Written examination: 24. June 2021

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

5 points are given for a correct "multiple choice" answer, and −1 point is given for a wrong answer. ONLY the following 5 answer options are valid: 1, 2, 3, 4, or 5. If a question is left blank or an invalid answer is entered, 0 points are given for the question. Furthermore, if more than one answer option is selected for a single question, which is in fact technically possible in the online system, 0 points are given for the question. The number of points needed to obtain a specific mark or to pass the exam is ultimately determined during censoring.

The final answers should be given by filling in and submitting the form online via CampusNet. The table provided here is ONLY an emergency alternative. Remember to provide your student number if you do hand in on paper.

Exercise	I.1	I.2	II.1	II.2	II.3	III.1	III.2	IV.1	IV.2	IV.3
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer	4	1	2	4	5	4	1	4	2	3

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

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

The exam paper contains 37 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

The mean monthly electricity bill for two-bedroom apartments in a particular city is reported to be 550 DKK. A sample of 36 two-bedroom apartments is selected. The average for the 36 apartments turned out to be $\bar{x} = 562$ DKK, with a sample standard deviation of $s = 40$ DKK. The null hypothesis is $H_0 : \mu = 550$, which should be tested against the two-sided alternative.

Question I.1 (1)

Let t_{obs} be the usual test statistic for a two-sided t-test. Using a significance level α what is the rejection rule?

1 □ Reject H_0 if $t_{\text{obs}} \leq t_{1-\alpha}$, $t_{1-\alpha}$ is the $1-\alpha$ quantile in a t-distribution with 34 degrees of freedom.

2 □ Reject H_0 if $t_{\text{obs}} \leq t_{1-\alpha}$, $t_{1-\alpha}$ is the $1-\alpha$ quantile in a t-distribution with 35 degrees of freedom.

3 □ Reject H_0 if $t_{\text{obs}} \geq t_{1-\alpha}$, $t_{1-\alpha}$ is the $1-\alpha$ quantile in a t-distribution with 35 degrees of freedom.

4* □ Reject H_0 if $t_{\text{obs}} \leq t_{\alpha/2}$ or if $t_{\text{obs}} \geq t_{1-\alpha/2}$, where $t_{\alpha/2}$ and $t_{1-\alpha/2}$ are $\alpha/2$ and $1-\alpha/2$ quantiles in a t-distribution with 35 degrees of freedom, respectively.

5 □ Reject H_0 if $t_{\text{obs}} \leq t_{\alpha/2}$ or if $t_{\text{obs}} \geq t_{1-\alpha/2}$, where $t_{\alpha/2}$ and $t_{1-\alpha/2}$ are $\alpha/2$ and $1-\alpha/2$ quantiles in a t-distribution with 34 degrees of freedom, respectively.

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

See chapter three.

```
----------------------------------- FACIT-END -------------------------------------
```

Question I.2 (2)

What is the value of the test statistic (t_{obs})?

1* □ $t_{\text{obs}} = 1.80$.

2 □ $t_{\text{obs}} = 1.96$.

3 □ $t_{\text{obs}} = 2.03$.

4 □ $t_{\text{obs}} = 11.38$.

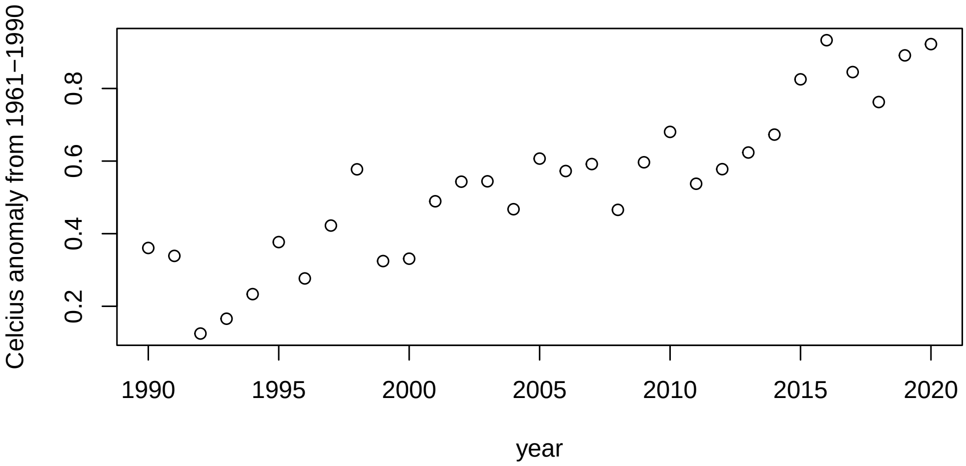
5 □ $t_{\text{obs}} = 11.22$.

```
----------------------------------- FACIT-BEGIN -----------------------------------


tobs <-(562-550)/(40/sqrt(36)) =1.8
With df =35, t.025 or 0.975 = 2.0301,
> qt(c(.025, 0.975),df=35)
[1] -2.030108  2.030108
tobs<-(562-550)/(40/sqrt(36))= 1.8; do not reject


------------------------------------ FACIT-END ------------------------------------
```

