Written examination: 26. May 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	III.1	III.2	III.3	IV.1	IV.2	V.1
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer										

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

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

The exam paper contains 21 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

In a cola tasting experiment there are 4 glasses with cola. Each glass contains either regular cola or light cola. You know that there are two glasses of each. A taster randomly chooses two glasses.

Question I.1 (1)

What is the probability that she gets regular cola in one of the glasses and light cola in the other?

1 ☐ 1/4

2 ☐ 1/3

3 ☐ 1/2

4 ☐ 2/3

5 ☐ 3/4

Question I.2 (2)

In another experiment, a glass of regular cola and a glass of light cola are given to each of 25 tasters. They are told to taste and answer if they think that there is a difference between the cola in the glasses. The answers are independent of each other.

From experience, one knows that it can be assumed that there is $p = 0.8$ probability that a taster can taste the difference between regular and light. Let X denote the number of the 25 tasters who say there is a difference. What will be the variance of X?

1 ☐ $\text{V}(X) = 5$

2 ☐ $\text{V}(X) = 4$

3 ☐ $\text{V}(X) = 3$

4 ☐ $\text{V}(X) = 2$

5 ☐ $\text{V}(X) = 1$

Continue on page 3

Exercise II

10 women measured their morning temperature on both July 1st and December 1st. From the measurements, one would like to investigate whether there is a difference in the morning temperature for women in the summer compared to the winter. It can be assumed that the summer measurements are normally distributed and that the winter measurements are normally distributed.

Question II.1 (3)

Which analysis will be most appropriate?

1 ☐ Test for the difference between two proportions

2 ☐ Regression analysis

3 ☐ (Un-paired) t-test

4 ☐ Paired t-test

5 ☐ Test in the binomial distribution

Question II.2 (4)

When the test was carried out a p-value of 0.4 was obtained. This means that:

1 ☐ There is a 40% probability that there is a difference between the morning temperature in the summer compared to the winter.

2 ☐ There is a 0.4% probability that there is a difference between the morning temperature in the summer compared to the winter.

3 ☐ The hypothesis cannot be tested.

4 ☐ There is definitely a difference between the morning temperature in the summer compared to the winter.

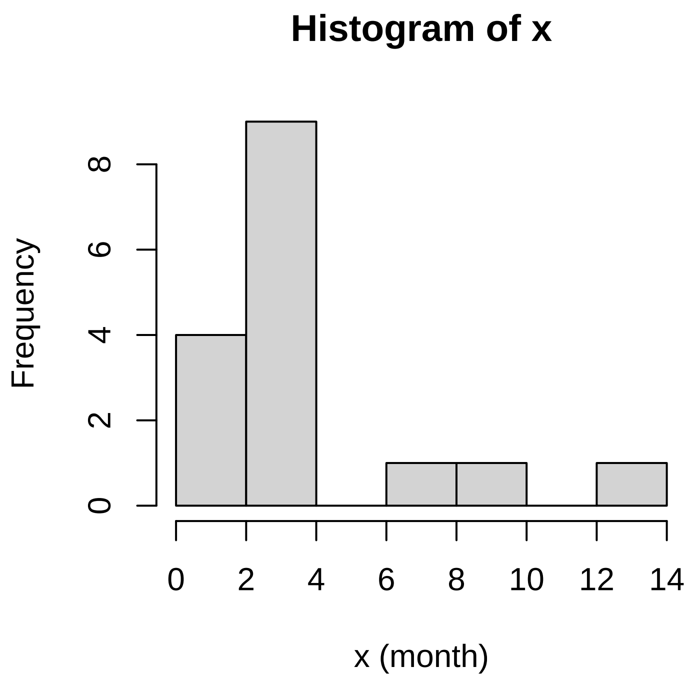
5 ☐ Under the null hypothesis, the probability of obtaining a value of the test statistic which is less extreme, than the value obtained, is 0.6.

Continue on page 4

Exercise III

A company has purchased a new 3D printer technology and they want to investigate whether it can be used to make components that are durable enough to be included in a specific product.

