Technical University of Denmark *Page 1 of 41 pages.*

Written examination: June 25 2020

Course name and number: **Introduktion to Matematical 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 9 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	I.3	II.1	II.2	II.3	II.4	II.5	III.1	III.2
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer	4	1	1	1	3	3	2	2	4	1

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

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

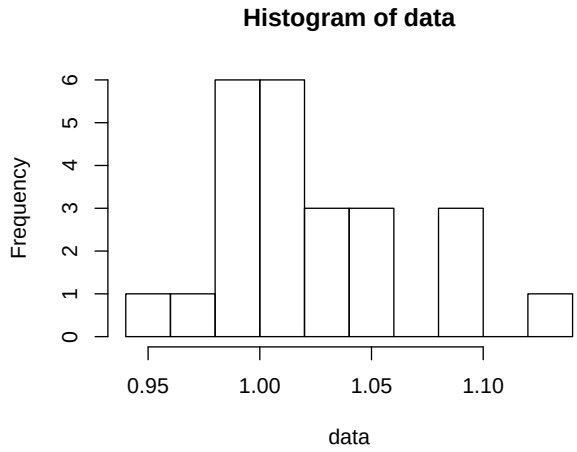
The exam paper contains 41 pages. Continue on page 2

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. 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

Samples of DNA were obtained from 24 different types of tissue. The distribution of nucleotide bases was measured, and the ratio of guanine to cytosine was calculated for every sample. We will refer to this ratio as the GC ratio.

We will assume that the GC ratio follows a normal distribution, a histogram of the data is seen below:



Question I.1 (1)

Which of the following numbers might be the estimated mean $\hat{\mu}$ for this data?

1 ☐ 0.031

2 ☐ 0.901

3 ☐ 0.968

4* ☐ 1.027

5 ☐ 1.139

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

As we do not have the actual values for this data, we have to deduce the sample mean from the histogram. The values 0.031, 0.901 and 1.139 are all outside the range of data. 0.968 is on the border of the histogram, so 1.027 must be the correct value.

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

Continue on page 4

Question I.2 (2)

Which of the following numbers might be the estimated standard deviation $\hat{\sigma}$ for this data?

1* ☐ 0.042

2 ☐ 0.180

3 ☐ 0.956

4 ☐ 1.017

5 ☐ 1.139

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

The standard deviation is the 'average distance' to the mean. 0.956, 1.017 and 1.139 are clearly too large. 0.180 spans the entire range of data, whereas 0.042 is just right.

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

Question I.3 (3)

We would like to study if the GC ratio differs significantly from 1 (the null-hypothesis).

The following formula was applied,

$$[\hat{\mu} - \hat{\sigma}/\sqrt{24} \cdot t_{0.975}, \hat{\mu} + \hat{\sigma}/\sqrt{24} \cdot t_{0.975}] = [1.0087, 1.0444]$$

where $t_{0.975}$ is the 97.5% quantile in the t-distribution with 23 degrees of freedom.

Which of the following statements is correct:

1* ☐ We reject the null hypothesis on a 5% significance level, since 1 is outside the 95% confindence interval.

2 ☐ We reject the null hypothesis on a 2.5% significance level, since 1 is outside the 97.5% confindence interval.

3 ☐ We reject the null hypothesis on a 2.5% significance level, since 0 is outside the 97.5% confindence interval.

4 ☐ We accept the null hypothesis, since $|1 - \hat{\mu}|$ is less than the standard deviation.

5 ☐ We do not have enough information to accept or reject the null hypothesis.

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

We recognize the formula as formula for the 95% cofidence interval, which corresponds to testing on a 5% significance level. Since 1 is not inside the confidence interval, the null hypothesis is rejected.

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

Continue on page 6

Exercise II

In this exercise, a model for predicting the electricity price on the whole sale market for electricity, the so-called Day-ahead market, is studied. Explanatory variables used are forecasts of solar and wind power.

The data used consists of 28 day averages from February 2020 for the electricity market area "DK1", which mainly covers Jutland.

The following 3 variables are used in the model:

- Electricity price (EUR/MWh)
- Solar power forecast (MW)
- Wind power forecast (MW)

The data is read into 3 vectors: `price`, `solar` and `wind`, and the following code is run:

```
  mean(price)

## [1] 17.4

  mean(solar)

## [1] 30.4

  mean(wind)

## [1] 2687

  range(price)

## [1] -1.39 38.33

  range(solar)

## [1]  6.08 64.00

  range(wind)

## [1]  508 4135
```

Continue on page 7

Afterwards the coefficients in a linear regression model are estimated by:

```
  fit <- lm(price ~ solar + wind)
  summary(fit)

##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.730 -2.785  0.861  2.368  7.453
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.566077   2.874305   13.42  6.3e-13 ***
## solar        0.063822   0.050260    1.27     A     B
## wind        -0.008606   0.000722  -11.91     C     D
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.01 on 25 degrees of freedom
## Multiple R-squared:  0.867,Adjusted R-squared:  0.857
## F-statistic: 81.7 on 2 and 25 DF,  p-value: 1.09e-11
```

Note that 4 fields have been replaced with letters.

Question II.1 (4)

What will be the result of one step in the backward selection procedure with significance level $\alpha = 0.05$ based on this model?