Continue on page 4

Exercise II

Historic values of the global mean atmospheric temperature can be downloaded from the UN IPCC website. The anual temperature the last 31 years was downloaded and is plotted below. The temperature is the so-called temperature abnormality, which simply is the difference from the average during the period 1961 to 1990.



The i'th observation of temperature is denoted T_i and year y_i, such that $i = 0, \ldots, 30$.

In order to analyse the development of temperature over time during the 31 years period the following model was fitted

$$T_i = \beta_0 + \beta_1 y_i + \varepsilon_i, \text{ where } \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.}$$

Continue on page 5

The obtained summary of the fit from R is:

```
##
## Call:
## lm(formula = temperature ~ year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13710 -0.07617  0.01296  0.05885  0.18984
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42.622490   3.736256  -11.41 3.07e-12 ***
## year          0.021526   0.001863   11.55 2.27e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0928 on 29 degrees of freedom
## Multiple R-squared:  0.8215,Adjusted R-squared:  0.8153
## F-statistic: 133.4 on 1 and 29 DF,  p-value: 2.271e-12
```

Question II.1 (3)

Using the formulation from evidence interpretation table in the book, what is the conclusion for the null hypothesis: There is no change in mean temperature during the period (both conclusion and argument must be correct)?

1 □ There is very strong evidence against H_0 since the p-value is $3.07 \cdot 10^{-12}$.

2* □ There is very strong evidence against H_0 since the p-value is $2.27 \cdot 10^{-12}$.

3 □ There is strong evidence against H_0 since the p-value is $3.07 \cdot 10^{-12}$.

4 □ There is strong evidence against H_0 since the p-value is $2.27 \cdot 10^{-12}$.

5 □ There is weak evidence against H_0 since the p-value is 0.0928.

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

The null hypothesis is $H_0 : \beta_1 = 0$, so we can simply read off the p-value from the summary for the slope (β_1), and find the correct formulation in Table 3.1.

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

Continue on page 6

Question II.2 (4)

What is the 99% confidence interval for the mean increase in temperature per year in the period?

1 □ [-52.9, -32.3]

2 □ [0.0184, 0.0247]

3 □ [0.421, 0.431]

4* □ [0.0164, 0.0267]

5 □ [0.323, 0.529]

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

We use the linear regression parameter confidence interval formula from Method 5.15. The 99.5% quantile in the *t*-distribution is calculated in R by

```
qt(0.995, df=31-2)

## [1] 2.756386
```

$$\hat{\beta}_1 \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_1} = 0.021526 \pm 2.756386 \cdot 0.001863$$

```
0.021526 + c(-1,1) * 2.756386 * 0.001863

## [1] 0.01639085 0.02666115
```

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

Question II.3 (5)

What is the empirical correlation coefficient between temperature and year in the period?

1 □ 0.178

2 □ 0.305

3 □ 0.695

4 □ 0.822

5* □ 0.906

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

We can use the relation between a simple linear regression model and correlation as described in Section 5.6. We can read the value of r^2 in the printed summary under `Multiple R-squared` and then, since the slope is positive, calculate the correlation by

```
sqrt(0.8215)

## [1] 0.9063664
```

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

Continue on page 8

Exercise III

The card game Casino is played with an ordinary deck of cards, i.e. 52 cards, 13 of each colour (clubs, hearts, diamonds and spade), and 4 of each value (e.g. 4 aces). The game starts by placing 4 cards face up on the table.

Question III.1 (6)

What is the probability that at least 3 of the 4 opening cards are aces?

