_0 : \delta = \mu_{\text{current}} - \mu_{\text{new}} = 0.05$ against a two-sided alternative hypothesis? (Here μ_{current} and μ_{new} refer to the mean of the logarithm of the fuel consumption.)

1 □ $p < 0.001$

2 □ $p = 0.442$

3 □ $p = 0.452$

4 □ $p = 0.908$

5 □ $p = 0.995$

16

A car manufacturer wants to find out if there is a difference in breaking strength in beams made with metal from different suppliers. Let Y represent the breaking strength of beams. In the following breaking strength are measured on beams each made with metal from a single supplier. Metal from four different suppliers were included in the study and the breaking strength was measured for 5 similar beams from each supplier:

Supplier A	Supplier B	Supplier C	Supplier D
92.0	131.0	74.1	90.4
111.6	103.5	52.8	95.2
98.4	100.0	82.5	87.6
87.7	84.7	94.7	63.2
134.9	134.5	107.3	119.5

Question IX.1 (21)

The engineers in the company have conducted the following analysis in R. What is the conclusion, at significance level $\alpha = 5\%$, about the difference in breaking strength of the test beams made with metal from the different suppliers (both conclusion and argument must be correct)?

```
anova(lm(y ~ Supplier))

## Analysis of Variance Table
##
## Response: y
##          Df Sum Sq Mean Sq F value Pr(>F)
## Supplier  3 2508.8  836.25   2.027 0.1507
## Residuals 16 6601.0  412.56
```

1 ☐ A significant difference in breaking strength <u>is not found</u>, since the p-value is <u>greater</u> than the significance level.

2 ☐ A significant difference in breaking strength <u>is found</u>, since the p-value is <u>greater</u> than the significance level.

3 ☐ A significant difference in breaking strength <u>is not found</u>, since the p-value is <u>less</u> than the significance level.

4 ☐ A significant difference in breaking strength <u>is found</u>, since the p-value is <u>less</u> than the significance level.

5 ☐ None of the above conclusions are correct.

The engineers now remembered that the strength tests were made on different days and that the conditions (e.g. weather) might differ between days. Luckily someone already thought

about this and in the tests during one day, exactly one beam from each supplier was tested. Therefore they could group the observations on days.

	Supplier A	Supplier B	Supplier C	Supplier D
Day 1	92.0	131.0	74.1	90.4
Day 2	111.6	103.5	52.8	95.2
Day 3	98.4	100.0	82.5	87.6
Day 4	87.7	84.7	94.7	63.2
Day 5	134.9	134.5	107.3	119.5

Which resulted in the following analysis (note, some of the values in the result have been replaced by letters and any eventual * in the result have been removed):

```
anova(lm(y ~ Supplier + Day))

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Supplier   3   A     836.25  3.8696    E
## Day        4   B    1001.92  4.6362    F
## Residuals 12   C        D
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question IX.2 (22)

What conclusion can be drawn, at the significance level $\alpha = 5\%$, from this analysis (both conclusion and argument must be correct)?

1 ☐ There is not a significant effect of neither supplier nor day, since the relevant p-values are 0.076 and 0.141 respectively.

2 ☐ There is a significant effect of supplier, but not of day, since the relevant p-values are 0.038 and 0.141 respectively.

3 ☐ There is a significant effect of both supplier and day, since the relevant p-values are 0.038 and 0.017 respectively.

4 ☐ There is not a significant effect of supplier, but there is a significant effect of day, since the relevant p-values are 0.076 and 0.017 respectively.

5 ☐ There is not a significant effect of neither supplier nor day, since the relevant p-values are 0.892 and 0.112 respectively.

Question IX.3 (23)

What is total sum of squares (SST)?

1 ☐ 216

2 ☐ 2509

3 ☐ 4008

4 ☐ 6516

5 ☐ 9110

A person (investor 1) decides to invest in stocks. The investor invest 10000 kr. in 10 different stocks (1000 kr. in each), after one year the investor sells all the stocks. The returns (measured in kr.) of the 10 stocks is given in the table below.

Stock	1	2	3	4	5	6	7	8	9	10
Return	1144	1218	1480	747	1178	-121	-382	-24	-652	-32

The investor wants to investigate if the investment has been a success. The success criterion is that the return is significantly different from (using a two-sided test) and bigger than a 2% (200 kr. in total, or 20 kr. per stock) return using significance level $\alpha = 5\%$.

The investor has dicided to use a test with no distributional assumption on the population, it is however assumed that the returns are independent. Some of the R-code below should be used for the next question.

```
x <- c(1144,   1218,   1480,    747,  1178, -121, -382,   -24, -652, -32)

k <- 10000
sim.samp <- replicate(k, sample(x, replace = TRUE))

quantile(apply(sim.samp, 2, mean), prob = c(0.025, 0.05, 0.95, 0.975),
         type = 2)

##    2.5%      5%     95%  97.5%
##    0.15   68.60 832.95 905.40

t.test(x)

##
##   One Sample t-test
##
## data:  x
## t = 1.8531, df = 9, p-value = 0.09687
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   -100.5569 1011.7569
## sample estimates:
## mean of x
##     455.6

t.test(x, mu = 200)

##
##   One Sample t-test
##
```

```
## data:  x
## t = 1.0396, df = 9, p-value = 0.3256
## alternative hypothesis: true mean is not equal to 200
## 95 percent confidence interval:
##   -100.5569 1011.7569
## sample estimates:
## mean of x
##     455.6
```

Question X.1 (24)

What is the result of the test (both the conclusion and the argument must be correct)?

1 □ The null-hypothesis is accepted as the p-value is 0.3

2 □ The null-hypothesis is accepted as $20 \in [0.15, 905.4]$

3 □ The null-hypothesis is rejected as the p-value is 0.1

4 □ The null-hypothesis is rejected as $0 < 0.15$

5 □ The null-hypothesis is rejected as $0 \in [-101, 1012]$

Another investor has invested in 10 different stocks, also with 1000 kr. in each stock. The two investors want to compare the returns of their investments. In the R-code below x is the returns for investor 1 and y is the returns for investor 2.