1* ☐ The `solar` variable must be taken out of the model

2 ☐ The `wind` variable must be taken out of the model

3 ☐ None of the explanatory variables must be removed

4 ☐ Both the `solar` and `wind` variables must be taken out of the model

5 ☐ Not enough information has been provided to determine this

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

The p-value for test of difference from zero for `solar` is

```r
2 * (1 - pt(1.27, df=25))
```

```
## [1] 0.216
```

and for the `wind` it is

```r
2 * (1 - pt(11.91, df=25))
```

```
## [1] 8.44e-12
```

Hence, the p-value for `solar` is highest, and it is above α and thus it should be removed.

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

Continue on page 9

Question II.2 (5)

With the estimated model, what is the prediction of the mean difference in electricity price between the day in the period with the lowest wind power and the day in the period with the highest wind power, if the solar power is assumed to be constant?

1 ☐ 2.98 EUR/MWh

2 ☐ 11.8 EUR/MWh

3* ☐ 31.2 EUR/MWh

4 ☐ 34.2 EUR/MWh

5 ☐ 38.6 EUR/MWh

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

```
0.008606 * (4135 - 508)

## [1] 31.2
```

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

Question II.3 (6)

Which of the following statements cannot be concluded for the estimated model?

1 ☐ The width of the prediction interval has its minimum at 30.4 MW solar power and 2687 MW wind power

2 ☐ The maximum value of the residuals is 7.453

3* ☐ The model can be applied to conclude if there are significant non-linear dependencies between the variables

4 ☐ The estimate of the variance of the error ϵ_i is $\hat{\sigma}^2 = 16.1$

5 ☐ The model has explained 86.7% of the variance

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

The model cannot be applied to conclude on non-linear dependencies (we would need to use e.g. curve linear regression). Hence answer no 3 is the correct answr to the question.

For the other answer: The width of the predicition inteval have its minimum at $\bar{x}$, hence answer no. 1 can be concluded.

The maximum of the residuals can be seen in the summary from R and 2 anlso hold a correct conclusion.

The estimate of the variance of ϵ_i is $\hat{\sigma}^2 = 4.01^2 = 16.1$ hence 4 is also a correct conclusion.

Since $R^2 = 0.867$ the model have explaind 86.7% of the variance.

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

Continue on page 11

Question II.4 (7)

During the model validation it is examined if a new variable `load`, which indicates the daily electricity consumption for the 28 days in question, should be included in the model. Again, the significance level is set to $\alpha = 5\%$ and a simple linear regression model is estimated with this new variable and the residuals from the previously estimated model:

```
summary(lm(fit$residuals ~ load))

##
## Call:
## lm(formula = fit$residuals ~ load)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.90  -1.35  -0.09   1.75   6.07
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -38.57812    6.61879   -5.83  3.8e-06 ***
## load          0.01465    0.00251    5.84  3.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.59 on 26 degrees of freedom
## Multiple R-squared:  0.568,Adjusted R-squared:  0.551
## F-statistic: 34.2 on 1 and 26 DF,  p-value: 3.68e-06
```

Which of the following conclusions should be drawn based on this result (both argument and conclusion must be correct)?

1 ☐ A model for predicting the electricity price should be made with `load` as the only explanatory variable

2* ☐ Since correlation between load and residuals is significantly different from zero, then `load` should be included in the model selection

3 ☐ Since the coefficient for `load` is positive, then `load` should not be included in the model selection

4 ☐ Since the coefficient for `load` is positive, then `load` should be included in the model selection

5 ☐ Based on the given information none of the above conclusions can be drawn

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

Since the regression parameter between load and the residual is significant so is the correlation between the two, and hence the incuding load will result in a more acurate model (which should then possibly be reduced). I.e. answer 2 is correct.

For the oter answers note that: we cannot conclude thatload should be the only regressor and hence answer 1 is wrong.

The sign of the parameter do not give statistical information and hence 3 and 4 are wrong.

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

Continue on page 13

Regardless of the outcome above, a model including `load` is now estimated. I.e. using the model

$$\texttt{price}_i = \beta_0 + \texttt{load}_i\beta_1 + \texttt{solar}\beta_2 + \texttt{wind}_i\beta_3 + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2) \tag{1}$$

or written in matrix notation

$$\texttt{price} = \boldsymbol{X}\boldsymbol{\beta}; \quad \epsilon_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{2}$$

As an aid for the further calculations, the following R code is given (which can be copied into R), where `XX` is the matrix $\boldsymbol{X}^T\boldsymbol{X}$

```
XX <- c(28, 73743, 851, 75239, 73743, 195282372, 2267307, 197399368,  851,
        2267307, 32733, 2159250  ,  75239, 197399368, 2159250, 235403829)
XX <- matrix(XX, ncol=4)
XX

##       [,1]      [,2]    [,3]      [,4]
## [1,]    28     73743     851     75239
## [2,] 73743 195282372 2267307 197399368
## [3,]   851   2267307   32733   2159250
## [4,] 75239 197399368 2159250 235403829
```

Question II.5 (8)

The variance estimate is calculated as $\hat{\sigma}^2$, what is the standard error for the parameter β_0?