1 □ 0.00057

2 □ $2.7 \cdot 10^{-6}$

3 □ 0.025

4* □ 0.00071

5 □ $3.7 \cdot 10^{-6}$

`------------------------------------ FACIT-BEGIN ------------------------------------`

This is the hypergeometric distribution, and we draw 4 balls fram the urn with 4 possible success, out of a total of 52 balls (or 48 non-success). Forther we need the probability $P(X \geq 3) = 1 - P(X \leq 2) = 1 - F(2)$, in R we can calculate this probability by

```
1-phyper(2,4,48,4)

## [1] 0.0007129005
```

`------------------------------------- FACIT-END -------------------------------------`

Question III.2 (7)

What is the probability that all the 4 opening cards are clubs in exactly 1 out of 4 games?

1* □ 0.010

2 □ 0.39

3 □ 0.42

4 □ 0.44

5 □ 0.015

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

The probability of all 4 opening cards being clubs in a random game is $p = 13/52 \cdot 12/51 \cdot 11/50 \cdot 10/49$, and the random variable counting the number of times all 4 card are clubs in 4 games follow a binomial with sucess probability p and $n = 4$. This is calculated by

```
p <- 13/52*12/51*11/50*10/49
p

## [1] 0.002641056

dbinom(1,size=4,prob=p)

## [1] 0.01048074
```

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

Continue on page 10

Exercise IV

COVID-19 restrictions have enforced the integration of online platforms for students' group activities. The table below shows the responses to a survey question on students' most preferred platform for group activities. The survey was conducted during three consecutive semesters (Encoded, Spring as F and Fall or Autumn as E).

	F2020	E2020	F2021	*Sum*
Teams	796	729	669	*2194*
Messenger	1017	1025	1000	*3042*
Discord	278	313	315	*906*
Slack	855	1013	1033	*2901*
Sum	*2946*	*3080*	*3017*	*9043*

Question IV.1 (8)

We want to test the null hypothesis that in Spring 2021, the proportion of students preferring slack for group activities was 34.0%. The usual large sample test statistics for the hypothesis is

1 □ $(1025 - 0.66 \cdot 3017)/\sqrt{3017 \cdot 0.34 \cdot 0.66} = -36.83$

2 □ $(1033 - 0.34 \cdot 3017)/\sqrt{3017 \cdot 0.34 \cdot 0.34} = 0.38$

3 □ $(1013 - 0.34 \cdot 3017)/\sqrt{9043 \cdot 0.34 \cdot 0.66} = 0.16$

4* □ $(1033 - 0.34 \cdot 3017)/\sqrt{3017 \cdot 0.34 \cdot 0.66} = 0.28$

5 □ $(1033 - 0.66 \cdot 3017)/\sqrt{3017 \cdot 0.34 \cdot 0.34} = -51.31$

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

Using equation (7-16) with $x = 1033$, $n = 3017$, and $p_0 = 0.34$:

```
(1033 - 0.34 * 3017) / sqrt(3017 * 0.34 * 0.66)

## [1] 0.2774841
```

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

Question IV.2 (9)

We want to test if there is a significant change in the proportion of students preferring Teams between Fall 2020 and Spring 2021.

Which line of code executes the desired test in R?

1 □ `binom.test(x = 669, n = 3017, p = 729/3080)`

2* □ `prop.test(x = c(729, 669), c(3080, 3017), correct = FALSE)`

3 □ `binom.test(x = 669, n = 3017, p = (669+729)/(3017+3080))`

4 □ `prop.test(x = 669, n = 3017, p = 729/3080, correct = FALSE)`

5 □ None of the above

`----------------------------------- FACIT-BEGIN -----------------------------------`

The task is to compare proportions between two populations (Section 7.3) (and not, e.g., to test whether a proportion has a specific value). I.e. answer 2.

`------------------------------------ FACIT-END ------------------------------------`

Question IV.3 (10)

One could also conduct a hypothesis test to investigate whether the distribution of preference of different digital platforms has changed over the three semesters for which data is given. The number of degrees of freedom in the distribution of the test statistic is:

1 □ 4

2 □ 8

3* □ 6

4 □ 10

5 □ 12

`----------------------------------- FACIT-BEGIN -----------------------------------`

Comparison of distributions in different groups (Method 7.22). With four technologies (rows) and three semesters (columns), $(r-1)\cdot(c-1) = 3\cdot 2 = 6$.

