Written examination: 14. August 2019

Course name and number: **Introduction to Statistics (02323)**

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 11 exercises. To answer the questions, you need to fill in the "multiple choice" form (6 separate pages) on CampusNet with the numbers of the answers that you believe to be correct.

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 online via CampusNet. 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	II.2	II.3	III.1	III.2	IV.1	IV.2	IV.3
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer										

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

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

The exam paper contains 24 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.*

Exercise I

Assume that $X_1, \ldots, X_{25}$ are independent random variables, which are normal distributed with $N(5, 2^2)$.

Question I.1 (1)

Which of the following values has the property: The probability that X_1 is lower than this value is 15% (remember that the answer can be rounded)?

1 ☐ -0.85

2 ☐ 0.85

3 ☐ 2.93

4 ☐ 3.93

5 ☐ 5.43

Question I.2 (2)

What is the probability that the sample mean $\bar{X} = \frac{1}{25}\sum_{i=1}^{25} X_i$ is greater than 4.5?

1 ☐ $P(\bar{X} > 4.5) = 0.89$

2 ☐ $P(\bar{X} > 4.5) = 0.85$

3 ☐ $P(\bar{X} > 4.5) = 2.05 \times 10^{-10}$

4 ☐ $P(\bar{X} > 4.5) = 0.18$

5 ☐ $P(\bar{X} > 4.5) = 0.55$

Continue on page 3

Exercise II

Given Lambert Beer's law the absorbance of light through a liquid solution can be calculated as

$$A = \gamma \cdot l \cdot c$$

where γ is a constant, l the path length through the liquid and c the concentration of solution.

Question II.1 (3)

Given that the standard deviation of the path length σ_l and the standard deviation of the concentration σ_c are known, the standard deviation of the absorbance can be approximated by which of the following formulas?

1 ☐ $(\frac{\partial A}{\partial c})^2\sigma_l^2 + (\frac{\partial A}{\partial l})^2\sigma_c^2$

2 ☐ $\sqrt{(\frac{\partial A}{\partial c})^2\sigma_l^2 + (\frac{\partial A}{\partial l})^2\sigma_c^2}$

3 ☐ $(\frac{\partial A}{\partial l})^2\sigma_l^2 + (\frac{\partial A}{\partial c})^2\sigma_c^2$

4 ☐ $\sqrt{(\frac{\partial A}{\partial l})^2\sigma_l^2 + (\frac{\partial A}{\partial c})^2\sigma_c^2}$

5 ☐ $\sqrt{(\frac{\partial A}{\partial c})^2\sigma_l^2} + \sqrt{(\frac{\partial A}{\partial l})^2\sigma_c^2}$

Question II.2 (4)

In an experiment the mean path length is determined to be 1 cm with a standard deviation of 0.1 cm. The average concentration is determined to be 0.65 M with a standard deviation of 0.09 M. γ is given as 0.3 $M^{-1}cm^{-1}$. Which of the following simulations can be used to determine the standard deviation of the absorbance?

1 ☐

```
k = 10000
e = 0.3
l = rnorm(k, 1, 0.1)
c = rnorm(k, 0.65, 0.09)
A = e*l*c
var(A)
```

2 ☐

```
e = 0.3
l = rnorm(1, 1, 0.1^2)
c = rnorm(1, 0.65, 0.09^2)
A = e*l*c
sd(A)
```

3 ☐
```
e = 0.3
l = rnorm(1, 1, 0.1^2)
c = rnorm(1, 0.65, 0.09^2)
A = e*l*c
var(A)
```

4 ☐
```
k = 10000
e = 0.3
l = rnorm(k, 1, 0.1^2)
c = rnorm(k, 0.65, 0.09^2)
A = e*l*c
sd(A)
```

5 ☐
```
k = 10000
e = 0.3
l = rnorm(k, 1, 0.1)
c = rnorm(k, 0.65, 0.09)
A = e*l*c
sd(A)
```

Question II.3 (5)

In the question above a random sample from a normal distribution was simulated using the command `rnorm`. Which of the commands below can be used to simulate a random sample from the standard normal distribution of length `n`?