An experiment has been carried out where components, printed with the new technology, have been used in a batch of test products. These products have then been subjected to a test that determines their lifetime. It is assumed that the lifetime follows an exponential distribution, so let $X \sim Exp(\lambda)$ denote the lifetime in months. A sample has been collected for $n = 16$ products. A histogram of the sample is:



The observed life times has been saved in the vector `x` and the following R code is run:

```
## Number of simulations
k <- 10000
nx <- length(x)
## Simulate k times
simxsamples <- replicate(k, rexp(nx, 1/mean(x)))
## Calculate the sample mean
simmeans <- apply(simxsamples, 2, mean)
## Quantiles of the means
quantile(simmeans, c(0.005,0.995))

##  0.5% 99.5%
##  1.70  6.42

quantile(simmeans, c(0.025,0.975))
```

```
##  2.5% 97.5%
##  2.07  5.68

quantile(simmeans, c(0.05,0.95))

##   5%  95%
## 2.26 5.26
```

Question III.1 (5)

It was pre-planned to investigate whether it can be shown, at significance level $\alpha = 1\%$, that the average lifetime m_X is over 2 months for the components.

Can this be concluded on the basis of the collected sample and the calculations above (both conclusion and argument must be correct)?

1 ☐ Since 2 is contained in the calculated 99% confidence interval it cannot be concluded.

2 ☐ Since 2 is not contained in the calculated 99% confidence interval it can be concluded.

3 ☐ Since 2 is contained in the calculated 95% confidence interval it cannot be concluded.

4 ☐ Since 2 is not contained in the calculated 95% confidence interval it can be concluded.

5 ☐ With the given information it is not possible to answer this question.

Question III.2 (6)

What is the sample mean of the collected sample?

1 ☐ $\bar{x} = 3.40$

2 ☐ $\bar{x} = 3.76$

3 ☐ $\bar{x} = 3.875$

4 ☐ $\bar{x} = 4.06$

5 ☐ With the given information it is not possible to answer this question.

Question III.3 (7)

A new sample of lifetimes has been collected where a new material has been used to print the components. They are subsequently subjected to the same tests and the observed lifetimes are stored in the vector y. There are $n_Y = 17$ observations in the new sample.

The following R code is run afterwards:

```r
## Number of simulations
k <- 10000
nx <- length(x)
ny <- length(y)
## Simulate k times
simxsamples <- replicate(k, rexp(nx, 1/mean(x)))
simysamples <- replicate(k, rexp(ny, 1/mean(y)))
## Calculate the simulated statistics
simdifmeans <- apply(simysamples, 2, mean) - apply(simxsamples, 2, mean)
simdifmedians <- apply(simysamples, 2, median) - apply(simxsamples, 2, median)
## Quantiles of the simulated statistics
quantile(simdifmeans, c(0.025,0.975))

##  2.5% 97.5%
## 0.733 9.443

quantile(simdifmeans, c(0.05,0.95))

##   5%  95%
## 1.30 8.59

quantile(simdifmedians, c(0.025,0.975))

##   2.5%  97.5%
## -0.428  8.265

quantile(simdifmedians, c(0.05,0.95))

##     5%    95%
## 0.0837 7.3868
```

Which of the following conclusions can be drawn on the basis of these calculations?

1 ☐ At $\alpha = 5\%$ significance level it can be concluded that the 50% quantile of the product lifetime is higher with components of the new material.

2 ☐ At $\alpha = 10\%$ significance level it can be concluded that the 50% quantile of the product lifetime is higher with components of the new material.

3 ☐ At $\alpha = 5\%$ significance level it can be concluded that there is at least 50% probability that the product lifetime is higher with components of the new material.

4 ☐ At $\alpha = 10\%$ significance level it can be concluded that there is at least 50% probability that the product lifetime is higher with components of the new material.

5 ☐ With the given information no conclusions can be drawn.

Continue on page 7

Exercise IV

Assume that X is normally distributed with mean 10 and variance 4, Y is normally distributed with mean 20 and variance 25, and X and Y are independent.

