Written examination: 15. August 2018

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

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 14 exercises. To answer the questions, you need to fill in the "multiple choice" form (6 seperate 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	II.1	II.2	III.1	III.2	IV.1	V.1	VI.1	VI.2	VI.3
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer										

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

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

The exam paper contains 25 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

Frank is working in a store, which he is not allowed to close before all customers have gone. The time a customer uses in the store is exponentially distributed with an average of 20 minutes. A customer has just arrived and Frank has a date with his girlfriend in 25 minutes. It takes him 5 minutes to close the store and 5 minutes to go to the restaurant where they will meet.

Question I.1 (1)

If it is assumed that no more customers come to the store, what is the probability that Frank will be late for his date?

1 ☐ 0.29

2 ☐ 0.37

3 ☐ 0.47

4 ☐ 0.61

5 ☐ 1

Continue on page 3

Exercise II

Assume that X and Y are independent and normal distributed random variables, where $X \sim N(3, 2)$ og $Y \sim N(4, 1)$.

Question II.1 (2)

What is the variance of $Z = 2X - 3Y$?

1 ☐ 1

2 ☐ 5

3 ☐ 11

4 ☐ 13

5 ☐ 17

Question II.2 (3)

Assume that the covariance between X and Y is 1.

What is the variance of $Z = 2X - 3Y$?

1 ☐ 1

2 ☐ 5

3 ☐ 11

4 ☐ 13

5 ☐ 17

Continue on page 4

Exercise III

Assume that $X_i \sim N(\mu_1, \sigma_1^2)$ and that $Y_i \sim N(\mu_2, \sigma_2^2)$. Also, assume that all random variables are independent. In addition, let S_1^2 and S_2^2 be the usual variance estimators

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$
$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

where $\bar{X}$ and $\bar{Y}$ as usual is the average of respectively $(X_1, ..., X_{n_1})$ and $(Y_1, ..., Y_{n_2})$. In addition assume that $\mu_1 = \mu_2 = 2$, $n_1 = n_2 = 9$ and $\sigma_1^2 = 2\sigma_2^2 = 2$.

Question III.1 (4)

What is the mean and variance of the random variable $Z = S_1^2/2 + S_2^2/2$?

1 ☐ $E[Z] = 2$, og $V[Z] = \frac{1}{2}$

2 ☐ $E[Z] = 1$, og $V[Z] = \frac{1}{2}$

3 ☐ $E[Z] = 1$, og $V[Z] = \frac{1}{8}$

4 ☐ $E[Z] = 2$, og $V[Z] = \frac{1}{8}$

5 ☐ $E[Z] = \frac{3}{2}$, og $V[Z] = \frac{5}{16}$

Question III.2 (5)

Which of the following statements about the ratio between S_1^2 and S_2^2 is true?

