Technical University of Denmark *Page 1 of 44 pages.*

Written examination: June 27, 2019

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

Duration: 4 hours

Aids and facilities allowed: All

The questions were answered by

_____ _____ _____

(student number) (signature) (table number)

This exam consists of 30 questions of the "multiple choice" type, which are divided between 8 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	I.4	I.5	I.6	II.1	II.2	II.3	II.4
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer	1	5	3	3	3	5	2	5	3	4

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

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

The exam paper contains 44 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

A jogger who has a certain course that he runs frequently, decides to use this trip to measure if there is progress in his fitness. As a measure of fitness, he decides to use the average speed on this course.

The R output below shows the result of a simple linear regression with time measured in weeks (`time`), since the start of a training program, as explanatory variable and the speed measured in km/h (`speed`).

```
##
## Call:
## lm(formula = speed ~ time, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.686 -0.646 -0.375  0.696  2.705
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.4929     0.4191   25.04   <2e-16 ***
## time          0.0223     0.0138    1.62     0.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.11 on 38 degrees of freedom
## Multiple R-squared:  0.0647,Adjusted R-squared:  0.0401
## F-statistic: 2.63 on 1 and 38 DF,  p-value: 0.113
```

Continue on page 3

Question I.1 (1)

At significance level $\alpha = 0.05$ what can be concluded about the development of jogger's fitness (all parts of the answer must be correct)?

1* ☐ No significant development over time can be detected as $0.113 > 0.05$

2 ☐ A significant development over time can be shown as $0.0223 < 0.05$

3 ☐ No significant development over time can be detected as $0.0647 > 0.05$

4 ☐ A significant development over time can be demonstrated as $0.04 < 0.05$

5 ☐ A significant development over time can be demonstrated as $0.014 < 0.05$

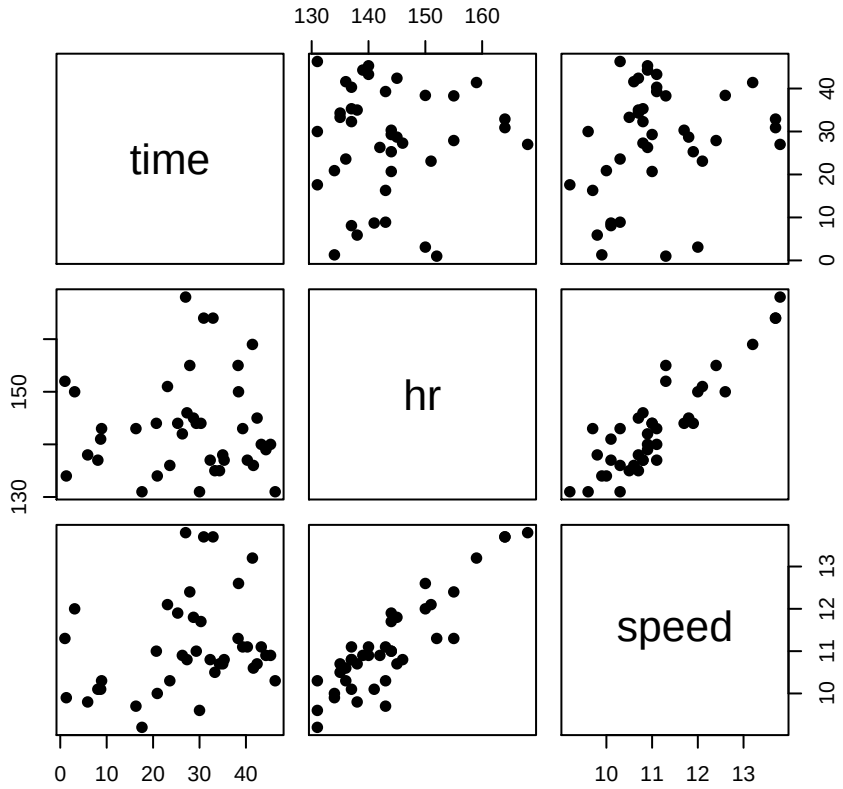
---------- FACIT-BEGIN ----------

No development corresponds to testing the hypothesis that the effect of time is zero, and the p-value for that test is 0.113 which is greater than the siginficance level, hence answer no. 1 is correct. See 5.13

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

In addition to the speed of the individual trips, the jogger has also measured his average pulse (beats/min) during the runs (`hr`), and he wants to include this in the model, so that the speed is modeled as a function of time and pulse. As part of the exploratory analysis, pairwise scatterplots are shown in Figure 1.

Figure 1



Continue on page 5

Question I.2 (2)

Based on Figure 1, one can conclude that:

1 ☐ There is a clear positive correlation between time and pulse and therefore there may be a collinarity problem

2 ☐ The correlation between pulse and speed is high, and therefore there is clearly a collinarity problem

3 ☐ There is a clear negative correlation between time and pulse, and therefore there is no collinarity problem

4 ☐ There is a clear negative correlation between time and speed, and therefore there is no collinarity problem

5* ☐ The correlation between pulse and speed is high, but there are no collinarity problems

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

The correlation between time and pulse is very weak (if there is any), hence answer no 1 is not correct.

The correlation between pulse and speed is quite strong, but since speed is the dependent variable this will not create any collinarity problems.

The correlation between time and puse is weak and it is not clear if it is negative, hence answer no 3 is not correct.

The correlation between time and speed is weak (hence not clear if it is negative) hence no 4 is wrong.

The correlation between pulse and speed is high, but collinarity is related to correlation between time and pulse (which is weak) and hence there is no collinarity problems. And answer no 5 is correct.