1 ☐ $\hat{\sigma}^2\sqrt{\frac{28}{24}}$

2* ☐ $\sqrt{7.23\hat{\sigma}^2}$

3 ☐ $2.87\hat{\sigma}$

4 ☐ $\frac{\hat{\sigma}}{\sqrt{28}}$

5 ☐ $\sqrt{28\hat{\sigma}^2}$

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

The variance-covariance matrix for the parameters are given by

$$V(\boldsymbol{\beta}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{3}$$

and the standard error for $\hat{\beta}_0$ is given by

$$\sqrt{V(\boldsymbol{\beta})_{1,1}} = \sqrt{\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{1,1}} \tag{4}$$

$(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{1,1}$ can be calculated in R by

```r
solve(XX)[1,1]

## [1] 7.230107
```

and hence $se(\beta_0) = \sqrt{7.23\hat{\sigma}^2}$.

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

Continue on page 15

Exercise III

In developing a new drug, a number of candidate substances were tested in a laboratory experiment. The experiment was designed so that a higher value of the score `y` indicates that the substance works better. The experiment was performed on 7 different drugs (A to G, in R given in the factor `drug`) and with 5 repetitions. The following values of the score were observed:

A	B	C	D	E	F	G
9.0	9.5	9.7	11.3	12.0	13.4	9.5
11.3	11.8	10.4	11.3	13.2	12.4	11.5
10.1	10.1	10.8	11.3	11.9	11.1	9.2
9.9	8.6	8.8	9.5	11.8	11.2	10.3
8.9	8.8	10.2	11.3	12.1	13.0	9.6

An ANOVA analysis was done by the drug developers and it can be assumed that the necessary assumptions for this test are met. The result of the analysis is:

```
anova(lm(y ~ drug))

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## drug       6 36.107  6.0179  6.9115 0.0001397 ***
## Residuals 28 24.380  0.8707
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question III.1 (9)

What will be the conclusion at significance level $\alpha = 5\%$ based on the analysis result (both argument and conclusion must be correct)?

1 ☐ None of the conclusions below are correct

2 ☐ A significant difference in the effect is found, since the p-value is above the significance level

3 ☐ A significant difference in the effect is not found, since the p-value is below the significance level

4* ☐ A significant difference in the effect is found, since the p-value is below the significance level

5 ☐ A significant difference in the effect is not found, since the p-value is above the significance level

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

The p-value is 0.00014 which is below 0.05 and hence there is a significant effect (i.e. difference between some of the drugs). This is answer no. 4. The also exclude the other answers as being correct.

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

Question III.2 (10)

What is the critical value, at significance level $\alpha = 0.05$, for the test carried out in the ANOVA analysis?

1* ☐ 2.445

2 ☐ 0.262

3 ☐ 0.913

4 ☐ 1.691

5 ☐ 1.701

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

Its the one-way ANOVA test described in Theorem 8.6. The critical value of the test is the $1 - \alpha = 0.95$ quantile in the F-distribution with $k - 1$ and $n - k$ degrees of freedom. The value is found by:

```
qf(0.95, 7-1, 7*5-7)

## [1] 2.445259
```

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

Continue on page 17

Exercise IV

In the board game Risk, players attack each other by moving armies (consisting of divisions) across borders on a map, after which a battle is fought between the two armies. The rules are such that the attacking party rolls with 3 dice and the defending party rolls 2 dice. The attacker uses only the two dice with the highest numbers. Usual 6-sided dice are used.

Question IV.1 (11)

A player is about to attack and wants to calculate the probability of getting at least 5 with at least 2 of 3 dice, which of the following R commands calculates this probability?

1 ☐ `dbinom(2, size = 2, prob = 1/3)`

2* ☐ `1 - pbinom(1, size = 3, prob = 1/3)`

3 ☐ `1 - dbinom(2, size = 3, prob = 1/6)`

4 ☐ `pbinom(2, size = 2, prob = 1/3)`

5 ☐ `1 - dbinom(1, size = 2, prob = 1/6)`

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

The probability of rolling at leats 5 (i.e. 5 or 6) is 1/3 with 3 dice we should look the binomial distribution

$$X \sim Binom(3, 1/3) \tag{5}$$

and the probability in question is

$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - F(1) \tag{6}$$

which is calculated by

```
1 - pbinom(1, size = 3, prob = 1/3)

## [1] 0.2592593
```

i.e. answer no. 2

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

Question IV.2 (12)

A player is about to attack and wants to calculate the probability of getting at least 4 with at least 2 of 3 dice, while the defender gets a maximum of 3 with each of his 2 dice, what is the desired probability?

1 ☐ 0.25

2* ☐ 0.125

3 ☐ 0.5

4 ☐ 0.0625

5 ☐ 0.75

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

The twow events are both binomial and independent and hence we can multiply the probabilities. The probability of getting at least 4 with on dice is 1/2 and the probability of getting at least 4 with 2 out of 3 is

$$P(X \geq 2) = 1 - P(X \leq 1) \tag{7}$$

which can be calculated by

```
(1 - pbinom(1, size = 3, prob = 1/2))
```

```
## [1] 0.5
```

and the probability of getting at most 3 with 2 out of 2 is $\left(\frac{1}{2}\right)^2 = \frac{1}{4}$ i.e. the answer is $\frac{1}{2}\frac{1}{4} = \frac{1}{8} = 0.125$ or directly in R