1 ☐ `pnorm(runif(n))`

2 ☐ `qnorm(runif(n))`

3 ☐ `dnorm(runif(n))`

4 ☐ `qnorm(punif(n))`

5 ☐ `pexp(n)`

Continue on page 5

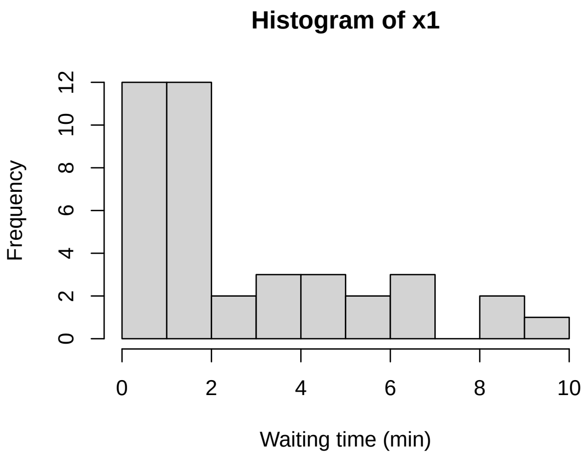
Exercise III

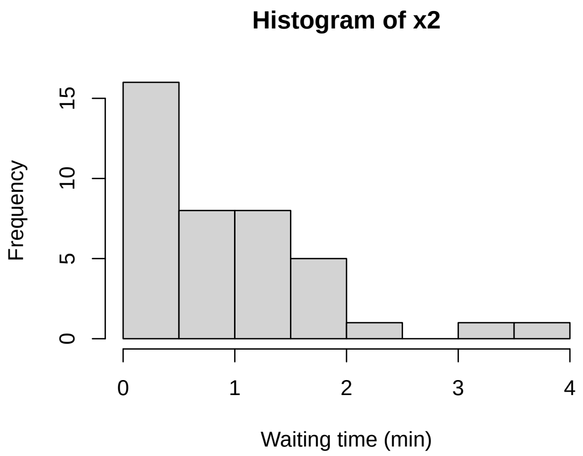
The human ressource department of a supermarket chain is interested in comparing waiting times for customers in two local shops. The waiting times (in minutes) of 40 customers have been measured in the two shops during an afternoon from 4 PM to 5 PM.

Let $X_{1,i}$ represent the i'th observed waiting time in Store 1. It can be assumed to follow an exponential distribution $X_{1,i} \sim Exp(\lambda_1)$ where $i = 1, \ldots, 40$.

Let $X_{2,i}$ represent the i'th observed waiting time in Store 2. It can be assumed to follow an exponential distribution $X_{2,i} \sim Exp(\lambda_2)$ where $i = 1, \ldots, 40$.

The data from each sample is stored in `x1` and `x2`, respectively, and a histogram of each sample is plotted below:





The average waiting times (in minutes) for the two shops are:

```
mean(x1)

## [1] 2.76

mean(x2)

## [1] 0.897
```

Question III.1 (6)

Estimate the rate parameters λ_1 and λ_2. The rates should be calculated in customers per hour (h^{-1}).

1 ☐ $\hat{\lambda}_1 = 2.76\, h^{-1}$ and $\hat{\lambda}_2 = 0.90\, h^{-1}$

2 ☐ $\hat{\lambda}_1 = 0.36\, h^{-1}$ and $\hat{\lambda}_2 = 1.11\, h^{-1}$

3 ☐ $\hat{\lambda}_1 = 19.54\, h^{-1}$ and $\hat{\lambda}_2 = 25.45\, h^{-1}$

4 ☐ $\hat{\lambda}_1 = 21.74\, h^{-1}$ and $\hat{\lambda}_2 = 66.89\, h^{-1}$

5 ☐ It is not possible to estimate λ_1 and λ_2.

Question III.2 (7)

At two other shops similar samples were collected. The rates were estimated to $\hat{\lambda}_1 = 0.23$ min^{-1} for the first shop and $\hat{\lambda}_2 = 0.39$ min^{-1} for the second.

To find out if there was a significant difference in mean waiting time at two shops the following calculations was carried out in R:

```
k <- 10000
simX1_samples <- replicate(k, rexp(40, 0.23))
simX2_samples <- replicate(k, rexp(40, 0.39))
sim_dif_means <- apply(simX1_samples, 2, mean) - apply(simX2_samples, 2, mean)

quantile(sim_dif_means, c(0.005, 0.995))

##   0.5%  99.5%
## -0.121  3.955

quantile(sim_dif_means, c(0.025, 0.975))

##  2.5% 97.5%
##  0.30  3.42

quantile(sim_dif_means, c(0.05, 0.95))

##    5%   95%
## 0.547 3.156
```

Which of the following statements is correct?