```
k <- 10000
sim.x <- replicate(k, sample(x, replace = TRUE))
sim.y <- replicate(k, sample(y, replace = TRUE))
sim.diff <- replicate(k, sample(x - y, replace = TRUE))

quantile(apply(sim.x, 2, mean) - apply(sim.y, 2, mean), prob = c(0.025,0.975),
         type = 2)

##       2.5%       97.5%
## -1562.82480    99.26984

quantile(apply(sim.diff, 2, mean), prob = c(0.025,0.975))

##       2.5%       97.5%
## -1186.2022   -285.1908

t.test(x, y, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = -3.0451, df = 9, p-value = 0.0139
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -1298.7101  -191.5961
## sample estimates:
## mean difference
##        -745.1531
```

The two investors want to compare the returns of their investments using a statistical method with as few assumptions as possible, e.g. with no distributional assumptions at all, it is however assumed that all observations are independent.

Question X.2 (25)

Is there, based on the R-output above, a significant (using significance level $\alpha = 5\%$) difference between the two investors returns (both the conclusion and the argument should be correct)?

1 □   Yes, since $0 \notin [-1186, -285]$

2 □   Yes, since $0.014 < 0.05$

3 □   Yes, since $0 \notin [-1299, -192]$

4 □   No, since $0.014 < 0.05$

5 □   No, since $0 \in [-1562, 99]$

The engineers at a construction company is developing a model for the lifespan of buildings. The model can be written as

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i,$$

where Y_i (Life) is the lifespan (in years) of building i, x_{1i} (Econ_RF) the economical risk factor of building i, x_{2i} (Matr_RF) the material risk factor of building i, and x_{3i} (Dsgn_RF) the design risk factor of building i. Furthermore, the model assumes that the errors, ϵ_i, are independent and identically distributed with a $N(0, \sigma^2)$ distribution.

The engineers have fitted the model using the method of least squares, and the output from R is given below:

```
##
## Call:
## lm(formula = Life ~ Econ_RF + Matr_RF + Dsgn_RF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5709  -3.3908  -0.0053   3.3407  15.3690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 151.76054    1.20254 126.199   <2e-16 ***
## Econ_RF      -0.35638    0.01270 -28.055   <2e-16 ***
## Matr_RF      -0.90717    0.01398 -64.878   <2e-16 ***
## Dsgn_RF      -0.11806    0.01289  -9.156   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.992 on 165 degrees of freedom
## Multiple R-squared:  0.9681,Adjusted R-squared:  0.9675
## F-statistic:  1669 on 3 and 165 DF,  p-value: < 2.2e-16
```

You may assume that the model assumptions are satisfied.

Question XI.1 (26)

One particular building had a lifespan of 81.3 years. The building had an economical risk factor of 64.7, a material risk factor of 55.2, and a design risk factor of 28.1. What is the residual associated with this building?

1 ☐ The residual cannot be determined without further information

2 ☐ 6.0 years

3 ☐ −6.0 years

4 ☐ 5.0 years

5 ☐ −5.0 years

Question XI.2 (27)

How many buildings (observations) were included in the data set used to fit the model?

1 ☐ 165

2 ☐ 166

3 ☐ 167

4 ☐ 168

5 ☐ 169

The model can be written in matrix-vector notation as

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

with the i'th row of $\boldsymbol{X}$ being $\begin{bmatrix} 1 & x_{1i} & x_{2i} & x_{3i} \end{bmatrix}$.

Question XI.3 (28)

What are the diagonal elements of the matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$?

1 $\square$ $(0.058, 6.5 \cdot 10^{-6}, 7.8 \cdot 10^{-6}, 6.7 \cdot 10^{-6})$

2 $\square$ $(1.2, 0.013, 0.014, 0.013)$

3 $\square$ $(36, 0.0040, 0.0049, 0.0041)$

4 $\square$ $(1.4, 0.00016, 0.00020, 0.00017)$

5 $\square$ $(17.2, 154380, 127405, 149863)$

Let $X_i \sim N(\mu_1, \sigma^2)$, be i.i.d. random variables with $i \in \{1, 2, ..., n_1\}$ and $Y_j \sim N(\mu_2, 2\sigma^2)$, be i.i.d. random variables with $j \in \{1, 2, ..., n_2\}$, and further let X_i and Y_j be independent. Also let S_x^2 and S_y^2 denote the usual sample variances, and finally let

$$S^2 = \alpha S_x^2 + (1 - \alpha)S_y^2.$$

Question XII.1 (29)

Assuming $n_1 > 1$, $n_2 > 1$, which value of α minimizes $V(S^2)$?

1 □ $\alpha = \frac{2n_1 - 2}{2n_2 + n_1 - 3}$

2 □ $\alpha = \frac{4n_1 - 4}{4n_1 + n_2 - 5}$

3 □ $\alpha = \frac{n_2 - 1}{n_1 + n_2 - 2}$

4 □ $\alpha = \frac{2n_2 - 2}{2n_2 + n_1 - 3}$

5 □ $\alpha = \frac{n_1 - 1}{n_1 + n_2 - 2}$

Question XII.2 (30)

Assuming $n_1 = n_2 = 2$ what is q such that $P(S_y^2/S_x^2 \leq q) = 0.99$?

1 □ 6995

2 □ 7475

3 □ 8023

4 □ 8087

5 □ 8104

The exam is finished. Enjoy the summer!