Question IV.1 (8)

Then $2Y - 2X + 4$ has the variance:

1 ☐ 36

2 ☐ 58

3 ☐ 84

4 ☐ 116

5 ☐ None of the values above.

Question IV.2 (9)

What is the standard deviation of $f(X,Y) = 2Y^2 + X^3/3$ (tip: if you solve this using simulation, remember to have many repetitions and choose the answer with the result being approx. $\pm$ 10 from the stated number in the answer)?

1 ☐ $\sigma_{f(X,Y)} \approx 100$

2 ☐ $\sigma_{f(X,Y)} \approx 250$

3 ☐ $\sigma_{f(X,Y)} \approx 350$

4 ☐ $\sigma_{f(X,Y)} \approx 450$

5 ☐ $\sigma_{f(X,Y)} \approx 5 \cdot 10^4$

Continue on page 8

Exercise V

The association between pressure (p) and depth (h) in an open liquid container may be described theoretically by the equation

$$p = p_0 + \rho g h \,,$$

where p_0 is atmospheric pressure, ρ is the density of the liquid, and g is the acceleration due to gravity. An experiment was conducted with the purpose of determining the density of a special liquid. 10 measurements of depth (in m) and pressure (in Pa) were conducted in this liquid, and the results were assigned to two vectors in R, `depth` and `pressure`, respectively. Furthermore, the following R code was run:

```
model1 <- lm(pressure ~ depth)
summary(model1)

##
## Call:
## lm(formula = pressure ~ depth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -119166  -73422   30513   53635  124689
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 1.019e+08  5.867e+04 1737.529  < 2e-16 ***
## depth       5.031e+03  9.455e+02    5.321 0.000711 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85880 on 8 degrees of freedom
## Multiple R-squared:  0.7797,Adjusted R-squared:  0.7521
## F-statistic: 28.31 on 1 and 8 DF,  p-value: 0.0007105
```

Question V.1 (10)

Give the estimate of the atmospheric pressure during the experiment:

1 ☐ $5.031 \cdot 10^3$ Pa

2 ☐ $5.867 \cdot 10^4$ Pa

3 ☐ $9.455 \cdot 10^7$ Pa

4 ☐ $1.019 \cdot 10^8$ Pa

5 ☐ $1.025 \cdot 10^8$ Pa

Question V.2 (11)

One would like to test the hypothesis that the expected atmospheric pressure is $1.005 \cdot 10^8$ Pa under the experimental conditions. Give the usual test statistic used to test this hypothesis:

1 ☐ $t_{\text{obs}} = 1738$

2 ☐ $t_{\text{obs}} = 5.321$

3 ☐ $t_{\text{obs}} = 23.86$

4 ☐ $t_{\text{obs}} = 28.31$

5 ☐ $t_{\text{obs}} = 0.000711$

Question V.3 (12)

Give a 95% confidence interval for the parameter which describes the association between depth and pressure:

1 ☐ $1.019 \cdot 10^8 \pm 2.306 \cdot 85880/(10-2)$

2 ☐ $1.019 \cdot 10^8 \pm 2.306 \cdot 85880$

3 ☐ $5031 \pm 2.306 \cdot 85880$

4 ☐ $1.019 \cdot 10^8 \pm 2.306 \cdot 5.867 \cdot 10^4$

5 ☐ $5031 \pm 2.306 \cdot 945.5$

Question V.4 (13)

Give an estimate of the density of the liquid during the experiment, when the acceleration due to gravity, g, is 9.82 N/kg:

1 ☐ 512 kg/m^3

2 ☐ 1004 kg/m^3

3 ☐ 307 kg/m^3

4 ☐ 802 kg/m^3

5 ☐ 610 kg/m^3

Continue on page 11

Exercise VI

A sample was taken with independent observations from a normally distributed population. One would like to test the hypothesis that the mean is zero against the alternative, that it is different from zero. The test statistic for the test follows a t-distribution. A p-value of 0.001 was obtained.

Question VI.1 (14)

What is then known about the 99% confidence interval for the mean?