1 ☐ Non-parametric bootstrapping was carried out. The 95% confidence interval is [-0.121, 3.955] and contains zero, hence the mean waiting times are significantly different.

2 ☐ Non-parametric bootstrapping was carried out. The 95% confidence interval is [-0.121, 3.955] and contains zero, hence the mean waiting times are NOT significantly different.

3 ☐ Parametric bootstrapping was carried out. The 95% confidence interval is [0.30, 3.42] and doesn't contain zero, hence the mean waiting times are significantly different.

4 ☐ Parametric bootstrapping was carried out. The 95% confidence interval is [0.30, 3.42] and doesn't contain zero, hence the mean waiting times are NOT significantly different.

5 ☐ Parametric bootstrapping was carried out. The 95% confidence interval is [0.547, 3.156] and doesn't contain zero, hence the mean waiting times are NOT significantly different.

Continue on page 8

Exercise IV

This exercise is about quality control in a company which produces hard disk drives for NAS ("Network Attached Storage"). The company would like to investigate the probability that a certain type of hard disk drive breaks down within the first three years of "typical use". The company chooses a random sample of 950 hard disk drives from their production line. They ask the customers who buy these drives to report it if a drive fails within the first three years of use. All the NAS hard disk drives are assumed to have the same probability p of failing within the first three years, and they are assumed to fail independently of each other.

Question IV.1 (8)

It was reported that altogether 92 of the hard disk drives failed within the first three years of their lifetime. Give the estimated standard error, $\hat{\sigma}_{\hat{p}}$, for the estimated proportion of hard disk drives which break down within the first three years.

1 ☐ $\hat{\sigma}_{\hat{p}} = 9.2 \cdot 10^{-5}$

2 ☐ $\hat{\sigma}_{\hat{p}} = 0.0031$

3 ☐ $\hat{\sigma}_{\hat{p}} = 0.0096$

4 ☐ $\hat{\sigma}_{\hat{p}} = 0.087$

5 ☐ $\hat{\sigma}_{\hat{p}} = 0.30$

Question IV.2 (9)

The company aims for 90% of their NAS hard disk drives to have a lifetime which exceeds three years. Using a statistical test, they would like to investigate whether they live up to this goal. Which statistical null hypothesis is then relevant to test?

1 ☐ $H_0: \ p = 0.1$

2 ☐ $H_0: \ p = 0.9$

3 ☐ $H_0: \ p \neq 0.1$

4 ☐ $H_0: \ p \neq 0.9$

5 ☐ None of the above hypotheses are applicable.

(The exercise text is continued)

Now, the company would like to compare the lifetime of their special NAS hard disk drives to the lifetime of regular hard disk drives (when these are used in a NAS setup). To this end, they present the following contingency table, which also includes data for the lifetime of 650 regular hard disk drives:

	NAS HDD	Regular HDD	Total
$<$ 1 year	10	7	17
1-2 years	33	45	78
2-3 years	49	69	118
$>$ 3 years	858	529	1387
Total	950	650	1600

This table summarizes how many of a given type of hard disk drive that failed within a certain age interval. For example, one can read from this table that 69 out of 650 regular hard disk drives broke down after 2-3 years of use. These data are to be used in the rest of the questions in this exercise.

Question IV.3 (10)

The company would like to investigate whether the two types of hard disk drives have the same probability of failing within the first three years of their lifetime. Which of the following snippets of R code carries out the relevant statistical test?

1 ☐ `prop.test(x = c(49, 69), n = c(950, 650), correct = FALSE)`

2 ☐ `prop.test(x = c(49, 69), n = c(858, 529), correct = FALSE)`

3 ☐ `prop.test(x = c(92, 950), n = c(121, 650), correct = FALSE)`

4 ☐ `prop.test(x = c(92, 121), n = c(950, 650), correct = FALSE)`

5 ☐ None of the above.

Question IV.4 (11)

The company could also have chosen to investigate whether the distribution of the number of drive failures in the four age intervals differs for the two types of hard disk drives. Under the corresponding null hypothesis H_0, what is the number of regular hard disk drives which are expected to fail after 1-2 years?