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

The jogger has now estimated a model with both pulse and time as explanatory variables, the result is seen in the R-output below.

```
summary(lm(speed ~ hr + time, data = dat))

##
## Call:
```

```
## lm(formula = speed ~ hr + time, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2849 -0.2066  0.0397  0.2738  0.7800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.00857    1.12243   -4.46  7.3e-05 ***
## hr           0.10831    0.00775   13.97  2.5e-16 ***
## time         0.02104    0.00557    3.78  0.00056 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.449 on 37 degrees of freedom
## Multiple R-squared:  0.851,Adjusted R-squared:  0.843
## F-statistic:  106 on 2 and 37 DF,  p-value: 5.08e-16
```

Continue on page 7

The model can be written as

$$Y_i = \beta_0 + \texttt{hr}_i\beta_1 + \texttt{time}_i\beta_2 + \epsilon_i \tag{1}$$

where ϵ_i is assumed to meet the usual conditions of the linear regression model.

Question I.3 (3)

What is the usual 95% confidence interval for the parameter β_1?

1 ☐ $[0.106; 0.111]$

2 ☐ $[-0.036; 0.252]$

3* ☐ $[0.093; 0.124]$

4 ☐ $[0.080; 0.137]$

5 ☐ $[0.101; 0.116]$

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

As define in 6.5 the confidence interval can be calculated by

$$\hat{\beta}_1 \pm t_{0.975}\hat{\sigma}_{\beta_1}$$

where we should use 37 degrees of freedom for the t-distribution. The parameter estimate and the standard error can be read from the summary and the numerical value can be calculated from

```
0.10831 + c(-1,1) * 0.00775 * qt(0.975,df=37)

## [1] 0.092607 0.124013
```

which is answer no 3.

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

The model can be written in matrix-vector notation as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

Question I.4 (4)

How many rows and columns does $\boldsymbol{X}$ have?

1 ☐ $\boldsymbol{X}$ has 37 rows and 3 columns.

2 ☐ $\boldsymbol{X}$ has 2 rows and 2 columns.

3* ☐ $\boldsymbol{X}$ has 40 rows and 3 columns.

4 ☐ $\boldsymbol{X}$ has 37 rows and 2 columns.

5 ☐ $\boldsymbol{X}$ has 3 rows and 3 columns.

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

As explained in chapter 6.6, the number of columns equals the number of parameters (i.e. 3), and the number of rows equals the number of observations (degrees of fredom plus number of parameters, i.e. 3+37=40), i.e. answer no. 3.

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

Continue on page 9

Question I.5 (5)

In order to predict the speed of a race at the time `time` = 55 weeks, the jogger would like to give an estimate of the speed if the average pulse is kept at 160 beats/min. What is the prediction of the speed under these conditions (when the model taking both heart rate and time into account is used)?

1 ☐ 10.49 km/h

2 ☐ 4.31 km/h

3* ☐ 13.48 km/h

4 ☐ 13.97 km/h

5 ☐ 11.7 km/h

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

The prediction is calculated by

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot 160 + \hat{\beta}_2 \cdot 55$$

the estimates are found in the in the summary above and the numerical value is calculated by

```
-5.008565+0.108308*160+0.021037*55
```

```
## [1] 13.478
```

which is answer no. 3.

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

The jogger also wants a confidence interval for the above prediction. A confidence interval can be written on the form

$$CI = \hat{Y}_{new} \pm C \tag{3}$$

As a help in the further calculations, the matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ calculated to

```
XX.Inv <- matrix(c(6.2,-4.3e-2,-3.7e-3,
                   -4.3e-2,3e-4,-3.5e-6,
                   -3.7e-4,-3.5e-6,1.5e-4),
                 ncol=3)
colnames(XX.Inv) <- rownames(XX.Inv) <- c("(Intercept)", "hr", "time")
XX.Inv
```

```
##             (Intercept)       hr     time
## (Intercept)      6.2000 -4.3e-02 -3.7e-04
## hr              -0.0430  3.0e-04 -3.5e-06
## time            -0.0037 -3.5e-06  1.5e-04
```

Continue on page 11

Question I.6 (6)

If a 95% confidence interval is chosen, what is C then?

1 ☐ 1.09

2 ☐ 0.24

3 ☐ 0.69

4 ☐ 1.40

5* ☐ 0.49

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

The value of C is given by

$$t_{0.975}\hat{\sigma}\sqrt{\boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T}$$

where $\boldsymbol{x}_{new} = [1, 160, 55]$ and hence the numerical value can be calculated by

```
x <- c(1,160,55)

sqrt(x %*% XX.Inv %*% x) * 0.4493 * qt(0.975, df = 37)

##          [,1]
## [1,] 0.48881
```

and the correct answer is no. 5. See equation 6-48 for more details.

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

Continue on page 12

Exercise II

The relation between a so-called "thermal sensation vote" (TSV), categorized as 0, 1 and 2, and the speed (FanSpeed) the employee set a fan to (also categorized as 0, 1, 2), have been measured in an office building. The result is shown in the table below

	FanSpeed=0	FanSpeed=1	FanSpeed=2	Sum
TSV=0	97	20	20	137
TSV=1	40	24	10	74
TSV=2	8	8	10	26
Sum	145	52	40	237

Question II.1 (7)

In this question we only consider persons who have specified TSV to 0. What is the usual 95% confidence interval for the probability that persons who specify a TSV of 0 also choose a fanpeed of 0?