```
(1 - pbinom(1, size = 3, prob = 1/2))*dbinom(0,size=2,prob=1/2)
```

```
## [1] 0.125
```

```
1/2^3
```

```
## [1] 0.125
```

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

Continue on page 19

Exercise V

The table below shows the number of girls and boys born in Denmark in year 2016 by women age 20, 25, 30 and 35 years, respectively:

	20	25	30	35	Total
Girls	281	1490	2333	1430	5534
Boys	282	1539	2536	1518	5875
Total	563	3029	4869	2948	11409

I.e. 281 girls were born by 20-year-old mothers in 2016.

Question V.1 (13)

What is the 95% confidence interval for the probability that a pregnant woman aged 30 years will give birth to a girl?

1 ☐ $[0.48, 0.52]$

2* ☐ $[0.47, 0.49]$

3 ☐ $[0.49, 0.51]$

4 ☐ $[0.46, 0.50]$

5 ☐ $[0.44, 0.48]$

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

Using Method 7.3:

```
n = 2333+2536
p = 2333/n
sigma_p = sqrt(p*(1-p)/n)
l_limit = p-1.96*sigma_p
u_limit = p+1.96*sigma_p

#Alternatively:
binom.test(2333, n, conf.level = 0.95)

##
##  Exact binomial test
##
## data:  2333 and n
```

```
## number of successes = 2333, number of trials = 4869, p-value = 0.003789
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4650352 0.4932975
## sample estimates:
## probability of success
##              0.4791538
```

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

Continue on page 21

Question V.2 (14)

We will now consider all of our data, and we want to investigate if a significant difference between the probability of getting a girl and the probability of getting a boy can be detected. What is the p-value and the conclusion from the corresponding test, when we use a significance level of $\alpha = 0.05$?

1 ☐ p-value $= 0.49$, and the relevant null hypothesis can be rejected.

2 ☐ p-value $= 0.94$, and the relevant null hypothesis cannot be rejected.

3* ☐ p-value $= 0.0014$, and the relevant null hypothesis can be rejected.

4 ☐ p-value $= 0.06$, and the relevant null hypothesis can be rejected.

5 ☐ p-value $= 0.51$, and the relevant null hypothesis cannot be rejected.

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

Testing if the probabilities are different is the same as testing if one of the probabilities is different from 0.5. Therefore we can use Method **??** (or follow Example **??**). We can get the result directly in R:

```
prop.test(x=5534, n=11409, p = 0.5, correct = FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  5534 out of 11409, null probability 0.5
## X-squared = 10.192, df = 1, p-value = 0.00141
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4758916 0.4942298
## sample estimates:
##         p
## 0.4850557
```

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

Question V.3 (15)

We would like to investigate, if the probability of giving birth to a girl can be assumed equal for women in the 4 investigated age groups. Under the corresponding null hypothesis, what is the expected number of girls born by women age 25?

1 ☐ 1514

2* ☐ 1469

3 ☐ 1490

4 ☐ 1383

5 ☐ 1560

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

The expected number can be calculated as follows (see Section 7.4):

```
p = 5534/11409
(e = p*3029)

## [1] 1469.234
```

which is answer no 2.

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

Continue on page 23

Question V.4 (16)

We would still like to investigate, if the probability of giving birth to a girl can be assumed equal for women in the 4 investigated age groups. What is the number of degrees of freedom in the relevant χ^2-test?

1 ☐ 4

2 ☐ 1

3 ☐ 8

4 ☐ 7

5* ☐ 3

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

Using Method 7.20 we get $c - 1 = 4 - 1 = 3$.

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

Question V.5 (17)

The test statistic, when we are investigating, if the probability of giving birth to a girl can be assumed equal for women in the 4 investigated age groups, is estimated to be 1.6943. Which command in R will provide the corresponding p-value? Assume that the correct number of degrees of freedom is stored in the variable `x`.

1 ☐ `1-pchisq(1.6943, df=x, lower.tail = FALSE)`

2 ☐ `1-pchisq(sqrt(1.6943), df=x, lower.tail = FALSE)`

3 ☐ `pchisq(1.6943^2, df=x, lower.tail = FALSE)`

4 ☐ `pchisq(1.6943^2, df=x)`

5* ☐ `1-pchisq(1.6943, df=x)`

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

Following Example 7.21:

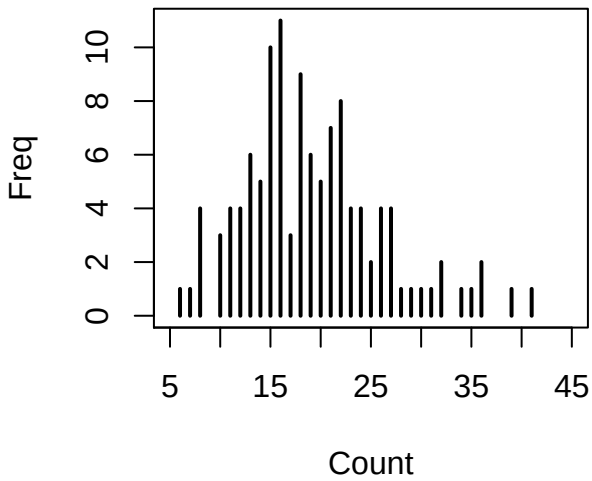
```r
study <- matrix(c(281, 1490, 2333, 1430, 282, 1539, 2536, 1518),
                ncol = 4, byrow = TRUE)
rownames(study) <- c("Girls", "Boys")
colnames(study) <- c("20", "25", "30", "35")

chisq.test(study, correct = FALSE)