1 ☐ 29

2 ☐ 33

3 ☐ 39

4 ☐ 45

5 ☐ None of the above numbers are the correct answer.

Question IV.5 (12)

Suppose that the company actually carries out a χ^2-test to investigate whether the distribution of the number of drive failures in the four age intervals differs for the two types of hard disk drives. How many degrees of freedom does the χ^2-distribution, which is used in this test, have?

1 ☐ 1 degree of freedom

2 ☐ 2 degree of freedom

3 ☐ 3 degree of freedom

4 ☐ 6 degree of freedom

5 ☐ 9 degree of freedom

Continue on page 11

Exercise V

On a small island it is known that the rate of blackouts in the electrical system is one per week. Define the random variable X which denotes the number of blackouts for some randomly chosen week. The number of blackouts per week is assumed to follow a poisson distribution.

Question V.1 (13)

What is the variance of X?

1 ☐ $\sigma^2 = \frac{1}{7}$

2 ☐ $\sigma^2 = 0.368$

3 ☐ $\sigma^2 = 1$

4 ☐ $\sigma^2 = 2.72$

5 ☐ $\sigma^2 = 7$

Question V.2 (14)

What is the probability that there will be no blackout on a randomly chosen day?

1 ☐ 0.13

2 ☐ 0.24

3 ☐ 0.53

4 ☐ 0.76

5 ☐ 0.87

Continue on page 12

Exercise VI

You would like to compare 5 groups with 6 observations in each. You will do this by making a one-way analysis of variance and test the hypothesis that all groups have same mean value. The observations are assumed to be independent of each other. The test statistic for this test is 4.30.

Question VI.1 (15)

What is the p-value for this test?

1 ☐ 0.009

2 ☐ 0.0002

3 ☐ 0.03

4 ☐ 0.00001

5 ☐ 0.05

Continue on page 13

Exercise VII

The analysis of variance results from a one-way analysis of variance are:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatm	2	1.19	A	B	C
Residuals	9	4.53	D		

Question VII.1 (16)

As you can see some numbers are missing. The number replaced by B is:

1 ☐ 0.26

2 ☐ 1.18

3 ☐ 1.53

4 ☐ 2.20

5 ☐ 7.40

Question VII.2 (17)

It is stated that there were equally many observations in each group. How many observations were there in one of the groups?

1 ☐ 2

2 ☐ 3

3 ☐ 4

4 ☐ 5

5 ☐ 9

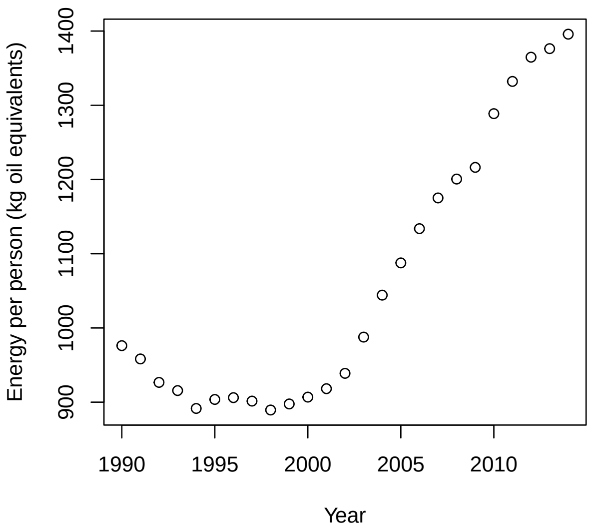
Continue on page 14

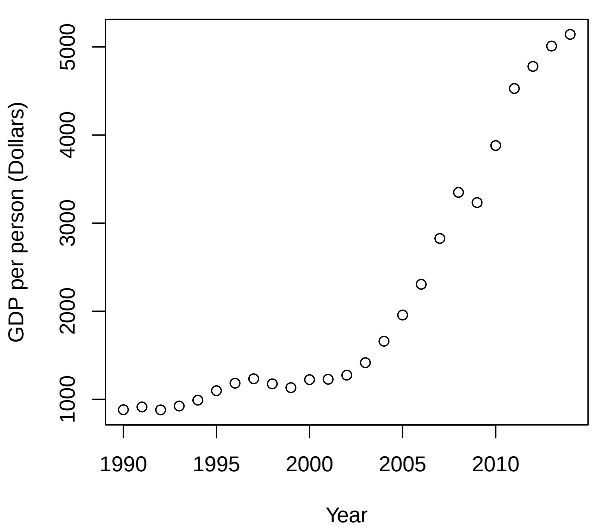
Exercise VIII

It is an engineering challenge to develop the technology that can cover the world's energy demand in a sustainable way. Considering The World Bank's population forecasts for 2050 one will reach the result that if everyone in 30 years should have the same energy demand, as the rich countries have now, then the energy demand will triple compared to 2014.