1 ☐ [0.76;0.90]

2* ☐ [0.63;0.78]

3 ☐ [0.67;0.75]

4 ☐ [0.79;0.86]

5 ☐ [0.70;0.71]

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

As defined in 7.3, the estimate of the probability iscalculated by $\hat{p} = \frac{x}{p}$, and the confidence interval is calculated by

$$\hat{p} \pm z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The numerical values can be calculated in R by

```
n <- 137
x <- 97
p.hat <- x / n
p.hat

## [1] 0.70803

se <- sqrt(p.hat * (1 - p.hat) / n)
round((CI <- p.hat + c(-1, 1) * qnorm(0.975) * se), digits = 2)

## [1] 0.63 0.78
```

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

Question II.2 (8)

Again, we consider only one group (i.e. one TSV) and only the probability of choosing FanSpeed = 0 or FanSpeed different from 0. A new study is planned and in this connection one want to examine how many persons should be in the new study if one want a margin of error of 0.1 and assume that the probability of choosing FanSpeed 0 is 0.3 and uses significance level $\alpha = 0.05$?

1 ☐ 90

2 ☐ 322

3 ☐ 277

4 ☐ 42

5* ☐ 81

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

As stated in 7.13, the number of persons needed in the new experiment is given by

$$n = p(1-p)\left(\frac{z_{1-\alpha/2}}{ME}\right)$$

With $p = 0.3$, $\alpha = 0.05$, and $ME = 0.1$ the numerical value can be calculated by

```
(n <- 0.3*0.7/0.1^2*qnorm(0.975)^2)

## [1] 80.671
```

Rounding to the integer value give 81 (answer no 5).

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

Continue on page 14

Question II.3 (9)

What is the 95% confidence interval for the difference between the probability of setting FanSpeed to 0 between the TSV 0 and TSV 1 groups, and what conclusion does it lead to?

1 ☐ There is a significant difference since $0 \notin [0.51, 0.78]$

2 ☐ There is not a significant difference since $0 \in [0.51, 0.78]$

3* ☐ There is a significant difference since $0 \notin [0.03, 0.3]$

4 ☐ There is not a significant difference since $0 \notin [0.09, 0.24]$

5 ☐ There is a significant difference since $0 \notin [0.09, 0.24]$

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

See method 7.15. The confidence interval is given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

with $\hat{p}_i = \frac{x_i}{n_i}$ and the values of n_i and x_i given in the table. The numerical values are calculated by:

```
n1 <- 137
n2 <- 74
x1 <- 97
x2 <- 40
phat1 <- x1/n1
phat2 <- x2/n2
round(phat1-phat2+c(-1,1)*qnorm(0.975)*
      sqrt(phat1*(1-phat1)/n1+phat2*(1-phat2)/n2),digits=2)

## [1] 0.03 0.30
```

Which is answer no. 3.

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

One now wants to test the entire distribution, ie. we consider the whole table above. To perform tests for the entire distribution, the expected number in each cell needs to be calculated, under the hypothesis of independence in the classification criteria.

Question II.4 (10)

What is the expected number at TSV = 0 and FanSpeed = 2?

1 ☐ 20

2 ☐ 5.7

3 ☐ 15.9

4* ☐ 23.1

5 ☐ 8

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

As explained in chapter 7.5.1, the expected number is calculated by

$$e_{ij} = \frac{n_j x_i}{n}$$

where n_j is the column total, x_i is the row total and n is the total number of observations. We are looking for e_{13}, i.e.

```
n3 <- 40
x1 <- 137
n <- 237
n3*x1/n

## [1] 23.122
```

which is anwer no. 4.

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

Question II.5 (11)

The usual test statistics for whether the distribution is the same in the 3 groups is calculated to 22.72, what is the conclusion on significance level $\alpha = 0.05$ (all parts of the answer needs to be correct)?

1 ☐ There is no significant difference since $4.6 \cdot 10^{-5} < 0.05$

2 ☐ There is no significant difference since $0.042 < 0.05$

3* ☐ There is a significant difference since $0.00014 < 0.05$

4 ☐ There is a significant difference since $0.042 < 0.05$

5 ☐ There is a significant difference since $4.6 \cdot 10^{-5} < 0.05$

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

See method 7.22. The test statistics should be compared with a χ^2-distribution with $(r-1)(c-1)$ degrees of freedom, in our case $r = c = 3$ hence the degrees of freedom is 4, and the p-value is calculated by

$$P(\chi^2 > 22.72)$$

with $\chi^2 \sim \chi^2(4)$, which can be calculated by

```
1-pchisq(22.72,df=4)

## [1] 0.00014402
```

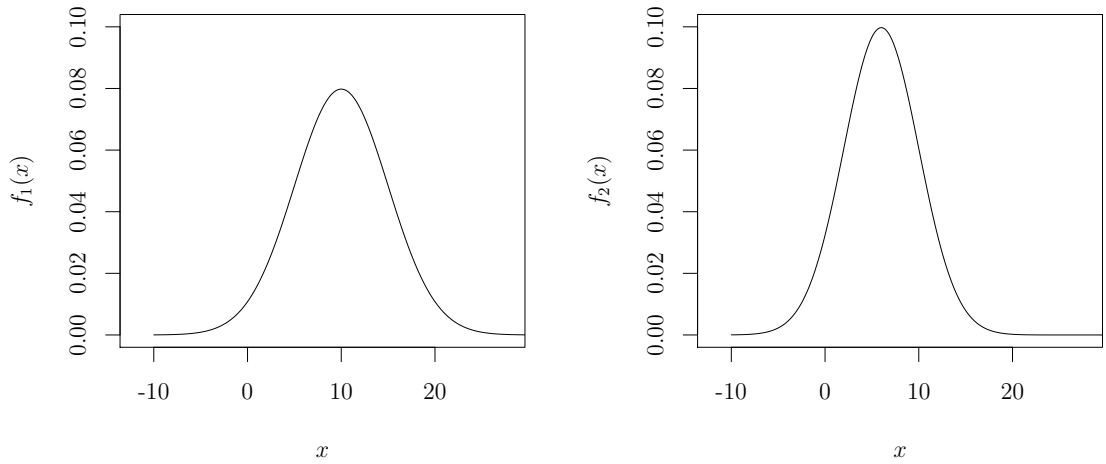
this is answer no. 3.