1 ☐ It contains zero.

2 ☐ It does not contain zero.

3 ☐ It contains zero, but not the estimate of the mean.

4 ☐ There is not enough information to know anything specific about the confidence interval.

5 ☐ It contains 0.01.

Question VI.2 (15)

If there were $n = 20$ observations in the sample, what do we then know about the observed test statistic?

1 ☐ $t_{\text{obs}} = -1.33$ or $t_{\text{obs}} = 1.33$

2 ☐ $t_{\text{obs}} = -1.73$ or $t_{\text{obs}} = 1.73$

3 ☐ $t_{\text{obs}} = -3.55$ or $t_{\text{obs}} = 3.55$

4 ☐ $t_{\text{obs}} = -3.58$ or $t_{\text{obs}} = 3.58$

5 ☐ $t_{\text{obs}} = -3.88$ or $t_{\text{obs}} = 3.88$

Continue on page 12

Exercise VII

The Danish Veterinary and Food Administration wants to reduce the proportion of resistant bacteria in pigs intestinal flora, as they pose a human risk. qPCR is one microbiological method to count the number of specific genes in a faeces sample. Below is the count of three genes: 16S, which is a reference gene, and two genes that encode resistance to tetracycline (tetO and tetM). Four samples were taken at different times (first Sample 1, then 2, 3 and finally 4) on the same farm and the researchers want to investigate whether changes have occurred.

	16S	tetO	tetM	Sum
Sample 1	4675	171	76	4922
Sample 2	2222	95	1	2318
Sample 3	2750	49	2	2801
Sample 4	2040	47	1	2088
Sum	11687	362	80	12129

A χ^2-test should be carried out to determine if the proportion of resistant genes has changed over time.

Question VII.1 (16)

The degrees of freedom in this test is:

1 ☐ 8

2 ☐ 12

3 ☐ 6

4 ☐ 9

5 ☐ It doesn't make sense to do a χ^2-test, when two of the observations are 1.

Question VII.2 (17)

Under the null hypothesis what is the expected number of tetM copies in Sample 4?

1 ☐ 20

2 ☐ 1

3 ☐ 13.77

4 ☐ 26.10

5 ☐ 696

Question VII.3 (18)

The test statistic turns out to be 132.3. The relevant p-value is found using which of the following calls in R?

1 ☐ `1 - dchisq(132.3, df=6)`

2 ☐ `1 - pchisq(132.3, df=6)`

3 ☐ `qchisq(132.3, df=6)`

4 ☐ `pchisq(132.3, df=6)`

5 ☐ `qchisq(1/132.3, df=6)`

Question VII.4 (19)

It has previously been planned to investigate whether the occurrence of tetO has changed between the first sample and fourth sample. For reasons not explained here, the observations of tetM should not be considered in this test. The following code has been run with the associated code output:

```
prop.test(x=c(171, 47), n=c(4675+171, 2040+47), correct=FALSE, conf.level=0.95)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(171, 47) out of c(4675 + 171, 2040 + 47)
## X-squared = 7.8067, df = 1, p-value = 0.005205
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.004550394 0.020982546
## sample estimates:
##     prop 1     prop 2
## 0.03528683 0.02252036
```

The usual $\alpha = 0.05$ significance level is used. What is the conclusion (both the conclusion and the argumentation must be correct)?

1 ☐ No significant change has been detected, since $0.02098 < 0.02252$.

2 ☐ A significant change has been detected, since $0.0052 < 0.95$, but it is not possible to conclude if the occurrence has increased or decreased.

3 ☐ A significant change has been detected, since $0.0052 < 0.05$, and the occurrence of tetO has increased.

4 ☐ A significant change has been detected, since $0.0052 < 0.95$, and the occurrence of tetO has increased.

5 ☐ A significant change has been detected, since $0.0052 < 0.05$, and the occurrence of tetO has decreased.

Continue on page 15

Exercise VIII

The IQ of a randomly selected individual is modeled by a normally distributed random variable. 50% of the population have an IQ over 100 (and 50% have an IQ below 100). Suppose 68% of the population have an IQ in the range of 85-115.