`------------------------------------ FACIT-END ------------------------------------`

Question IV.4 (11)

The semesters and the online platforms are assumed independent. The expected number of students using Messenger in Spring 2021 is estimated to be

1 □ 1000

2* □ $3017 \cdot 3042/9043 = 1015$

3 □ $(1017 + 1025 + 1000)/3 = 1014$

4 □ $(1017 + 1025)/2 = 1021$

5 □ $1000 \cdot 3017/3042 = 992$

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

As shown in the chapter 7.5.1 the expected number in a cell is calculated as:

$$\frac{\text{column total} \times \text{row total}}{\text{grand total}} = \frac{3017 \times 3042}{9043} = 1015$$

```
------------------------------------ FACIT-END ------------------------------------
```

Continue on page 13

Question IV.5 (12)

In order to test whether the change of preference of the digital platforms are significant over the three semesters, the following is given: The test statistic is calculated to be 31.316. The significance level is set to $\alpha = 0.05$. In the distribution used to assess the test statistic, the 0.95 and 0.975 quantiles are, respectively, 12.59 and 14.45. What may be concluded? (Both the conclusion and reasoning must be correct).

1 *□ The preference or users of the digital platforms over the semesters changed significantly, as the test statistic is greater than the given 0.95 quantile.

2 □ The distribution across platforms and semesters has changed significantly, as, under the null hypothesis, there is a 97.5% probability of observing a test statistic greater than 14.45.

3 □ The distribution across platforms and semesters has not changed significantly, as the test statistic is greater than the given 0.975 quantile.

4 □ The distribution across platforms and semesters has changed significantly, as, under the null hypothesis, there is a 95% probability of observing a test statistic greater than 12.59.

5 □ The numbers provided above cannot be used to argue statistically, whether the distribution across platforms and semester has changed.

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

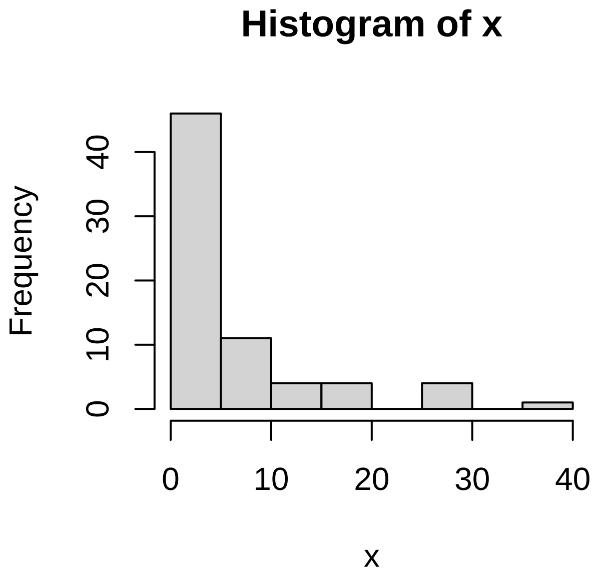
Method 7.22. The test statistic is greater than the given 0.95 quantile, so the null hypothesis of no difference over the semesters is rejected, and the change is concluded to be significant.

```
------------------------------------ FACIT-END ------------------------------------
```

Continue on page 14

Exercise V

In a study of blind use, the time (measured in hours) between interactions with the blinds are measured. The data is presented in the histogram below.



Some key-numbers are given in the R-output below (`x` is the data shown in the histogram)

```
m <- mean(x)
s <- sd(x)
n <- length(x)
c(m, s, n)

## [1]  5.996429  8.243370 70.000000
```

Question V.1 (13)

Assuming that the waiting times (time between interactions with the blinds) follow independent Exponential distributions with expectation equal the empirical average, what is the probability of observing a waiting time larger than 10?