1 ☐ $\frac{1}{2}\frac{S_1^2}{S_2^2} \sim F(8, 8)$

2 ☐ $2\frac{S_1^2}{S_2^2} \sim F(9, 9)$

3 ☐ $\frac{S_1^2}{S_2^2} \sim \chi^2(8)$

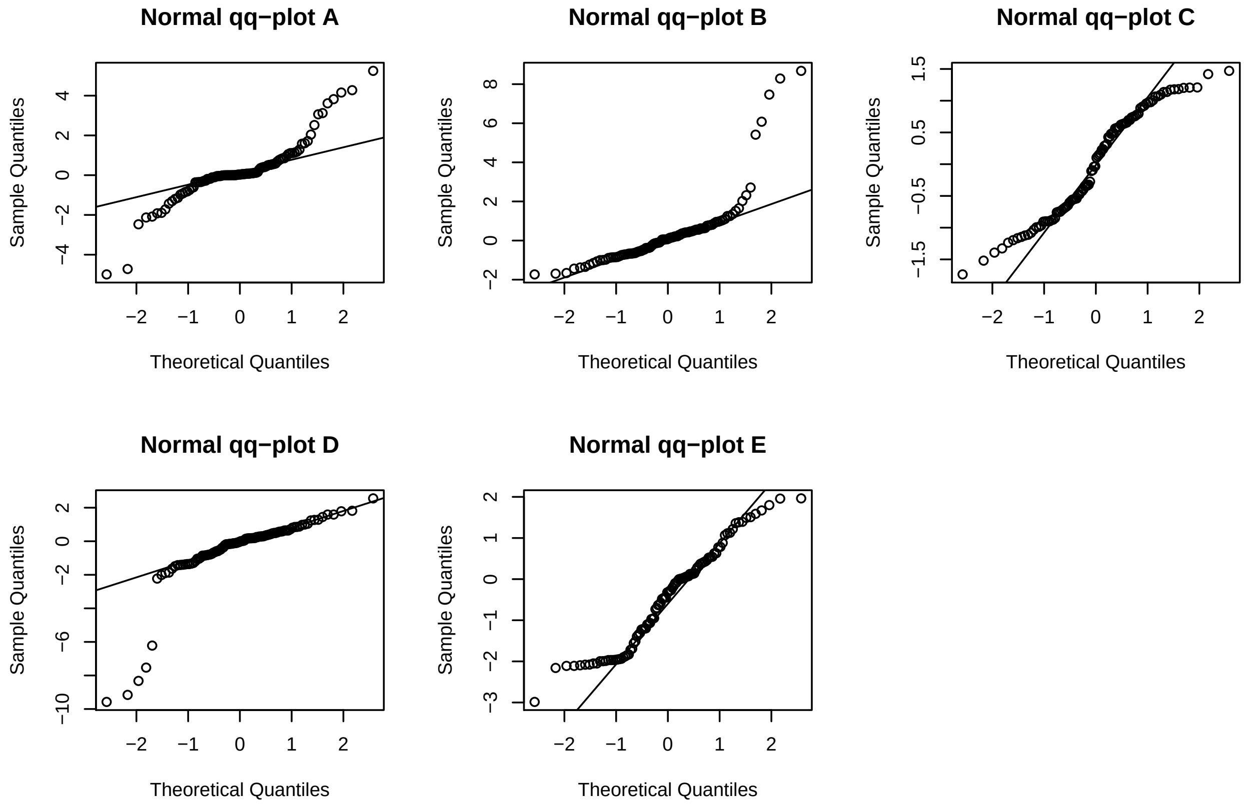
4 ☐ $2\frac{S_1^2}{S_2^2} \sim F(8, 8)$

5 ☐ $\frac{S_1^2}{S_2^2} \sim \chi^2(9)$

Continue on page 5

Exercise IV

Below are shown five qq-plots:



Question IV.1 (6)

Which of these five qq-plots shows that data contains several outliers with values below the sample mean and that the remaining values can be assumed normally distributed?

1 ☐ A

2 ☐ B

3 ☐ C

4 ☐ D

5 ☐ E

Continue on page 6

Exercise V

Assume that $X_1, \ldots, X_{100}$ are independent random variables, which all are normal distributed $N(\mu, \sigma^2)$. Further, assume the 100 random variables represent a sample, and from a realization of the sample $\bar{x} = 2.3$ and $s^2 = 0.25$ have been calculated.

Question V.1 (7)

What is now a 95% confidence interval for $\exp(\mu)$?

1 ☐ [9.03, 11.01]

2 ☐ [7.31, 12.64]

3 ☐ [1.32, 3.28]

4 ☐ [3.74, 26.57]

5 ☐ [9.97, 12.81]

Continue on page 7

Exercise VI

A sample has been randomly taken from a population and the following analysis has been run:

```
t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 3.638, df = 19, p-value = 0.00175
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.141765 4.235249
## sample estimates:
## mean of x
##  2.688507
```

Question VI.1 (8)

Would the null hypothesis

$$H_0 : \mu_X = 0$$

have been rejected on significance level $\alpha = 0.1$ (both conclusion and argument must be correct)?

1 ☐ No, since the p-value for the relevant test is 0.072 and thus greater than $\alpha = 0.1$.

2 ☐ Yes, since the p-value for the relevant test is 0.072 and thus greater than $\alpha = 0.1$.

3 ☐ Yes, since the p-value for the relevant test is 0.035 and thus greater than $\alpha = 0.1$.

4 ☐ No, since the p-value for the relevant test is 0.00175 and thus smaller than $\alpha = 0.1$.

5 ☐ Yes, since the p-value for the relevant test is 0.00175 and thus smaller than $\alpha = 0.1$.

Question VI.2 (9)

Would the null hypothesis

$$H_0 : \mu_X = 2$$

have been rejected on significance level $\alpha = 5\%$ (both conclusion and argument must be correct)?

1 ☐ No, since μ_0 is not contained in the 95% confidence interval.

2 ☐ No, since μ_0 is contained in the 95% confidence interval.