This exercise uses data retrieved from The World Bank, which categorizes the world's countries into the categories: low, middle and high income countries. The development of middle income countries is very important for the development of the world energy demand.

The following plot shows the Energy Consumption and Gross National Product (GDP) per year per person for middle income countries from 1990 to 2014:





The data consists of the plotted annual values stored in the vectors: `year` is the year, `energy` is energy demand and `gdp` is GDP. Only this data is used, thus all conclusions in the exercise apply only to middle income countries in this particular period.

First four summary statistics are calculated:

```
c(mean(energy), mean(gdp))

## [1] 1061 2169

c(sd(energy), sd(gdp))

## [1]  179 1465
```

Thereafter two different simple linear regression models are estimated:

```
summary(lm(energy ~ year))

##
## Call:
## lm(formula = energy ~ year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -122.49  -60.45    3.37   74.54  174.70
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42299.35    4669.35   -9.06  4.8e-09 ***
## year            21.66       2.33    9.29  3.0e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.1 on 23 degrees of freedom
## Multiple R-squared:  0.789,Adjusted R-squared:  0.78
## F-statistic: 86.2 on 1 and 23 DF,  p-value: 3.03e-09

summary(lm(energy ~ gdp))

##
## Call:
## lm(formula = energy ~ gdp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52.82 -29.23  -9.45  27.37  69.09
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.01e+02   1.35e+01    59.5   <2e-16 ***
## gdp         1.20e-01   5.18e-03    23.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.2 on 23 degrees of freedom
## Multiple R-squared:  0.959,Adjusted R-squared:  0.957
## F-statistic:  536 on 1 and 23 DF,  p-value: <2e-16
```

Question VIII.1 (18)

According to these results, what is estimated mean annual increase in energy demand in the period (in ”kg oil equivalents” per year)?

1 ☐ 0.120

2 ☐ 2.33

3 ☐ 3.37

4 ☐ 21.7

5 ☐ 801

Question VIII.2 (19)

What is the calculated correlation between the energy demand and the GDP?