Question VIII.1 (20)

What percentage of the population have an IQ of at least 140 and is thus considered geniuses according to this model?

1 ☐ 0.01%

2 ☐ 1%

3 ☐ 4%

4 ☐ 0.4%

5 ☐ 0.06%

Continue on page 16

Exercise IX

The data below have been collected from two groups:

Group 1: 10.5, 9.3, 10.7, 10.8, 11.2

Group 2: 8.9, 9.5, 10.2, 9.8, 10.3

All measurements are assumed to be taken independent. The Group 1 measurements are believed to originate from a normal distribution, and the measurements in Group 2 are assumed to originate from a normal distribution. In addition, it is assumed that the variances in the two normal distributions are identical.

Question IX.1 (21)

What is the sample mean of the Group 2 sample?

1 ☐ 9.74

2 ☐ 9.8

3 ☐ 10.2

4 ☐ 10.31

5 ☐ 48.5

Question IX.2 (22)

What will be the numerical value of the test statistic for the usual test of the hypothesis that there is no difference in mean of the two groups?

1 ☐ 0.8

2 ☐ 1.04

3 ☐ 1.86

4 ☐ 2.19

5 ☐ 2.55

Question IX.3 (23)

What is the 90% confidence interval for the mean in Group 1?

1 ☐ [9.61, 11.39]

2 ☐ [9.32, 11.68]

3 ☐ [8.92, 12.03]

4 ☐ [9.87, 12.03]

5 ☐ None of the intervals above are correct.

Question IX.4 (24)

A new experiment must be designed in order to achieve a greater power of the statistical test for the mean values. There is still an equal number of observations in each group. The researchers want to have 99% power to discover a difference in mean of at least 1 between the two groups, at significance level 1%. As a guess of the variance, the pooled variance estimate from the two samples are used.

What is the minimum number of observations needed from each group in order for the above requirements to be fulfilled?

1 ☐ At least 4

2 ☐ At least 6

3 ☐ At least 12

4 ☐ At least 18

5 ☐ At least 22

Continue on page 18

Exercise X

How much clothes a person wears (the clothing level) has a large influence on the level of comfort in offices. In the table below samples from three rooms of the average clothing level (on a scale 0 to 1) are presented:

	Room 1	Room 2	Room 3
	0.43	0.56	0.38
	0.36	0.71	0.39
	0.41	0.20	0.48
	0.42	0.57	0.52
	0.41	0.69	0.23
	0.54	0.55	0.37
	0.61	0.78	0.60
	0.53	0.42	0.46
	0.49	0.42	0.44
	0.69	0.59	0.44
Means	0.49	0.55	0.43

As an initial analysis, a one-way analysis of variance, with room as explanatory factor, is carried out. The result is shown in the R output below (where significant codes have been removed and some numbers are replaced by letters):

```
anova(lm(clo ~ room, data=Data))

## Analysis of Variance Table
##
## Response: clo
##           Df  Sum Sq  Mean Sq F value Pr(>F)
## room       2 0.06963 0.034813       A   0.1385
## Residuals 27 0.44147 0.016351
```

Question X.1 (25)

What is the value of A (rounded) in the R output above?

1 ☐ $A = 1.07$

2 ☐ $A = 2.00$

3 ☐ $A = 2.13$

4 ☐ $A = 4.00$

5 ☐ $A = 4.26$

Question X.2 (26)

What is the conclusion (at significance level $\alpha = 0.05$) about the difference in mean clothing level between the three rooms (both the conclusion and the argument must be correct)?

1 ☐ There is a significant difference since $0.016351 < 0.05$.

2 ☐ There is not a significant difference since $0.016351 < 0.05$.

3 ☐ There is a significant difference since $0.1385 > 0.05$.

4 ☐ There is not a significant difference since $0.1385 > 0.05$.

5 ☐ There is a significant difference since $0.034813 < 0.05$.