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

Continue on page 17

Exercise III

Let two independent random variables be $X_1 \sim N(10, 5^2)$ and $X_2 \sim N(6, 4^2)$. The probability density functions of the two variables are plotted in the figure below: X_1 to the left, and X_2 to the right:



Question III.1 (12)

Which of the following statements is not true?

1 ☐ $\mathrm{E}(X_1) > \mathrm{E}(X_2)$

2 ☐ $\mathrm{V}(X_1) > \mathrm{V}(X_2)$

3 ☐ $P(X_1 < 0) \neq P(X_2 < 0)$

4* ☐ $P(X_1 > 10) < P(X_2 > 10)$

5 ☐ $P(X_1 > 10) = P(X_2 < 6)$

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

Lets go through the answers:

- $\mathrm{E}(X_1) > \mathrm{E}(X_2)$ is true since $10 > 6$.
- $\mathrm{V}(X_1) > \mathrm{V}(X_2)$ is true since $5^2 > 4^2$.
- $P(X_1 < 0) \neq P(X_2 < 0)$ is true since $0.023 \neq 0.067$.

```
pnorm(0, mean=10, sd = 5)

## [1] 0.02275
```

```r
pnorm(0, mean=6, sd = 4)

## [1] 0.066807
```

- $P(X_1 > 10) < P(X_2 > 10)$ is false since $0.5 < 0.16$ (hence the correct answer).

```r
1 - pnorm(10, mean=10, sd = 5)

## [1] 0.5

1 - pnorm(10, mean=6, sd = 4)

## [1] 0.15866
```

- $P(X_1 > 10) = P(X_2 < 6)$ is true since $0.5 = 0.5$.

```r
1 - pnorm(10, mean=10, sd = 5)

## [1] 0.5

pnorm(6, mean=6, sd = 4)

## [1] 0.5
```

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

Question III.2 (13)

What is the probability that $X_1 > 3 \cdot X_2$?

1* ☐ 0.27

2 ☐ 0.48

3 ☐ 0.19

4 ☐ 0.52

5 ☐ 0.055

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

First rewrite

$$P(X_1 > 3 \cdot X_2) = P(X_1 - 3 \cdot X_2 > 0) \tag{4}$$

Then use Therorem 2.40, which states that any linear combination of normal distributed random variables is also normal distributed, and the identities for a linear combinations of random variables in Theorem 2.56. Thus the mean is

$$\mathrm{E}(X_1 - 3 \cdot X_2) = \mathrm{E}(X_1) - 3\,\mathrm{E}(X_2) = 10 - 3 \cdot 6 = -8 \tag{5}$$

and the variance is

$$\mathrm{V}(X_1 - 3 \cdot X_2) = \mathrm{V}(X_1) + 9\,\mathrm{V}(X_2) = 5^2 + 9 \cdot 4^2 = 169 \tag{6}$$

Now the seeked probability can be found by

```
1 - pnorm(0, mean=-8, sd=sqrt(169))

## [1] 0.26915
```

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

Continue on page 20

Question III.3 (14)

What is the probability that $X_1 - 10 > \sqrt{(X_2-6)^2} = |X_2 - 6|$?

1 ☐ 0.32

2* ☐ 0.29

3 ☐ 0.45

4 ☐ 0.5

5 ☐ 0.96

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

We are looking at the probabilty

$$P\left(X_1 - 10 > \sqrt{(X_2-6)^2}\right) = P\left(\frac{X_1 - 10}{\sqrt{(X_2-6)^2}} > 1\right) \tag{7}$$

where $X_1 \sim N(10, 5^2)$ and $X_2 \sim N(6, 4^2)$. In order to transform to standard normal distribution (see theorem 2.43) we can write

$$P\left(\frac{X_1 - 10}{\sqrt{(X_2-6)^2}} > 1\right) = P\left(\frac{5\frac{X_1-10}{5}}{4\sqrt{\left(\frac{X_2-6}{4}\right)^2}} > 1\right) \tag{8}$$

$$= P\left(\frac{\frac{X_1-10}{5}}{\sqrt{\left(\frac{X_2-6}{4}\right)^2}} > \frac{4}{5}\right) \tag{9}$$

now $Z = \frac{X_1-10}{5} \sim N(0,1)$, $\frac{X_2-6}{4} \sim N(0,1)$ and following theorem 2.79 $Q = \left(\frac{X_2-6}{4}\right)^2 \sim \chi^2(1)$, since X_1 and X_2 are independent we also have (see theorem 2.87)

$$T = \frac{\frac{X_1-10}{5}}{\sqrt{\left(\frac{X_2-6}{4}\right)^2}} = \frac{Z}{\sqrt{Q}} \sim t(1) \tag{10}$$

and the probability can be calculated as

$$P\left(X_1 - 10 > \sqrt{(X_2-6)^2}\right) = P\left(T > \frac{4}{5}\right) = 1 - P\left(T < \frac{4}{5}\right) = 1 - F_T\left(\frac{4}{5}\right) \tag{11}$$

where F_T is the distribution function for the t-distribution with 1 degree of freedom and can be calculated by

```r
1-pt(4/5,df=1)

## [1] 0.28522
```

As an alternative we could use parametric bootstrap to get the approximative result

```r
k <-1000000
x <- rnorm(k,mean=10,sd=5)
y <- rnorm(k,mean=6,sd=4)
sum((x-10)>sqrt((y-6)^2))/k

## [1] 0.28478
```

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

Continue on page 22

Exercise IV

In a leisure club for engineers different card and dice games are being played. To have a chance to win it is of course always important to know about the probabilities in a game, but it is especially important in this club for engineers, where all members have in-depth knowledge of statistics and probability theory.