1 ☐ 0.83

2 ☐ 0.93

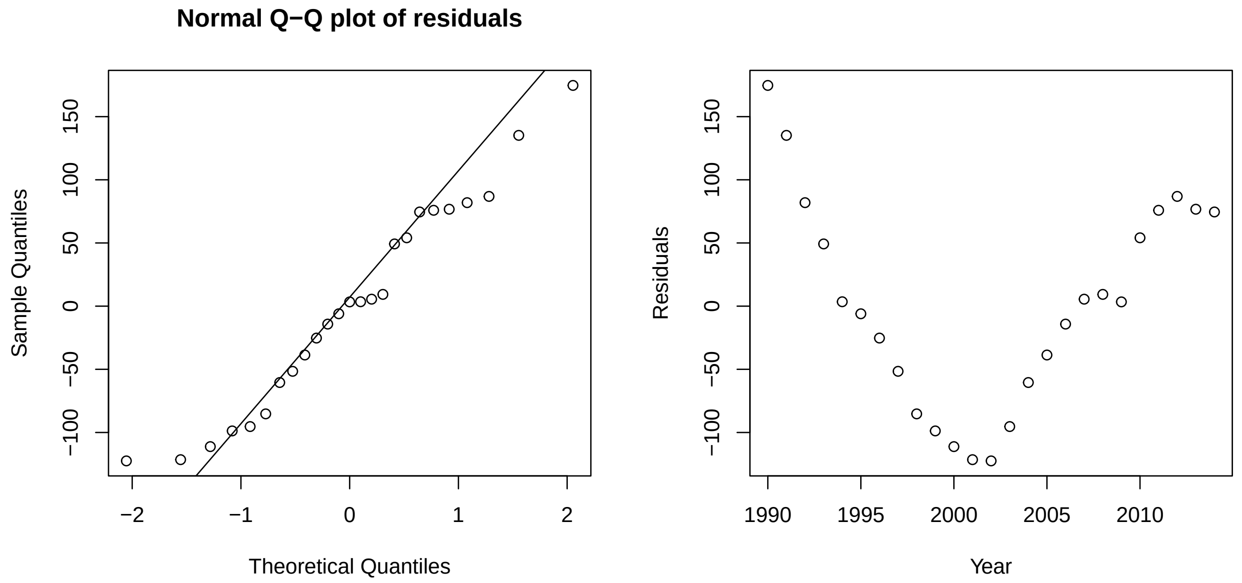
3 ☐ 0.98

4 ☐ 1.93

5 ☐ This cannot be calculated with the information given in the exercise.

Question VIII.3 (20)

The following two plots are generated for the residual analysis of the estimated model between the energy demand and the year:



Which of the following conclusions is most appropriate based on the two plots above (both the conclusion and the argument must be correct)?

1 ☐ The assumption of independent errors should be rejected, since the distribution of the residuals appears to be heavily right skewed.

2 ☐ The assumption of independent errors should be rejected, since the distribution of the residuals appears to be heavily left skewed.

3 ☐ The assumption of independent errors should be rejected, since a clear linear relation can be seen between the residuals and the years.

4 ☐ The assumption of independent errors should be rejected, since a clear non-linear relation can be seen between the residuals and the years.

5 ☐ None of the above conclusions with their associated argument are correct.

Question VIII.4 (21)

Are there, according to the book's definition, any extreme observations in the sample consisting of the residuals from the estimated model between the energy demand and the year (both conclusion of argument must be correct)?

1 ☐ Yes, since $-262.9 < 122.5$ and $174.7 < 277.0$.

2 ☐ No, since $-262.9 < 122.5$ and $174.7 < 277.0$.

3 ☐ Yes, since $135.0 < 297.2$.

4 ☐ No, since $135.0 < 297.2$.

5 ☐ Yes, since $0.5 < 0.789$.

Question VIII.5 (22)

The model is now extended to a multiple linear regression model, using both the year and the GDP as explanatory variables.

The following result is obtained by estimating the model:

```
summary(lm(energy ~ year + gdp))

##
## Call:
## lm(formula = energy ~ year + gdp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52.87 -29.05  -9.28  27.18  69.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.24e+02   4.98e+03    0.13     0.90
## year        8.93e-02   2.50e+00    0.04     0.97
## gdp         1.19e-01   1.26e-02    9.52    3e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38 on 22 degrees of freedom
## Multiple R-squared:  0.959,Adjusted R-squared:  0.955
## F-statistic:  256 on 2 and 22 DF,  p-value: 5.75e-16
```

When comparing the result from the model with only the year as explanatory variable (from the start of the exercise) and the result of the model with both the year and GDP, the following "absurd" conclusion can be drawn for the hypothesis of a dependence between year and energy demand:

There is very strong evidence of the hypothesis when the year alone is used as explanatory variable, while there is little or no evidence when both year and GDP are used.

However, this result is by no means absurd statistically, as it often can occur if the following is true:

1 ☐ GDP is decreasing in the period.

2 ☐ There is a relatively high non-linear relationship between the year and the energy demand in the observed data.

3 ☐ There is a relatively high non-linear relationship between the year and the GDP demand in the observed data.

4 ☐ There is a relatively high correlation between the year and the energy demand in the observed data.

5 ☐ There is a relatively high correlation between the year and the GDP demand in the observed data.

Continue on page 20

Exercise IX

In a study of two types of pig feeds, 20 pigs was divided into two (smaller) groups (x: Group 1 with 8 and Group 2 with 12 pigs). Those two groups received from the age of 3 months until they were slaughtered (6 months) each a different type of feed. The table below shows the pigs weight when slaughtered (kg):

x	113.3	117.9	111.9	109.6	109.6	111.5	97.8	103.3				
y	110.7	108.3	110.6	106.7	109.7	107.5	105.9	111.0	99.9	110.2	99.4	103.6

The following was calculated $\bar{x} = 109.4$, $\bar{y} = 107.0$, $s_x^2 = 6.2^2$ and $s_y^2 = 4.1^2$. It can be assumed that the weight when they were slaugtered followed a normal distribution in each group. Further, the pooled variance was calculated to $s_p^2 = 5.0^2$.