##
##  Pearson's Chi-squared test
##
## data:  study
## X-squared = 1.6943, df = 3, p-value = 0.6382
```

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

Continue on page 25

Exercise VI

The figure below shows the annual number of so-called "Major earthquakes" globally during the period 1900-2016.



There are a total of 117 observations (1900-2016), the average number being 19.2/year, with the standard deviation 6.97.

Question VI.1 (18)

Assuming that the number of major earthquakes per year is Poisson distributed with mean 19.2, which of the following statements is then true?

1 ☐ The time between major earthquakes follows an exponential distribution with mean value 19.2

2 ☐ The time between major earthquakes follows a Binomial distribution with mean value 19.2

3* ☐ The time between major earthquakes follows an exponential distribution with mean value 1/19.2

4 ☐ The time between major earthquakes follows a Poisson distribution with mean value 19.2

5 ☐ The time between major earthquakes follows a Poisson distribution with mean value 1/19.2

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

If the number of earthquakes follow a Possion distribution then the time between earthquakes follow an Exponential distribution with the time between events being $1/\lambda$, this is answer no. 3.

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

Continue on page 27

Question VI.2 (19)

With the same assumption as in the previous question, what is the probability of observing more than 40 major earthquakes over a two-year period?

1 ☐ 0.581

2* ☐ 0.358

3 ☐ 0.419

4 ☐ 0.5

5 ☐ 0.642

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

The expected number of earthquakes in 1 year is 19.2 and hence the expected number of earthquakes in 2 years is $2 \cdot 19.2$ and with the distribution still being a Poisson. We are asked for

$$P(X > 40) = 1 - P(X \leq 40) = 1 - F(40) \tag{8}$$

which can be calculated by

```
round(1 - ppois(40, lambda = 2 * 19.2), digits = 3)

## [1] 0.358
```

and this is answer no. 2

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

Question VI.3 (20)

We now want a 95 % confidence interval for the parameter λ (the mean value) based on parametric bootstrap. In order to answer the question the R code below has been executed, where `earthquakes` is a vector with the annual number of earthquakes

```
k <- 1000
n <- length(earthquakes)

sim <- replicate(k, rpois(n, lambda = mean(earthquakes)))
my.means <- apply(sim, 2, mean)
quantile(my.means, prob = c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 18.40171 19.94038

mean(my.means) + sd(my.means) * c(-1, 1) * 1.96 / sqrt(n)

## [1] 19.11887 19.26401

mean(earthquakes) + sd(earthquakes) * c(-1, 1) * 1.96 / sqrt(n)

## [1] 17.91720 20.44178

sim2 <- replicate(k,sample(earthquakes, replace=TRUE))
my.means2 <- apply(sim2, 2, mean)
quantile(my.means2, prob = c(0.025, 0.975))

##     2.5%    97.5%
## 17.94850 20.45363

mean(my.means2) + sd(my.means2) * c(-1, 1) * 1.96

## [1] 17.88962 20.45184
```

Which of the following is a 95% confidence interval for the mean based on parametric bootstrap?

1 ☐ [17.95, 20.45]

2 ☐ [19.12, 19.26]

3 ☐ [17.89, 20.45]

4 ☐ [17.92, 20.44]

5* ☐ [18.40, 19.94]

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

As we are asked for a parametric bootstrap we should use the simulation from the poisson distribution with the same size as the original dataset, this is done in the first lines (`sim`). for each simulation we should calculate the average (`my.means`) and finally the confidence interval is found by taking the 2.5% and 97.5% quantile, these are found as $[18.40, 19.94]$, i.e. answer number 5.

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

Question VI.4 (21)

Regardless of the answer to the previous question, the distribution assumption is considered questionable and therefore it is decided to base the further analysis on non-parametric bootstrap. An estimate (including 95% confidence interval) for the 75% quantile of annual major earthquakes is desired. For this purpose, the following R-code has been executed

```
k <- 1000

sim1 <- replicate(k, sample(earthquakes, replace = TRUE))
Q1 <- function(x){ quantile(x, 0.25)}
samp1 <- apply(sim1, 2, Q1)
quantile(samp1, prob = c(0.025, 0.975))

##  2.5% 97.5%
##    13    16

sim2 <- replicate(k, sample(earthquakes, replace = TRUE))
samp2 <- apply(sim2, 2, quantile, prob = 0.75)
quantile(samp2, prob = c(0.025, 0.975))

##  2.5% 97.5%
##    21    25

sim3 <- replicate(k, rpois(length(earthquakes), lambda = mean(earthquakes)))
Q3 <- function(x){ quantile(x, 0.75)}
samp3 <- apply(sim3, 2, Q3)
quantile(samp3, prob = c(0.025, 0.975))

##  2.5% 97.5%
##    21    23

sim4 <- replicate(k, rpois(length(earthquakes), lambda = mean(earthquakes)))
Q1 <- function(x){ quantile(x, 0.25)}
samp4 <- apply(sim4, 2, Q1)
quantile(samp4, prob = c(0.025, 0.975))

##  2.5% 97.5%
##    15    17

sim5 <- replicate(k, sample(earthquakes))
quantile(sim5, prob = c(0.025, 0.975))

##  2.5% 97.5%
##     8    36
```

What is the 95% confidence interval for the 75% quantile?

1 ☐ [21, 23]

2 ☐ [8, 36]

3* ☐ [21, 25]

4 ☐ [15, 17]

5 ☐ [13, 16]

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

We are asked for a non-parametric boot strap, and hence we should sample directly from data with replacement, this is `sim1`, `sim2`, `samp1` finds the 25% quantile while `samp2` find the non-parametric bootstrap confidence interval for the required 75% quantile, and the correct answer is [21,25] i.e. answer no. 3.

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

Question VI.5 (22)

We now want to investigate whether there has been a development in the number of earthquakes during the period. As a first investigation a 95 % non-parametric bootstrap confidence interval for the difference in mean value over two periods (called `earthquakes1` and `earthquakes2`) should be calculated. The R code below has been executed in order to investigate the question.