1 □ 0.81

2 □ 0.68

3 □ 0.031

4 □ 0.97

5* □ 0.19

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

The mean is 6 and hence the parameter in the exponantial distribution is 1/6, and the probability in question is $P(X > 10) = 1 - P(X < 10) = 1 - F(10)$, which is calculated by

```
1-pexp(10,1/6)

## [1] 0.1888756
```

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

As an aid for the next questions the R-code below has been evaluated

```
k <- 10000

t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 6, df = 69, p-value = 6e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.03 7.96
## sample estimates:
## mean of x
##         6

X <- matrix(rexp(n * k, m), ncol=n)
quantile(apply(X, 1, mean),c(0.01,0.025,0.05,0.1,0.9,0.95,0.975,0.99))

##    1%  2.5%    5%   10%   90%   95% 97.5%   99%
## 0.124 0.130 0.135 0.142 0.193 0.201 0.207 0.215

X <- t(replicate(k,sample(x, replace = TRUE)))
quantile(apply(X, 1, mean),c(0.01,0.025,0.05,0.1,0.9,0.95,0.975,0.99))

##    1%  2.5%    5%   10%   90%   95% 97.5%   99%
##  3.90  4.19  4.44  4.76  7.29  7.68  8.03  8.49

X <- matrix(rexp(n * k, 1 / m), ncol = n)
quantile(apply(X, 1, mean),c(0.01,0.025,0.05,0.1,0.9,0.95,0.975,0.99))
```

```
##    1%  2.5%    5%   10%   90%   95% 97.5%   99%
##  4.43  4.64  4.84  5.09  6.92  7.22  7.49  7.78

X <- matrix(rnorm(n * k, m, s), ncol=n)
quantile(apply(X, 1, mean),c(0.01,0.025,0.05,0.1,0.9,0.95,0.975,0.99))

##    1%  2.5%    5%   10%   90%   95% 97.5%   99%
##  3.70  4.08  4.40  4.74  7.25  7.64  7.97  8.36
```

Continue on page 17

Question V.2 (14)

Again assuming the Exponential distribution, and based on the R-code above, what is the 95% confidence interval for the mean waiting time based on parametric bootstrap?

1 □ $[0.13, 0.21]$

2 □ $[4.03, 7.96]$

3* □ $[4.64, 7.49]$

4 □ $[4.19, 8.03]$

5 □ $[4.08, 7.97]$

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

As this is a paramtric bootstrap, confidence intervals are calculated by calculated by

```
X <- matrix(rexp(n * k, 1 / m), ncol = n)
quantile(apply(X, 1, mean),c(0.01,0.025,0.05,0.1,0.9,0.95,0.975,0.99))
```

and with a 95% confidence interval we need the 0.025 and 0.975 quantiles, this is answer no. 3.

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

Question V.3 (15)

What would be the 98% confidence interval if the calculations does not rely on the Exponential distribution assumption (i.e. non-parametric bootstrap)?

1 □ $[4.03, 7.96]$

2 □ $[4.43, 7.78]$

3* □ $[3.90, 8.49]$

4 □ $[0.12, 0.22]$

5 □ $[3.7, 8.36]$

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

The non-parametric bootstrap is calculated sampling directly for data, which can be done by

```
X <- t(replicate(k,sample(x, replace = TRUE)))
quantile(apply(X, 1, mean),c(0.01,0.025,0.05,0.1,0.9,0.95,0.975,0.99))
```

further for a 98% CI we need the 0.01 and 0.99 quantile, this is andswer no. 3.

`------------------------------------ FACIT-END ------------------------------------`

Continue on page 19

Question V.4 (16)

The blind use data come from two different blinds, and it is desired to test whether there is a difference in the mean waiting time for the two blinds. In the R-code below `x1`, `n1`, `m1`, `s1` denote observed waiting times, number of observations, mean waiting time and empirical standard deviation for blind one, respectively, while `x2`, `n2`, `m2`, `s2` denote the same numbers for blind two. Which of the following pieces of R-code calculate a non-parametric 95% confidence interval for the difference in mean waiting time between the two blinds?

1 □

```r
D <- matrix(rexp((n1 + n2) * k, 1 / m1 - 1 / m2), ncol = n1 + n2)
quantile(apply(D, 1, mean),c(0.025, 0.975))
```

2* □

```r
D <- t(replicate(k, sample(x1, replace = TRUE))) -
    t(replicate(k,sample(x2, replace = TRUE)))
quantile(apply(D, 1, mean),c(0.025, 0.975))
```

3 □

```r
t.test(x1,x2)
```

4 □

```r
D <- matrix(rexp(n1 * k, 1 / m1), ncol = n1) -
    matrix(rexp(n2 * k, 1 / m2), ncol = n2)
quantile(apply(X, 1, mean),c(0.05, 0.95))
```

5 □

```r
D <- t(replicate(k,sample(x1 - x2, replace = TRUE)))
quantile(apply(D, 1, mean),c(0.025, 0.975))
```

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

Non-parametris bootstrap calculate confidence intervals directly from data, this is done only in answer 2, and 4. Answer 4 is a paired test, which does not make sense here, while 2 is is the correct two sample test.