Question IX.1 (23)

What is the 95% confidence interval for the mean weight of the pigs from Group 1 when slaughtered?

1 ☐ $[104.2, 114.6]$

2 ☐ $[105.2, 113.6]$

3 ☐ $[107.6, 111.2]$

4 ☐ $[101.7, 117.1]$

5 ☐ $[106.6, 112.2]$

Question IX.2 (24)

A 99% confidence interval for the variance of the weight in Group 1 is wanted. How is this calculated correctly?

1 ☐ $\left[\frac{7 \cdot 6.2^2}{20.3}, \frac{7 \cdot 6.2^2}{1.0}\right]$

2 ☐ $\left[\frac{8 \cdot 6.2}{20.3}, \frac{8 \cdot 6.2}{1.0}\right]$

3 ☐ $\left[\frac{9 \cdot 6.2}{20.3}, \frac{9 \cdot 6.2}{1.0}\right]$

4 ☐ $\left[\frac{8 \cdot 6.2^2}{20.3}, \frac{8 \cdot 6.2^2}{1.0}\right]$

5 ☐ $\left[\frac{7 \cdot 6.2}{20.3}, \frac{7 \cdot 6.2}{1.0}\right]$

Question IX.3 (25)

When testing for the difference in mean slaughter weight between Group 1 and Group 2, what is the result of the usual Welch test statistics?

1 ☐ $|t_{\text{obs}}| = 0.96$

2 ☐ $|t_{\text{obs}}| = 1.0$

3 ☐ $|t_{\text{obs}}| = 2.6$

4 ☐ $|t_{\text{obs}}| = 49.8$

5 ☐ $|t_{\text{obs}}| = 90.8$

Question IX.4 (26)

If, in a new experiment, it is wanted to obtain a strength of 80% to be able to detect one difference of 4 kg between the two groups of on a confidence level of 99%, and the weighted variance is used as a guess of the population's variance, how many pigs should at least be included in this experiment?

1 ☐ 22

2 ☐ 42

3 ☐ 52

4 ☐ 78

5 ☐ 104

Continue on page 22

Exercise X

The following sample has been collected and sorted:

1	2	3	4	5	6	7	8	9	10	11	12
0.4	1.0	1.4	2.0	2.5	2.7	3.0	3.3	4.2	4.5	6.5	7.6

Question X.1 (27)

What is the median of the sample?

1 ☐ 2.6

2 ☐ 2.7

3 ☐ 2.85

4 ☐ 3.0

5 ☐ 3.3

Question X.2 (28)

If the sample is stored in the vektor `x` in R, which of the following calls is calculating the standard deviation of the sample?