```
k <- 10000
n <- length(earthquakes)
n1 <- length(earthquakes1)
n2 <- length(earthquakes2)

sample1 <- replicate(k, sample(earthquakes1 - earthquakes2, replace = TRUE))
quantile(apply(sample1, 2, mean), probs = c(0.025, 0.975))

##   2.5%  97.5%
## 3.1995 7.9200

sample1 <- replicate(k, sample(earthquakes1, replace = TRUE))
sample2 <- replicate(k, sample(earthquakes2, replace = TRUE))
quantile(apply(sample1, 2, mean) -
         apply(sample2, 2, mean), probs = c(0.025, 0.975))

##   2.5%  97.5%
## 3.2200 8.1605

sample1 <- replicate(k,sample(sort(earthquakes1) -
                              sort(earthquakes2), replace = TRUE))
quantile(apply(sample1, 2, mean), probs = c(0.025, 0.975))
```

```
##  2.5% 97.5%
##   5.0   6.3
```

```
mean(earthquakes1 - earthquakes2) + c(-1, 1) * sd(earthquakes) * 1.96 / sqrt(n)
```

```
## [1] 4.377711 6.902289
```

```
mean(earthquakes1 - earthquakes2) + c(-1, 1) * 1.96 *
    sqrt(var(earthquakes1) + var(earthquakes2)) * sqrt(1/n1 + 1/n2)
```

```
## [1] 2.086492 9.193508
```

Which of the following intervals is a non-parametric bootstrap 95% confidence interval for the difference in mean?

1 ☐ [4.38, 6.90]

2* ☐ [3.22, 8.16]

3 ☐ [2.09, 9.19]

4 ☐ [3.20, 7.92]

5 ☐ [5.00, 6.30]

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

Since these are two different periods we should look have two independent samples (with replacements) from the two preiods. This is only done in the second result lines, the average for each simulaton is also calculated here and the correct quantile of the difference is extracted, hence the correct result is [3.22,8.16] or answer no. 2.

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

Continue on page 32

Exercise VII

A plant cooperative has designed a germination system for households, and it has been placed at many of their members. They are now in search of the right soil mix, and therefore they have run an experiment in which several of their members were testing five different soil mixtures. Each participant made five germinations simultaneously - one with each of the five soil mix.

The germination time in hours were recorded for all germinations using a simple camera system and the following observations have been collected:

	Mix 1	Mix 2	Mix 3	Mix 4	Mix 5
Participant 1	9.9	8.7	6.8	10.0	10.0
Participant 2	5.2	3.1	8.8	10.0	5.0
Participant 3	10.9	8.1	10.3	10.1	9.9
Participant 4	6.1	10.6	7.6	7.1	9.8
Participant 5	8.9	8.9	3.8	8.0	8.8
Participant 6	5.9	6.4	5.9	6.0	7.7
Participant 7	8.9	8.9	8.3	10.2	7.5
Participant 8	10.3	13.4	9.3	9.6	9.7

An analysis was carried out (note that some of the values have been replaced by letters and potential stars have been removed):

```
anova(lm(Time ~ Mix + Participant))

## Analysis of Variance Table
##
## Response: Time
##             Df Sum Sq Mean Sq F value Pr(>F)
## Mix          4    7.3    1.82    A     B
## Participant  7   77.5   11.07    C     D
## Residuals   28   89.1    3.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Continue on page 33

Question VII.1 (23)

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

1* ☐ There is no significant effect of mix, however there is a significant effect of participant, since the relevant p-values are 0.68 and 0.0083, respectively

2 ☐ There is neither a significant effect of mix nor participant, since the relevant p-values are 0.093 and 0.17, respectively

3 ☐ There is both a significant effect of mix and participant as the relevant p-values are 0.0034 and 0.014, respectively

4 ☐ There is both a significant effect of mix and participant as the relevant p-values are 0.023 and 0.57, respectively

5 ☐ There is a significant effect of mix, however, there is no significant effect of participant, since the relevant p-values are 0.0045 and 0.85, respectively

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

In order to solve the question we need the F-test statistics and the p-values these are calculated by

```
(F <- c(1.82, 11.07) / 3.18)

## [1] 0.572 3.481

(pv <- 1-pf(F,df1=c(4,7), df2=28))

## [1] 0.68491 0.00826
```

These are the p-values in answer no. 1, and on a 5% level this imply that there is no effect of mix but that the participant have a significant effect. Hence answer no. 1 is correct.

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

Question VII.2 (24)

How much of the total variation is explained by the model used?