3 ☐ Yes, since μ_0 is not contained in the 95% confidence interval.

4 ☐ Yes, since μ_0 is contained in the 95% confidence interval.

5 ☐ This cannot be decided with the given information.

Question VI.3 (10)

How many observations are in the sample?

1 ☐ $n = 10$

2 ☐ $n = 18$

3 ☐ $n = 19$

4 ☐ $n = 20$

5 ☐ $n = 21$

Continue on page 9

Exercise VII

In a linear regression problem, the following design matrix has been established

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}$$

where we denote the model parameters $\boldsymbol{\beta} = [\beta_0, \beta_1]$, hence the model is given by $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.

Question VII.1 (11)

Which statement about the resulting model is correct (independent of data)?

1 ☐ $\hat{\beta}_0 = \bar{Y}$

2 ☐ $\hat{Y}_3 = \hat{Y}_1 + \hat{Y}_2$

3 ☐ $\hat{\beta}_1 = \bar{Y}$

4 ☐ $\hat{Y}_1 = \hat{Y}_2 + \hat{Y}_3$

5 ☐ $\hat{Y}_2 = \hat{Y}_1$

Question VII.2 (12)

If the residual variance is known and equal to 1 (i.e. $\sigma^2 = 1$), what is then the covariance between the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$?

1 ☐ 2

2 ☐ -0.8

3 ☐ 0.8

4 ☐ 1

5 ☐ -1

Continue on page 10

Exercise VIII

The witch trials in Salem were a series of hearings and litigations in the British colony of Massachusetts in America, between February 1692 and May 1693. During these trials, 185 people, 141 women and 44 men were accused of being witches. 19 of the accused, 14 women and 5 men, were hanged.

	Accused men	Accused women	Total
Hanged	5	14	19
Not hanged	39	127	166
Total	44	141	185

We wish to test the hypothesis

$$H_0 : p_1 = p_2,\ H_1 : p_1 \neq p_2$$

where p_1 is the proportion of accused women who were hanged and p_2 is the proportion of accused men who were hanged.

Question VIII.1 (13)

What will be the usual test statistic when we want to test the hypothesis?

1 ☐ $z_{\text{obs}} = (\frac{5}{44} + \frac{14}{141})/\sqrt{\frac{19}{185} \cdot (1 - \frac{19}{185}) \cdot (\frac{1}{44} - \frac{1}{141})} = 5.61$

2 ☐ $z_{\text{obs}} = (\frac{14}{19} + \frac{127}{166})/\sqrt{\frac{19}{185} \cdot (1 - \frac{19}{185}) \cdot (\frac{1}{19} - \frac{1}{166})} = 22.9$

3 ☐ $z_{\text{obs}} = (\frac{5}{44} - \frac{14}{141})/\sqrt{\frac{19}{185} \cdot (1 - \frac{19}{185}) \cdot (\frac{1}{44} + \frac{1}{141})} = 0.274$

4 ☐ $z_{\text{obs}} = (\frac{127}{166} - \frac{14}{19})/\sqrt{\frac{19}{185} \cdot (1 - \frac{19}{185}) \cdot (\frac{1}{19} + \frac{1}{166})} = 0.384$

5 ☐ $z_{\text{obs}} = (\frac{127}{166} - \frac{14}{19})/\sqrt{\frac{19}{185} \cdot (1 - \frac{19}{185}) \cdot (\frac{1}{19} - \frac{1}{166})} = 0.431$

Question VIII.2 (14)

Under the assumption that the null hypothesis is true, what is then the expected number of accused and hanged women?

1 ☐ $\frac{19 \cdot 166}{185} = 17.0$

2 ☐ $\frac{5 \cdot 141}{44} = 16.0$

3 ☐ $\frac{14 \cdot 141}{141} = 14.0$

4 ☐ $\frac{14 \cdot 141}{185} = 10.7$

5 ☐ $\frac{19 \cdot 141}{185} = 14.5$

Continue on page 12

Exercise IX

A producer of rat poison wants to test which of 4 types of rat poisons rats are most likely to eat. The 4 types of rat poisons have neutral flavor, vanilla-butter flavor, roastbeef flavor and bread flavor.

The producer plans to test the 4 types of rat poisons on a number of rats, giving each of the rats 4 tastings of rat poison, one with each taste. For each rat it is then detected which of the 4 types of rat poisons the rat ate first.