Question IV.1 (15)

In the popular card game casino, each player is dealt 4 cards in each round. In this and the next question, it can be assumed that the cards are mixed such that the order of the cards is completely random in the stack from which the cards are drawn. The game is without jokers, i.e. there are 52 cards and 13 of each suit.

Before the game starts, a club member wants to make a calculation of the following: What is the probability that a player will be dealt a hand in the first round, where all 4 cards are spades?

1 ☐ 0.11%

2* ☐ 0.26%

3 ☐ 1.24%

4 ☐ 5.23%

5 ☐ 9.79%

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

See chapter 2.3.2. We use a hypergeometric distribution since the cards are not put pack. Draw 4 cards with 13 out of 52 spades, and 52-13 not spades and number of spades drawn equal 4:

```
dhyper(x=4, m=13, n=52-13, k=4)

## [1] 0.0026411

## or
13/52 * 12/51 * 11/50 * 10/49

## [1] 0.0026411
```

which equal 0.26%.

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

Question IV.2 (16)

During a 2-player casino game, as described in the previous question, one player has remembered that 10 spades have been played to the table. There are 16 cards left in the stack. Which of the following R calls calculates the probability that the player is dealt a hand where none of the 4 cards are spades?

1 ☐ `13/16 * 12/16 * 11/16 * 10/16`

2 ☐ `dbinom(13, size = 4, prob = 13/16)`

3 ☐ `dbinom(0, size = 4, prob = 3/16)`

4 ☐ `dhyper(0, m = 4, n = 16, k = 3)`

5* ☐ `dhyper(4, m = 13, n = 3, k = 4)`

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

Again a hyper geometric distribution (chapter 2.3.2), Directly calculated by:

```
13/16 * 12/15 * 11/14 * 10/13
```

```
## [1] 0.39286
```

or with the hypergeometric distribution:

```
dhyper(0, 3, 13, 4)
```

```
## [1] 0.39286
```

```
dhyper(4, 13, 3, 4)
```

```
## [1] 0.39286
```

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

Continue on page 24

Question IV.3 (17)

The dice game yatzy with 6 dices is also played in the gaming club. Each dice can be assumed to be fair so that there is exactly 1/6 probability of every outcome. In each round, a player can roll the dices up to 3 times, however no matter what, in each round a "first roll" with all 6 dices rolled simultaneously must be carried out. There are 18 rounds in a yatzy game.

Hence, each player rolls 18 times a "first roll" where all 6 dices are rolled simultaneously.

A yatsy is when all 6 dices show the same number.

If only a single player is considered, what is then the probability to get at least one yatsy on a "first roll" during a single game?

1 ☐ 1.4%

2 ☐ 0.69%

3* ☐ 0.23%

4 ☐ 0.15%

5 ☐ 0.0038%

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

First calculate the probability to roll a yatzy in a "first roll". There are 6 possible yatzy (either with ones, twos, ..., sixes), so consider that the first dice can settle on any number (1 to 6) and the the remaining 5 have to settle on that number, hence the probablity of yatzy in the first roll is

```
(1/6)^5
```

```
## [1] 0.0001286
```

and we can calculate the probabity of rolling at least one yatzy by

$$P(X > 0) = 1 - P(X = 0)$$

with $X \sim Binom((1/6)^5, 18)$ or numerically

```
(1 - dbinom(x=0, size=18, prob=(1/6)^5))
```

```
## [1] 0.0023123
```

which is answer no. 3.

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

Continue on page 26

Exercise V

A factory produces chocolate easter eggs with three different types of filling: chocolate, marzipan, or crunchy. In the following, we consider a dataset which contains the weights of altogether 9 different eggs, three with each type of filling. The data is read into R in two vectors:

```
weight <- c(3.02, 2.98, 2.95, 3.13, 3.06, 3.12, 2.88, 2.92, 2.86)
filling <- c("chocolate", "chocolate", "chocolate", "marzipan", "marzipan",
             "marzipan", "crunchy", "crunchy", "crunchy")
```

Let Y_{ij} be the weight of the jth egg with filling i (i = chocolate, marzipan or crunchy, $j = 1, 2,$ or 3). A statistical model of the following form is used to describe the data: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, where the ε_{ij} are independent and $N(0, \sigma^2)$-distributed, and μ, the α_i's and σ^2 are unknown parameters.

Question V.1 (18)

Which of the following interpretations of the statistical model is correct?

1 ☐ The model assumes that all eggs have the same mean weight.

2 ☐ The model describes the linear relation between the type of filling and the mean weight of an egg.

3 ☐ The model says that weight differences between eggs are solely due to random noise.

4* ☐ The model allows the mean weight of an egg to depend on its filling.

5 ☐ None of the above interpretations are correct.

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

The model is a one-way ANOVA model where the different types of filling determine the groups. The mean of eggs is affected by both the overall mean μ and the group differences α (see equation 8-4)

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

Question V.2 (19)

Give the usual estimate of μ.