1 ☐ $\frac{4+7}{28} = 0.393$

2 ☐ $\frac{7.3-1.82+77.5-11.07}{89.1-3.18} = 0.837$

3 ☐ $\frac{7.3+77.5}{89.1} = 0.952$

4* ☐ $\frac{7.3+77.5}{7.3+77.5+89.1} = 0.488$

5 ☐ $\frac{1.82+11.07+3.18}{7.3+77.5+89.1} = 0.0924$

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

The total sum of variance is $7.3 + 77.5 + 89.1 = 174$ of which $7.3 + 77.5 = 84.8$ is explained, hence the proportion explained is 48.7%.

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

Continue on page 35

Exercise VIII

As part of an UX design study, users were randomly presented for one of two webpage layouts (A and B). The users had to perform a specific action; the time spent (in seconds) was measured.

The following code reads the data into R, `A` is the users who were presented for layout A, whereas `B` is the other users who were presented for layout B.

```
A <- c(8.360, 6.377,  7.385, 6.245, 8.766, 6.848, 6.074, 6.310, 5.946, 8.270)
B <- c(8.806, 6.513, 10.209, 5.495, 6.513, 8.529, 8.354, 5.681, 7.553, 6.834)
```

Question VIII.1 (25)

What is the 99% confidence interval for the difference in means between the two groups A and B?

1 ☐ [6.674, 7.783]

2* ☐ [-2.093, 1.312]

3 ☐ [-1.409, 0.628]

4 ☐ [-1.627, 0.846]

5 ☐ [-2.116, 1.335]

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

Either calculate this interval by hand using theorem 3.54 from the book or use `t.test(A, B, conf.level = 0.99)` in R.

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

Question VIII.2 (26)

The usual t-test gave a p-value of 0.513. Find the right conclusion on significance level $\alpha = 0.05$ (both conclusion and argument must be correct):

1 ☐ The two variances are significantly different as the p-value is large.

2 ☐ The two means are significantly different as the p-value is large.

3* ☐ The two means can reasonably be assumed to be equal as the p-value is large.

4 ☐ The two means can reasonably be assumed to be equal as the p-value is small.

5 ☐ The two means are significantly different as the p-value is small.

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

The usual t-test is a comparison of mean values. In our case, the p-value is large, much larger than 0.05. Therefore we accept the null hypothesis, and the two means can reasonably be assumed to be equal.

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

Continue on page 37

Question VIII.3 (27)

The webpage designers would like to repeat the study with more people. It is assumed that the standard deviation within each group is 1.2, and that a t-test on level $\alpha = 0.05$ should be carried out.