Question IX.1 (15)

How many rats should the producer use if: she assumes that the population's proportion is 0.5, and wants to decide a 95% confidence interval with a mean width of 4% for the proportion of rats, which prefer rat poison with neutral taste?

1 ☐ 1691

2 ☐ 601

3 ☐ 9604

4 ☐ 6764

5 ☐ 2401

Question IX.2 (16)

The producer decides to give tastings to 1000 rats. The number of rats that select each of the 4 types of rat poison is listed in the following table:

Neutral	Vanilla-butter	Roastbeef	Bread	Total
265	224	269	242	1000

Specify the 95% confidence interval for the proportion of rats that prefer rat poison with neutral flavor.

1 ☐ $0.265 \pm 1.96 \cdot \sqrt{\frac{0.265 \cdot (1-0.265)}{265}} = [0.212, 0.318]$

2 ☐ $0.265 \pm 1.6449 \cdot \sqrt{\frac{0.265 \cdot (1-0.265)}{265}} = [0.220, 0.310]$

3 ☐ $0.265 \pm 1.96 \cdot \sqrt{\frac{0.265 \cdot (1-0.265)}{735}} = [0.233, 0.297]$

4 ☐ $0.265 \pm 1.6449 \cdot \sqrt{\frac{0.265 \cdot (1-0.265)}{735}} = [0.238, 0.297]$

5 ☐ $0.265 \pm 1.96 \cdot \sqrt{\frac{0.265 \cdot (1-0.265)}{1000}} = [0.238, 0.292]$

Question IX.3 (17)

The producer changes the taste of vanilla-butter, and then repeats the experiment. This time the outcome of the experiment was:

Neutral	Vanilla-butter	Roastbeef	Bread	Total
219	284	263	234	1000

You must decide whether there is a significant difference in the proportion of rats that prefer rat poison with vanilla-butter taste in the two experiments corresponding to the zero hypothesis

$$H_0 : p_1 = p_2$$

where p_1 is the proportion of rats that prefer rat poison with vanilla-butter flavor in the first experiment and p_2 is the proportion of rats that prefer rat poison with vanilla-butter taste in the second experiment.

Indicate which of the following calls that can be used to calculate this in R:

1 ☐ `prop.test(x=c(224,284), n=c(2000,2000), correct=FALSE)`

2 ☐ `prop.test(x=c(0.224,0.284), n=c(2000,2000), correct=FALSE)`

3 ☐ `prop.test(x=60, n=1000, correct=FALSE)`

4 ☐ `prop.test(x=c(224,284), n=c(1000,1000), correct=FALSE)`

5 ☐ `prop.test(x=60, n=2000, correct=FALSE)`

Continue on page 14

Exercise X

An experiment was carried out using 33 cylinder-shaped containers, which were all filled to the brim with a powder material. Among other things, the (interior) radius and height (in cm) of each container were measured, and the weight of the content (in g) was determined. The measurements of the containers' radiuses are assigned to the vector `radius` in R, while the vector `ratio` contains measurements of the ratio between the weight of each container's content and the container's height (in g/cm).

Afterwards, the following code was executed in R:

```
radius2 <- radius^2
model1 <- lm(ratio ~ radius + radius2)
summary(model1)

##
## Call:
## lm(formula = ratio ~ radius + radius2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.269 -20.380  -2.684  15.071  64.217
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.9362    38.3505  -0.129    0.898
## radius       -0.4221     6.7042  -0.063    0.950
## radius2       2.8765     0.2645  10.876 6.25e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.19 on 30 degrees of freedom
## Multiple R-squared:  0.993,Adjusted R-squared:  0.9926
## F-statistic:  2141 on 2 and 30 DF,  p-value: < 2.2e-16
```

Question X.1 (18)

Which of the following assertions is the only one which describes the statistical model corresponding to `model1` correctly?

1 ☐ The model is a simple linear regression model. It has two dependent variables, weight and height. Radius is the explanatory variable.

2 ☐ The model is a simple linear regression model. The ratio between weight and height is the dependent variable, while radius is the explanatory variable.

3 ☐ The model is a multiple linear regression model. Radius and radius squared are the dependent variables, while the ratio between weight and height is the explanatory variable.

4 ☐ The model is a multiple linear regression model. The ratio between weight and height is the dependent variable. There are two explanatory variables, radius and radius squared.

5 ☐ The model is a multiple linear regression model. It has two dependent variables, weight and height, and two explanatory variables, radius and radius squared.