Question X.3 (27)

What is a pre-planned 95% confidence interval for the difference between the mean value in Room 1 and Room 2 (i.e. it was planned to make this confidence interval only, before the sample was collected)?

1 ☐ $[0.12, 0.45]$

2 ☐ $[0.03, 0.25]$

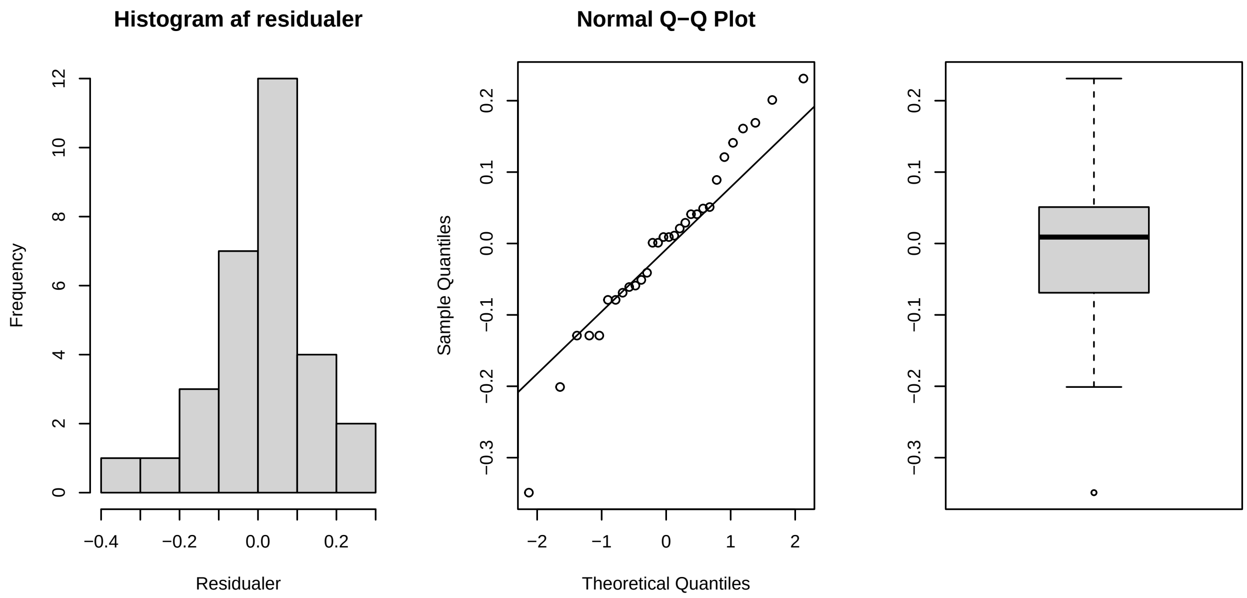
3 ☐ $[-0.17, 0.09]$

4 ☐ $[-0.06, 0.18]$

5 ☐ $[-0.30, 0.42]$

Question X.4 (28)

The following histogram, normal qq-plot and box-plot are of the residuals:



What can rightly be judged based on these plot from the books definition of outliers?

1 ☐ That it is clear that the distribution of residuals is left-skewed.

2 ☐ That the residuals appears normally distributed without any outliers.

3 ☐ That the residuals appears normally distributed, though with a single outlier.

4 ☐ That it is clear that the distribution of residuals is right-skewed.

5 ☐ That the residuals do not follow a normal distribution.

Continue on page 21

Exercise XI

The following sample has been sorted:

$$10, 25, 25, 36, 37, 41, 54, 64, 68, 83$$

Question XI.1 (29)

What is the median of the sample?

1 ☐ 37

2 ☐ 38

3 ☐ 39

4 ☐ 40

5 ☐ 41

Question XI.2 (30)

What is the sample variance?

1 ☐ $\mathrm{V}(x) = 22.60$

2 ☐ $\mathrm{V}(x) = 510.7$

3 ☐ $\mathrm{V}(x) = 1521$

4 ☐ $\mathrm{V}(x) = 1962$

5 ☐ $\mathrm{V}(x) = 2052$

The exam is finished. Have a great summer!