The following code was run in R:

```
  power.t.test(n=40, delta=0.5, sd=1.2, sig.level=0.05)

##
##      Two-sample t test power calculation
##
##               n = 40
##           delta = 0.5
##              sd = 1.2
##       sig.level = 0.05
##           power = 0.4524276
##     alternative = two.sided
##
## NOTE: n is number in *each* group

power.t.test(power=0.80, delta=0.4, sd=1.2, sig.level=0.05)

##
##      Two-sample t test power calculation
##
##               n = 142.2466
##           delta = 0.4
##              sd = 1.2
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Continue on page 38

Which of the following statements is true?

1* ☐ The chance that a study with $n = 40$ in each group finds a significant difference between the means, if the real difference is 0.5, is around 45%.

2 ☐ The chance that a study with $n = 142$ in each group finds a significant difference between the means, if the real difference is 0.5, is around 80%.

3 ☐ The risk that a study with $n = 40$ in each group does not find a significant difference between the means, if the real difference is 0.5, is around 45%.

4 ☐ The risk that a study with $n = 142$ in each group does not find a significant difference between the means, if the real difference is 0.5, is around 80%.

5 ☐ The probability that $\mu_A \neq \mu_B$ is around 45% when $n = 40$, and increases to 80% when $n = 142$.

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

$\mu_A \neq \mu_B$ is not a probability statement and does not depend on n.

The power is expressing the probability of rejecting the hypothesis (NOT the risk of no rejection), and statement 2 is using a scenario based difference of delta=0.5, but the R-call is showing the result for delta=0.4. So the only true statement is: statement 1.

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

Continue on page 39

Exercise IX

Let $X_i \sim N(\mu_1, \sigma_1^2)$ and $Y_i \sim N(\mu_2, \sigma_2^2)$ be independent random variables, further assume that samples of the size n_1 and n_2 respectively are taken from each of the two populations. Let S_p^2 denote the usual pooled variance estimator.

Question IX.1 (28)

What is $V(S_p^2)$?

1 ☐ $\frac{1}{2}(\sigma_1^2 + \sigma_2^2)$

2* ☐ $2\frac{(n_1-1)\sigma_1^4+(n_2-1)\sigma_2^4}{(n_1+n_2-2)^2}$

3 ☐ $\frac{\sigma_1^4}{(n_1-1)^2} + \frac{\sigma_2^4}{(n_2-1)^2}$

4 ☐ $2\frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}{n_1+n_2-2}$

5 ☐ $(\sigma_1^4 + \sigma_2^4)$

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

The pooled variance is given by

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \tag{9}$$

and hence the variance of S_p^2 can be written as

$$V(S_p^2) = \frac{(n_1-1)^2 V(S_1^2) + (n_2-1)^2 V(S_2^2)}{(n_1+n_2-2)^2} \tag{10}$$

Further $V(S_i^2) = \frac{2\sigma_i^4}{n_i-1}$, hence

$$V(S_p^2) = \frac{(n_1-1)^2 \frac{2\sigma_1^4}{n_1-1} + (n_2-1)^2 \frac{2\sigma_2^4}{n_2-1}}{(n_1+n_2-2)^2} \tag{11}$$

$$= 2\frac{(n_1-1)\sigma_1^4 + (n_2-1)\sigma_2^4}{(n_1+n_2-2)^2} \tag{12}$$

which is answer no. 2

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

Now define

$$Q = \frac{n_1(\bar{X} - \mu_1)^2 + n_2(\bar{Y} - \mu_2)^2}{S_p^2} = \frac{D_1 + D_2}{S_p^2} \tag{13}$$

Question IX.2 (29)

Considering Q as a function of D_1, D_2 and S_p^2. What is the error propagation law expression for the variance of Q, expressed by σ_1^2, σ_2^2, $\mu_s = E[S_p^2]$, and $\sigma_s^2 = V[S_p^2]$. It is further given that $E[D_1] = \sigma_1^2$, $E[D_2] = \sigma_2^2$, $V[D_1] = 2\sigma_1^4$, and $V[D_2] = 2\sigma_2^4$?

1 ☐ $\frac{1}{\mu_s^2}\left(2\sigma_1^2 + 2\sigma_2^2 + \frac{(\sigma_1^2+\sigma_2^2)^2}{\mu_s}\sigma_s^2\right)$

2 ☐ $\frac{1}{\mu_s^2}\left(2\sigma_1^2 + 2\sigma_2^2 + \frac{\sigma_1^2+\sigma_2^2}{\mu_s^4}\sigma_s\right)$

3 ☐ $\frac{1}{\mu_s}\left(2\sigma_1^2 + 2\sigma_2^2 + \frac{(\sigma_1^2+\sigma_2^2)}{\mu_s}\sigma_s^2\right)$

4* ☐ $\frac{1}{\mu_s^2}\left(2\sigma_1^4 + 2\sigma_2^4 + \frac{(\sigma_1^2+\sigma_2^2)^2}{\mu_s^2}\sigma_s^2\right)$

5 ☐ $\frac{1}{(\sigma_1^2+\sigma_2^2)^2}\left(2\sigma_1^4 + 2\sigma_2^4 + \frac{(\sigma_1^2+\sigma_2^2)^2}{\mu_s}\sigma_s^2\right)$

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

We need the derivative of Q wrt. S_p^2, D_1 and D_2 these are given by

$$\frac{\partial Q}{\partial D_1} = \frac{1}{S_p^2} \tag{14}$$

$$\frac{\partial Q}{\partial D_2} = \frac{1}{S_p^2} \tag{15}$$

$$\frac{\partial Q}{\partial S_p^2} = -\frac{D_1 + D_2}{(S_p^2)^2} \tag{16}$$

hence by using independence between S_p^2, $\bar{X}$ and $\bar{Y}$ and incerting in the error propagation formula we get

$$V[Q] \approx \left(\frac{1}{\mu_s}\right)^2 2\sigma_1^4 + \left(\frac{1}{\mu_s}\right)^2 2\sigma_2^4 + \left(\frac{\sigma_1^2 + \sigma_2^2}{\mu_s^2}\right)^2 \sigma_s^2 \tag{17}$$

$$= \frac{1}{\mu_s^2}\left(2\sigma_1^4 + 2\sigma_2^4 + \frac{(\sigma_1^2 + \sigma_2^2)^2}{\mu_s^2}\sigma_s^2\right) \tag{18}$$

$$\tag{19}$$

which is answer number 4.

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

Question IX.3 (30)

Now suppose that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, what is the exact expression for $E(Q)$?

1* ☐ $2\frac{n_1+n_2-2}{n_1+n_2-4}$

2 ☐ $4\frac{n_1+n_2}{n_1+n_2-2}\sigma^2$

3 ☐ $4\frac{n_1+n_2}{n_1+n_2-2}$

4 ☐ $\frac{n_1+n_2-2}{n_1+n_2-4}\sigma^2$

5 ☐ $2\frac{n_1+n_2}{n_1+n_2-2}\sigma^2$

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

Remark 2.77 imply that the enumarator and the denomiator are independent, also under the assumptions $\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$, the enumerator is

$$Q_1 = n_1(\bar{X}-\mu_1)^2 + n_2(\bar{Y}-\mu_2)^2 \tag{20}$$

Now since $\bar{X} \sim N(\mu_1, \sigma^2/n_1)$ we have (same argument apply to $\bar{Y}$)

$$\frac{(\bar{X}-\mu_1)^2}{\sigma^2/n_1} \sim \chi^2(1) \tag{21}$$

and therefore

$$Q_1/\sigma^2 \sim \chi^2(2) \tag{22}$$

and

$$F = \frac{(Q_1/\sigma^2)/2}{(S_p^2/\sigma^2)} = \frac{1}{2}\frac{Q_1}{S_p^2} \sim F(2, n_1+n_2-2) \tag{23}$$

and therefore (with $Q = 2F$)

$$E[Q] = 2E[F] = 2\frac{n_1+n_2-2}{n_1+n_2-4} \tag{24}$$

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

The exam is finished. Have a great summer!