Question X.2 (19)

Using `model1` as a starting point, give an estimate of the weight of the powder material content of a filled container, which has both a radius and a height of 10 cm.

1 ☐ The ratio between weight and height is estimated to be 271.5 g/cm, which gives an estimated weight of 27.15 g.

2 ☐ The ratio between weight and height is estimated to be 278.5 g/cm, which gives an estimated weight of 2785 g.

3 ☐ The ratio between weight and height is estimated to be 271.5 g/cm, which gives an estimated weight of 2715 g.

4 ☐ The ratio between weight and height is estimated to be 278.5 g/cm, which gives an estimated weight of 27.85 g.

5 ☐ The weight of the powder material is estimated to be 278.5 g.

Question X.3 (20)

Set the significance level to $\alpha = 0.05$. One would like to investigate whether the model given by `model1` may be reduced to a simple linear regression model, in which radius only enters squared. What is the conclusion based on the R output above (both the argument and the conclusion must be correct)?

1 ☐ The relevant p-value is 0.898 and thus greater than the significance level. The model cannot be reduced as desired.

2 ☐ The relevant p-value is 0.950 and thus greater than the significance level. The model cannot be reduced as desired.

3 ☐ The relevant p-value is 0.950 and thus greater than the significance level. The model can be reduced as desired.

4 ☐ The relevant p-value is $6.25 \cdot 10^{-12}$ and thus smaller than the significance level. The model cannot be reduced as desired.

5 ☐ The relevant p-value is $6.25 \cdot 10^{-12}$ and thus smaller than the significance level. The model can be reduced as desired.

Question X.4 (21)

The following code was executed in R (note that some parts of the output have been replaced by `x`):

```
model2 <- lm(ratio ~ radius2)
summary(model2)

##
## Call:
## lm(formula = ratio ~ radius2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.639 -20.469  -2.871  14.762  63.867
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -7.277      9.264       x        x x
## radius2        2.860      0.043       x        x x
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.7 on x degrees of freedom
## Multiple R-squared:  0.993,Adjusted R-squared:  0.9928
## F-statistic:     x on 1 and  x DF,  p-value:        x
```

Use the statistical model given by `model2` as a starting point. This model has the form

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

Here, Y_i is the ratio between weight and height, x_i is radius squared, and ε_i, $i = 1, \ldots, 33$, are independent and identically $N(0, \sigma^2)$-distributed.

Give a 95% confidence interval for the model's slope parameter:

1 ☐ $2.860 \pm 1.696 \cdot 0.043 = [2.79, 2.93]$

2 ☐ $2.860 \pm 2.040 \cdot 29.7 = [-57.73, 63.45]$

3 ☐ $2.860 \pm 1.696 \cdot \sqrt{29.7} = [-6.38, 12.10]$

4 ☐ $2.860 \pm 2.040 \cdot \sqrt{29.7} = [-8.26, 13.98]$

5 ☐ $2.860 \pm 2.040 \cdot 0.043 = [2.77, 2.95]$

Question X.5 (22)

Again, use the model given by `model2` as a starting point and use the R output from the previous question in the following. Put the significance level to $\alpha = 0.05$. Test the (null) hypothesis that the model's intercept can be set to 0. What may be concluded (both the conclusion and argument must be correct)?

1 ☐ The model's intercept cannot be set to 0 as the relevant p-value is 0.78, and thus the hypothesis is rejected.

2 ☐ The model's intercept cannot be set to 0 as the relevant p-value is $2 \cdot 10^{-7}$, and thus the hypothesis is rejected.

3 ☐ The model's intercept can be set to 0 as the relevant p-value is 0.44, and thus the hypothesis is accepted.

4 ☐ The model's intercept can be set to 0 as the relevant p-value is 0.78, and thus the hypothesis is accepted.

5 ☐ The model's intecept cannot be set to 0 as the relevant p-value is 0.44, and thus the hypothesis is rejected.

Continue on page 18

Exercise XI

Question XI.1 (23)

One way to estimate the proportion of antimicrobial resistant bacteria in a sample, e.g. from a pig, is to spread the same volume of the sample on agar plates with and without the antibiotic. The volume is chosen such that the expected count is 20 on a plate without antibiotics.