1 ☐ $\hat{\mu} = 2.89$

2 ☐ $\hat{\mu} = 2.95$

3* ☐ $\hat{\mu} = 2.99$

4 ☐ $\hat{\mu} = 3.04$

5 ☐ $\hat{\mu} = 3.10$

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

An estimate of the overall mean μ may be found by taking the overall average of the weights. The following R-code gives the correct result (when rounded to two decimals):

```
mean(weight)

## [1] 2.9911
```

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

Continue on page 28

Question V.3 (20)

The following code was run in R:

```
anova(lm(weight ~ filling))

## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value  Pr(>F)
## filling    2 0.0707  0.0353    29.4 0.00079 ***
## Residuals  6 0.0072  0.0012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Which of the following answer options is correct? (Both the argument and conclusion must be true).

1 ☐ As 29.454 is less than the 0.99 quantile of an F-distribution with 6 and 2 degrees of freedom, there is a significant difference between the weight of eggs with different fillings at a 1% significance level.

2 ☐ As 29.454 is greater than 0.95, there is a significant difference between the weight of eggs with different fillings at a 5% significance level.

3 ☐ As 29.454 is less than the 0.99 quantile of an F-distribution with 6 and 2 degrees of freedom, there is no significant difference between the weight of eggs with different fillings at a 1% significance level.

4 ☐ As 0.0007899 is less than 0.05, there is no significant difference between the weight of eggs with different fillings at a 5% significance level.

5* ☐ None of the above answer options are correct.

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

Lets go through the answers:

1: If the F-test statistic is smaller than the 0.99 quantile of the F-distribution (the critical value), the conclusion would be that there is no significant difference between the weight of eggs with different filling. So 1 is false.

2: You cannot compare the F-test statistic with 0.95. You need to either calculate a critical value (0.99 quantile of the F-distribution) or the p-value. So 2 is false.

3: Since we can read from the p-value (0.0008) that there is a significant difference, we also know that the F-test statistic is not smaller than the critical value. So 3 is also false.

4: If the p-value is smaller than the significance level we conclude that there is a difference. Therefore 4 is wrong.

Since none of the above answers are correct, answer 5 is correct.

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

Continue on page 30

Question V.4 (21)

Give a post-hoc pairwise 95% confidence interval for the expected weight difference between eggs with chocolate and marzipan filling (i.e for $\mu_{marzipan} - \mu_{chocolate}$).

1* ☐ [0.05; 0.19]

2 ☐ [0.03; 0.21]

3 ☐ [-0.05; 0.29]

4 ☐ [0.11; 0.13]

5 ☐ None of the above intervals are correct.

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

Using Method 8.9 with $t_{0.975} = 2.447$ being the 0.975 quantile of the t-distribution with $9-3 = 6$ degrees of freedom, and $MSE = 0.0012$ from the R-output above, the interval may be computed as follows (and then rounded to two decimals):

```
round(3.103333-2.983333 + c(-1,1) * qt(0.975,df=6) * sqrt(0.0012 * (1/3 + 1/3)),
      digits=2)

## [1] 0.05 0.19
```

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

We are now informed that the eggs also have different colours: red, blue, or green. For each type of filling, there is exactly one egg of each colour. Information about the colours of the nine eggs in the dataset are read into R in a third vector named `colour`.

In next question, a statistial model of the following form is used to describe the data: $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, where the ε_{ij}s are independent and $N(0, \sigma^2)$-distributed, and μ, the α_is, the β_is, and σ^2 are the unknown parameters. The following code has been run in R, and parts of the output have been replaced with letters:

```
anova(lm(weight ~ filling + color))

Analysis of Variance Table

Response: vaegt
          Df   Sum Sq  Mean Sq F value   Pr(>F)
filling    2 0.070689 0.035344 25.9673 0.005114 **
```

```
color      2 0.001756 0.000878       A        B
Residuals  4 0.005444 0.001361
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Continue on page 32

Question V.5 (22)

Determine the correct values of A and B in the R-output above.

1 ☐ $A = 0.645$, $B = 0.428$

2 ☐ $A = 1.550$, $B = 0.428$

3 ☐ $A = 0.645$, $B = 0.683$

4 ☐ $A = 1.550$, $B = 0.317$

5* ☐ $A = 0.645$, $B = 0.572$

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

The F-test statistic (A) and the corresponding p-value (B) may be calculated as described in table 8.2.2:

$$F_{obs} = \frac{MS(Tr)}{MSE} P - value = P(F > F_{obs})$$

where the test statistic follows an F-distribution with $k-1$ and $n-k$ degrees of freedom.

By using the values from the output above we compute (then rounded to the relevant number of decimals):

```
# Test statistic
( Fstat <- 0.000878/0.001361 )

## [1] 0.64511

# P-value
1-pf(Fstat, df1=2, df2=4)

## [1] 0.5717
```

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

Continue on page 33

Exercise VI

The concentration of glucose before (c_0) and after (c_1) an enzymatic reaction has been measured 10 times by a reducing sugar analysis. For each of the 10 experiments the difference (c_{diff}) has been determined, such that $c_{diff} = c_1 - c_0$. The concentration difference can be assumed to be normally distributed. The concentrations and differences are shown in the table below, along with their average and standard deviations.

	1	2	3	4	5	6	7	8	9	10	$\bar{c}_i$	s_i
c_0	22.2	18.5	15.1	22.7	16.2	22.2	17.7	21.5	15.5	19.0	19.06	2.94
c_1	13.3	18.9	10.6	14.7	14.9	15.1	16.6	8.7	13.0	7.2	13.30	3.59
c_{diff}	-8.9	0.4	-4.5	-8.0	-1.3	-7.1	-1.1	-12.8	-2.5	-11.8	-5.76	4.65

Data can be entered in R with the code shown below

```
c0 <- c(22.2, 18.5, 15.1, 22.7, 16.2, 22.2, 17.7, 21.5, 15.5, 19.0)
c1 <- c(13.3, 18.9, 10.6, 14.7, 14.9, 15.1, 16.6, 8.7, 13.0, 7.2)
dif <- c(-8.9, 0.4, -4.5, -8.0, -1.3, -7.1, -1.1, -12.8, -2.5, -11.8)
```

Question VI.1 (23)

Based on the data above, what is the 95% confidence interval for the mean concentration difference?