1 ☐ `sqrt(sum((x-mean(x))^2)/(length(x)-1))`

2 ☐ `sd(x)*length(x)`

3 ☐ `sqrt(sd(x))`

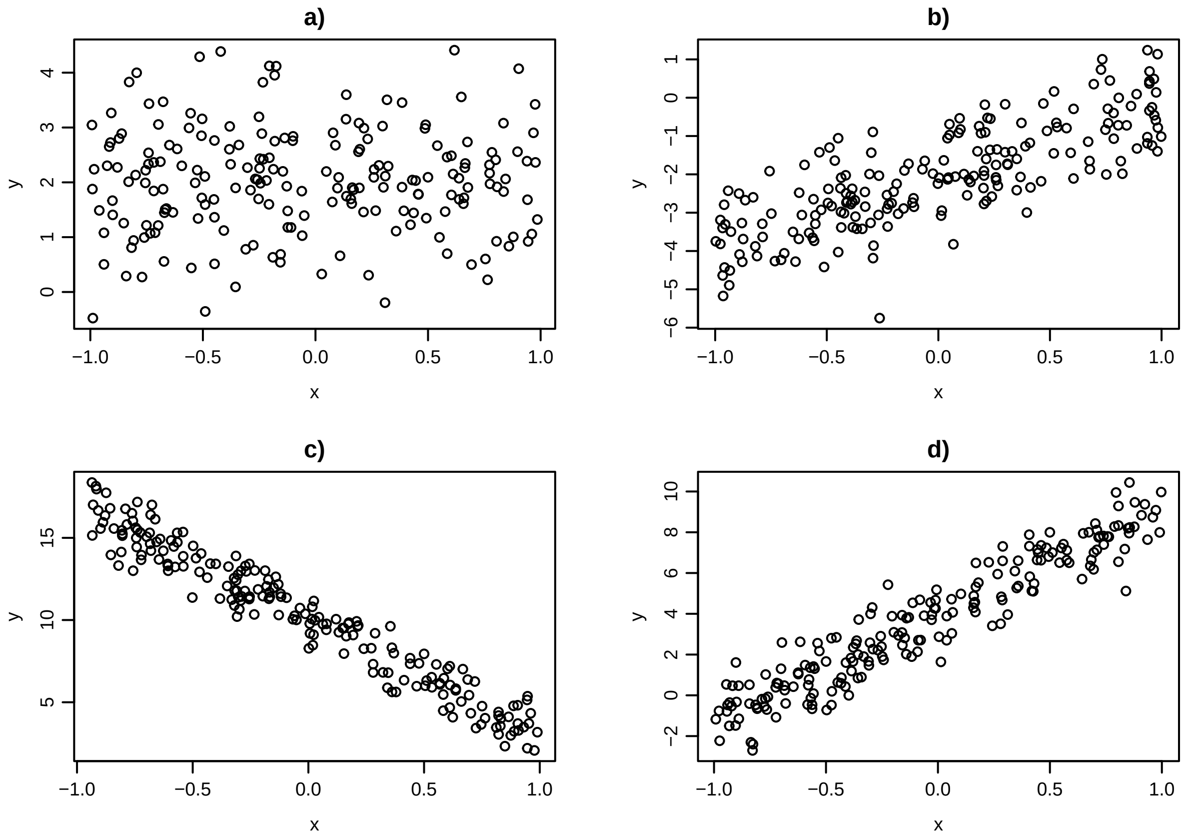
4 ☐ `var(x)^2/length(x)`

5 ☐ `sd(x)^2/length(x)`

Continue on page 23

Exercise XI

Below are four scatter plots of y and x observations:



Question XI.1 (29)

Which four correlation coefficients (in the order: `a), b), c), d)`) fits best with the observations in the figure?

1 ☐ 0.02, 0.79, 0.95, -0.97

2 ☐ 0.02, 0.95, 0.79, -0.97

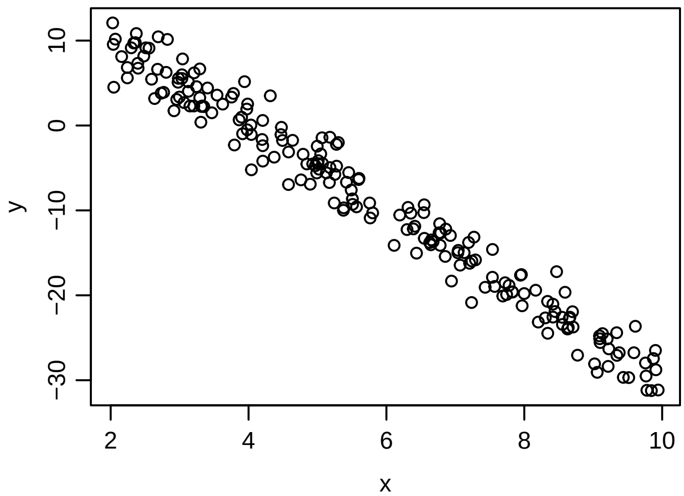
3 ☐ -0.97, 0.02, 0.79, 0.95

4 ☐ 0.02, 0.95, -0.97, 0.79

5 ☐ 0.02, 0.79, -0.97, 0.95

Question XI.2 (30)

Another sample of x and y data is plotted below:



A linear regression is carried out on the values in the plot with the R code

```
summary(lm(y ~ x))
```

and the result for the coefficients estimates from the summary is:

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      A       0.0716   138.3   <2e-16 ***
## x                B       0.1236   -54.9   <2e-16 ***
```

The estimated values of the coefficients have been replaced with letters.

Which of the following answers is the only which is not with very unlikely?

1 ☐ A is 10 and B is -2.

2 ☐ A is 20 and B is -5.

3 ☐ A is 4 and B is -5.

4 ☐ A is 4 and B is -2.

5 ☐ A is 10 and B is -8.

The exam is finished. Enjoy the final weeks of the summer!