The proportion of antimicrobial resistant bacteria is estimated as the observed proportion:

$$\text{"proportion"} = \frac{\text{"Count on plate with antibiotics"}}{\text{"Count on plate without antibiotics"}}$$

If the antibitic doesn't inhibit/kill all bacteria then the "proportion" is above 0% and there is evidence of resistance.

It is of interest to investigate the distribution of "proportion" if the true proportion of resistant bacteria is 75%. It can be assumed that the count on a plate is Poisson distributed.

The following code has be run:

```
k <- 10000
noAntibiotica <- rpois(k, lambda = 20)
withAntibiotica1 <- rpois(k, lambda = 15)
withAntibiotica2 <- rpois(k, lambda = noAntibiotica*0.75)

quantile(noAntibiotica, c(0.025, 0.05, 0.95, 0.975))

##  2.5%    5%   95% 97.5%
##    12    13    28    29

quantile(withAntibiotica1, c(0.025, 0.05, 0.95, 0.975))

##  2.5%    5%   95% 97.5%
##     8     9    22    23

quantile(withAntibiotica1 / withAntibiotica2, c(0.025, 0.05, 0.95, 0.975))

##  2.5%    5%   95% 97.5%
## 0.417 0.500 2.286 2.750

quantile(withAntibiotica1 / noAntibiotica, c(0.025, 0.05, 0.95, 0.975))

##  2.5%    5%   95% 97.5%
## 0.353 0.407 1.333 1.500

quantile(withAntibiotica2 / noAntibiotica, c(0.025, 0.05, 0.95, 0.975))

##  2.5%    5%   95% 97.5%
## 0.385 0.435 1.100 1.167
```

The simulated values of the 2.5% and 97.5% quantiles in the distribution of "proportion" are found to be:

1 ☐ $q_{0.025} = \frac{1}{12}, q_{0.975} = 1 - \frac{1}{29}$

2 ☐ $q_{0.025} = 0.417, q_{0.975} = 2.750$

3 ☐ $q_{0.025} = 0.353, q_{0.975} = 1.500$

4 ☐ $q_{0.025} = 0.385, q_{0.975} = 1.167$

5 ☐ $q_{0.025} = 0.435, q_{0.975} = 1.100$

Question XI.2 (24)

A sample from a pig is plated. 16 were counted on the plate without the antibiotic and 22 were counted on the plate with the antibiotic.

Which of the following conclusions is the only meaningful one regarding the proportion of resistant bacteria in the pig?

1 ☐ The proportion is 137.5%.

2 ☐ The most likely proportion is 80%.

3 ☐ The most likely proportion is 100%.

4 ☐ The most likely proportion is 72.7%.

5 ☐ The only possibility is that something went wrong in the laboratory, such that the observations are invalid.

Continue on page 20

Exercise XII

A chain of supermarkets wants to assess the effect of a advertising campaign. In 15 stores they have counted the number of items sold in the week before and after the campaign.

The table below presents the collected data, which is stored the vectors `before` and `after` in R:

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15
before	31	83	136	493	28	505	510	127	138	19	35	37	268	64	224
after	38	79	132	551	33	560	547	124	152	15	41	39	278	66	290

Furthermore, three different functions are applied to each of the two vectors and the result is:

	before	after
mean	179.9	196.3
sd	182.1	202.1
var	33160.6	40861.0

Question XII.1 (25)

Which of the following codes will calculate a non-parametric bootstrap 95% confidence interval for the mean of the relative change in the sales (`k` is set to 10000)?

1 ☐
```
simSamples <- replicate(k, sample((after-before)/before, replace = TRUE))
quantile(apply(simSamples, 2, mean), c(0.025, 0.975))
```

2 ☐
```
simSamples <- replicate(k, sample((after-before)/(before+after), replace = TRUE))
quantile(apply(simSamples, 2, mean), c(0.025, 0.975))
```

3 ☐
```
simSamples <- replicate(k, sample(after-before, replace = TRUE))/
              replicate(k, sample(after, replace = TRUE))
quantile(apply(simSamples, 2, mean), c(0.025, 0.975))
```

4 ☐
```
simSamples <- replicate(k, sample(after-before, replace = TRUE)) -
              replicate(k, sample(after, replace = TRUE))
quantile(apply(simSamples, 2, mean), c(0.025, 0.975))
```

5 ☐
```r
simSamples <- replicate(k, sample(after-before, replace = FALSE))/
              replicate(k, sample(after, replace = TRUE))
quantile(apply(simSamples, 2, mean), c(0.025, 0.975))
```