1 ☐ $-5.76 \pm 2.23 \cdot \frac{4.65}{\sqrt{10}}$

2* ☐ $-5.76 \pm 2.26 \cdot \frac{4.65}{\sqrt{10}}$

3 ☐ $-5.76 \pm 2.26 \cdot \frac{4.65}{\sqrt{9}}$

4 ☐ $-5.76 \pm 1.81 \cdot \frac{21.61}{\sqrt{10}}$

5 ☐ $-5.76 \pm 2.26 \cdot \frac{21.61}{\sqrt{20}}$

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

This is a paired set up and hence we should use the one-sample set up for the differences in concentration (theorem 3.33). The confidence interval is given by

$$\bar{x} \pm t_{0.975} \frac{s_{diff}}{\sqrt{n}}$$

where the degrees of freedom for the t-distribution is n-1=9. Which is

```r
qt(0.975,df=9)

## [1] 2.2622
```

I.e. the correct answer is

$$-5.76 \pm 2.26 \frac{4.65}{\sqrt{10}}$$

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

Continue on page 35

Question VI.2 (24)

A one-sample t-test with a two-sided alternative hypothesis has been performed to find out if the mean concentration difference is significantly different from 0 when significance level $\alpha = 0.05$ is used. The usual test statistic was calculated to 3.918, which of the following statements is correct (all parts of the answer must be correct)?

1* ☐ There is a significant difference since $0.0035 < 0.05$.

2 ☐ There is a significant difference since $0.0044 < 0.05$.

3 ☐ There is no significant difference since $0.0044 < 0.05$.

4 ☐ There is a significant difference since $0.0028 < 0.05$.

5 ☐ There is no significant difference since $0.0028 < 0.05$.

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

As seen from method 3.36 the p-value is calculated as

$$2P(T > |t_{obs}|) = 2P(T > 3.918)$$

where $T \sim t(9)$ hence the p-value is

```
(1-pt(3.918,df=9))*2

## [1] 0.0035217
```

which is smaller than the sigfificance level and hence there is a significant difference.

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

It is now decided that the distribution assumption is questionable. Therefore it is decided to construct a confidence interval for the difference based on non-parametric bootstrap, for this purpose the following R code have been executed

```
set.seed(1342)
k <- 10000

simsamples <- replicate(k, sample(dif, replace = TRUE))
simmeans <- apply(simsamples, 2, mean)
quantile(simmeans, c(0.025, 0.975))
```

```
##    2.5%   97.5%
## -8.5203 -3.0898

quantile(simmeans, c(0.05, 0.95))

##    5%   95%
## -8.12 -3.49

simsamples <- replicate(k, sample(dif, replace=TRUE))
simmeans <- apply(simsamples, 2, median)
quantile(simmeans, c(0.025, 0.975))

##  2.5% 97.5%
## -9.45 -1.30

simsamples0 <- replicate(k, sample(c0, replace=TRUE))
simsamples1 <- replicate(k, sample(c1, replace=TRUE))
simmeans <- apply(simsamples1, 2, mean) - apply(simsamples0, 2, mean)
quantile(simmeans, c(0.025, 0.975))

##    2.5%   97.5%
## -8.4800 -3.0198

quantile(simmeans, c(0.05, 0.95))

##    5%   95%
## -8.05 -3.49
```

Question VI.3 (25)

What is the 95% confidence interval based on non-parametric bootstrap?

1* ☐ [-8.52; -3.09]

2 ☐ [-8.12; -3.49]

3 ☐ [-9.45; -1.3]

4 ☐ [-8.48; -3.02]

5 ☐ [-8.05; -3.49]

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

The bootstrap should be based on the vector `dif`, which should be sampled with replacement, and the average of each sample should be calculated, and finally we should calculate the 0.025

and 0.975 quantiles, this is done in the first 3 lines after `k`, and hence the result is [-8.52,-3.09], which is answer no. 1.

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

Continue on page 38

Exercise VII

During a sports competition at an elementary school an ambitious group of boys from 5th grade challenges a group of girls from 7th grade. The 23 boys (n_1) claim that they are faster on average when sprinting over a distance of 100m compared to the 26 girls (n_2). The average time for the boys and girls turns out to be 15.06s and 15.40s, respectively. Assume the data in the two groups to be normally distributed.

Question VII.1 (26)

Based on the R output below, where the variables `boys` and `girls` contain all running times for boys and girls respectively. Which of the following statements is correct?

```
t.test(x = boys, y = girls)

##
##  Welch Two Sample t-test
##
## data:  boys and girls
## t = -1.25, df = 42, p-value = 0.22
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.89847  0.21151
## sample estimates:
## mean of x mean of y
##    15.057    15.400
```

1 ☐ The 95% confidence interval contains zero, hence the boys are significantly faster than the girls.

2 ☐ The p-value is larger than the significance level $\alpha = 0.05$, hence the boys are significantly faster than the girls.

3* ☐ The 95% confidence interval contains zero, hence the boys are not significantly faster than the girls.

4 ☐ The p-value is larger than the significance level $\alpha = 0.05$, hence the boys are significantly slower than the girls.

5 ☐ The two sample t-test cannot be applied in this case because $n_1 \neq n_2$.

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

The confidence interval is calculated to [-0.9,0.2] which contaian zero hence there is no significant difference, this is answer no. 3. The p-value is larger than the significance level and hence there is no significant difference (i.e. 2 and 4 is wrong).

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

Continue on page 40

Question VII.2 (27)

State the null and two-sided alternative hypothesis for the two sample t-test above (in the answers, μ_1 is the mean value of boys and μ_2 is the mean of girls).

1 ☐ $H_0 : \mu_1 = 0$ and $H_1 : \mu_1 \neq 0$.

2 ☐ $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 < \mu_2$.

3 ☐ $H_0 : \hat{\mu}_1 = \hat{\mu}_2$ and $H_1 : \hat{\mu}_1 < \hat{\mu}_2$.

4 ☐ $H_0 : \hat{\mu}_1 = \hat{\mu}_2$ and $H_1 : \hat{\mu}_1 \neq \hat{\mu}_2$.

5* ☐ $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$.

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

The null hypothesis is $H_0 : \mu_1 = \mu_2$ since we are checking if the two true means are the same (and because it is the true mean there is no hat) and the alternative is that they are not the same $H_1 : \mu_1 \neq \mu_2$, hence 5 is correct. Answer no 4 is wrong since the hypothesis is about the true values not the estimates.

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

The standard deviation for boys and girls respectively is calculated below.

```
sd(boys)

## [1] 1.0548

sd(girls)

## [1] 0.8419
```

Question VII.3 (28)

If we assume that the true standard deviation for boys and girls is the same (i.e. $X_{1,i} \sim N(\mu_1, \sigma^2)$ and $X_{2,i} \sim N(\mu_2, \sigma^2)$), we also define S_1^2 and S_2^2 in the usual way (with $n_1 = 23$, and $n_2 = 26$), what is $P\left(\frac{S_1}{S_2} > \frac{1.05}{0.84}\right)$ then?

1 ☐ 0.704

2* ☐ 0.141

3 ☐ 0.196

4 ☐ 0.293

5 ☐ 0.853

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

See theorem 2.98. The probability can be written as

$$P\left(\frac{S_1}{S_2} > \frac{1.05}{0.84}\right) = P\left(\frac{S_1^2}{S_2^2} > \frac{1.05^2}{0.84^2}\right)$$

Further $F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$ and the probability is

$$P\left(F > \frac{1.05^2}{0.84^2}\right) = 1 - P\left(F < \frac{1.05^2}{0.84^2}\right)$$

which can be calculated by

```
1-pf((1.05/0.84)^2,df1=23-1,df2=26-1)

## [1] 0.1406
```

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

Continue on page 42

Exercise VIII

When predicting electricity produced by wind power, the so-called "power curve" is an important component. The power curve describes the relationship between wind speed and produced power. A simple model for the power curve is

$$p(W) = \frac{1}{1 + e^{-\gamma(W-\alpha)}} \tag{12}$$

where p is the relative power (i.e. $p \in (0, 1)$) for a given wind turbine or wind turbine park and W is the wind speed. The wind speed is subject to some uncertainty either due to measurement uncertainty or due to uncertainties in predicting wind speed.

Question VIII.1 (29)

If we assume that the wind speed of a given prediction has a mean value of μ_w and variance σ_w^2, then what is the approximation of the standard deviation of the output (p) based on non-linear error propagation?

1 ☐ $\gamma^2 \sqrt{\frac{e^{-2\gamma(\mu_w-\alpha)}}{\left(1+e^{-\gamma(\mu_w-\alpha)}\right)^4}\sigma_w^2}$

2* ☐ $\gamma e^{-\gamma(\mu_w-\alpha)} p(\mu_w)^2 \sigma_w$

3 ☐ $-\gamma \sqrt{\frac{e^{-2\gamma(\mu_w-\alpha)}}{\left(1+e^{-\gamma(\mu_w-\alpha)}\right)^4}\sigma_w^2}$

4 ☐ $\gamma^2 e^{-\gamma(\mu_w-\alpha)} p(\mu_w) \sigma_w$

5 ☐ $\gamma \sqrt{\frac{e^{-\gamma(\mu_w-\alpha)}}{\left(1+e^{-\gamma(\mu_w-\alpha)}\right)^2}\sigma_w^2}$

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

The derivative of the power curve is

$$p'(w) = \frac{\gamma e^{-\gamma(W-\alpha)}}{\left(1 + e^{-\gamma(W-\alpha)}\right)^2} \tag{13}$$

$$= \gamma e^{-\gamma(W-\alpha)} p(\mu_w)^2 \tag{14}$$

and as shown in method 4.3 the variance can be approximated by

$$\sigma_p^2 = \gamma^2 e^{-2\gamma(W-\alpha)} p(\mu_w)^4 \sigma_w^2 \tag{15}$$

taking the square root gives answer no 2.

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

If we assume a specific distribution for wind speed then error propagation by simulation is an alternative to the result above. One common assumption is that the wind speed follows a log-normal distribution. It is now assumed that $W \sim LN(\log(7), 0.1)$, $\gamma = 1$ and $\alpha = 5$. Further the following R code have been executed

```
set.seed(321)
k <- 10000
w <- rnorm(k, log(7), 0.1)
p <- 1/(1 + exp(-(w - 5)))
sqrt(1/(k - 1) * var(p))

## [1] 4.3063e-05

sd(p)

## [1] 0.0043061

mean(p)^2

## [1] 0.0020478

w <- rlnorm(k, log(7), 0.1)
p <- 1/(1 + exp(-(w - 5)))
sqrt(1/(k - 1) * var(p))

## [1] 0.00078496

sd(p)

## [1] 0.078492

mean(p)^2

## [1] 0.74701
```

Question VIII.2 (30)

Based on the above, what is the estimate of the standard deviation of p?

1 ☐ $4.3 \cdot 10^{-5}$

2 ☐ 0.0020

3 ☐ 0.00078

4* ☐ 0.078

5 ☐ 0.75

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

Since $W \sim LN(\log(7), 0.1)$ we should look at the second part of the code and the result is simply the standard deviation of the calculated `p`'s, i.e. 0.078 or answer no. 4.

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

The exam is finished. Have a great summer!