Question XII.2 (26)

Under the assumption of normality a 95% confidence interval for the weekly sale prior to the campaign is found to be (despite the realism of the assumption):

1 ☐ $\left[\frac{33161}{26.1}, \frac{33161}{5.63}\right]$

2 ☐ $\left[\sqrt{\frac{6499468}{23.7}}, \sqrt{\frac{6499468}{6.57}}\right]$

3 ☐ $\left[\frac{464248}{26.1}, \frac{464248}{5.63}\right]$

4 ☐ $\left[\frac{2549}{26.1}, \frac{2549}{5.63}\right]$

5 ☐ $\left[\sqrt{\frac{2549}{23.7}}, \sqrt{\frac{2549}{6.57}}\right]$

Continue on page 22

Exercise XIII

A paper manufacturer wants to find out if there is a difference in paper quality produced with wood from different suppliers. In the production, a variable Y is measured and it is known that the quality of the paper depends on this value: the higher the value measured, the higher the quality of the paper is. The following values are collected from separate production runs with wood from 3 different suppliers:

Supplier A	Supplier B	Supplier C
9.3	14.0	10.4
9.2	10.5	10.4
8.0	10.5	9.6
6.9	8.3	8.5

Question XIII.1 (27)

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 paper quality produced with wood from the 3 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   2 12.302  6.1508  2.4126 0.1449
## Residuals  9 22.945  2.5494
```

1 ☐ A significant difference in quality is not found, since the p-value is less than the significance level.

2 ☐ A significant difference in quality is found, since the p-value is greater than the significance level.

3 ☐ A significant difference in quality is not found, since the p-value is greater than the significance level.

4 ☐ A significant difference in quality is found, since the p-value is less than the significance level.

5 ☐ None of the above conclusions are correct.

Question XIII.2 (28)

Referring to the analysis in the previous question. What is the proportion of the total variation explained by the effect of the three suppliers?

1 ☐ $\frac{2.5494}{12.302+22.945+6.1508+2.5494} = 0.058\%$

2 ☐ $\frac{6.1508}{12.302+22.945+6.1508+2.5494} = 14.0\%$

3 ☐ $\frac{2.5494}{6.1508+2.5494} = 29.3\%$

4 ☐ $\frac{12.302}{12.302+22.945} = 34.9\%$

5 ☐ $\frac{6.1508}{6.1508+2.5494} = 70.7\%$

Continue on page 24

Exercise XIV

This exercise is a continuation of the previous exercise. The engineers now remember that the runs were made on 4 different plants. This was taken into account in experiment design, such that in each run the wood from suppliers was shifted between the plants. It is therefore possible to take the effect of the plant into account in the analysis. The data has now been set up such that it is divided according to both factors:

	Supplier A	Supplier B	Supplier C
Plant 1	9.3	14.0	10.4
Plant 2	9.2	10.5	10.4
Plant 3	8.0	10.5	9.6
Plant 4	6.9	8.3	8.5

and thereafter the following analysis has been carried out (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 + Plant))

## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## Supplier   2 12.3017  6.1508    A       B
## Plant      3 17.3867  5.7956    C       D
## Residuals  6  5.5583  0.9264
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question XIV.1 (29)

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

1 ☐ There is both a significant effect of supplier and plant, since the relevant p-values are 0.030 and 0.028 respectively.

2 ☐ There is a significant effect of supplier, but not of plant, since the relevant p-values are 0.030 and 0.056 respectively.

3 ☐ There is not a significant effect of neither supplier nor plant, since the relevant p-values are 0.060 and 0.056 respectively.

4 ☐ There is not a significant effect of supplier, but there is a significant effect of plant, since the relevant p-values are 0.060 and 0.028 respectively.

5 ☐ There is not a significant effect of neither supplier nor plant, since the relevant p-values are 0.12 and 0.17 respectively.

Question XIV.2 (30)

What assumptions must be validated before the results of the analysis can be used?

1 ☐ No assumptions must be validated.

2 ☐ The assumption of variance homogeneity of the errors must be validated, but due to CLT the normality assumption doesn't have to be validated.

3 ☐ Validation should be carried out, however it is not possible due to the low number of observations.

4 ☐ The assumption of variance homogeneity of the errors doesn't have to be validated, but the assumption of normal distribution of the errors must be validated.

5 ☐ Both the assumption of variance homogeneity and normal distribution of the errors must be validated.

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