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

Continue on page 20

Exercise VI

A study was designed to investigate whether the use of recorded video lectures and group exercises affect the exam score of students. 9 students (denoted `x`) did not use or attend either of the two activities. Other 9 students (denoted `y`) used or attended both types of learning activities. The table below shows the final exam score of the students:

Students	1	2	3	4	5	6	7	8	9
x	54	54	38	47	48	62	54	55	52
y	67	70	71	60	68	65	68	67	64

Assuming that the two groups are independent or unrelated, the following code is now run in R, in order to test whether the differences in the exam scores between the groups can be assumed to be zero ($H_0 : \delta = 0$):

```
x <- c(54, 54, 38, 47, 48, 62, 54, 55, 52)
y <- c(67, 70, 71, 60, 68, 65, 68, 67, 64)
```

The output from the standard statistical analysis is given below. Please note that some numbers in the standard output have been replaced by the letters A, B and C.

```
t = -6.0836, df = A , p-value = B
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -20.537160  C
sample estimates:
mean of x mean of y
 51.55556  66.66667
```

Continue on page 21

Question VI.1 (17)

Assuming the data are normally distributed, what conclusion can be made when applying a significance level of $\alpha = 0.05$ (all elements of the answer should be correct)?

The use or non-use of learning resources and activities:

1 □ affects the exam score since the average score of x and y are different

2 □ do not show any significant effect on the exam score since the upper limit of the confidence interval is 9.685062

3 □ do not show any significant effect on the exam score since the lower limit of the confidence interval is -20.537160

4* □ have an effect on the exam score since the p-value is $6.039 \cdot 10^{-5}$

5 □ have an effect on the exam score since the p-value is $4.996 \cdot 10^{-4}$

`----------------------------------- FACIT-BEGIN -----------------------------------`

A paired t-test is designed to compare the means of the same group or item under two separate scenarios. An unpaired t-test compares the means of two independent or unrelated groups.

The standard statistical test for this setup is an unpaired Welch Two Sample t-test.The easiest way to solve this is by copying and running

```
x <- c(54, 54, 38, 47, 48, 62, 54, 55, 52)
y <- c(67, 70, 71, 60, 68, 65, 68, 67, 64)
t.test(x,y)

##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -6.0836, df = 11.725, p-value = 6.039e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -20.537160  -9.685062
## sample estimates:
## mean of x mean of y
##  51.55556  66.66667
```

From the p-value we can find the correct answer. Rejection rules are:

- $p < 0.001$ Very strong evidence against H_0

- $0.001 \geq p < 0.01$ Strong evidence against H_0
- $0.01 \geq p < 0.05$ Some evidence against H_0
- $0.05 \geq p < 0.1$ Weak evidence against H_0
- $p \geq 0.1$ Little or no evidence against H_0

```
----------------------------------- FACIT-END -----------------------------------
```

Continue on page 23

Exercise VII

In a study of the growth of trees, the circumference [mm] of the trees are measured as a function of age for a number of trees. At each age (point in time) the circumference is measured for all trees. As an initial analysis it is tested if there is a difference in circumference as a function of age using a one-way anova analasis. The result is seen in the R-output below (where some numbers have been replaced by letters).

```
> anova(lm(circumference ~ age))
Analysis of Variance Table

Response: circumference
          Df Sum Sq Mean Sq F value    Pr(>F)
age        6  A     16008.4       D  1.661e-10 ***
Residuals 28  B           C
```

Question VII.1 (18)

What is the total number of observations in the study?

1 □ 28

2 □ 33

3* □ 35

4 □ 29

5 □ 34

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

The number of observations is 28+6+1=35.

```
------------------------------------ FACIT-END ------------------------------------
```

Continue on page 24

The usual model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \tag{1}$$

where ϵ_{ij} are iid. random variables.

Question VII.2 (19)

What is the usual estimate of σ?

1 □ 126.5

2 □ 27.4

3 □ 51.7

4* □ 24.1

5 □ 457.4

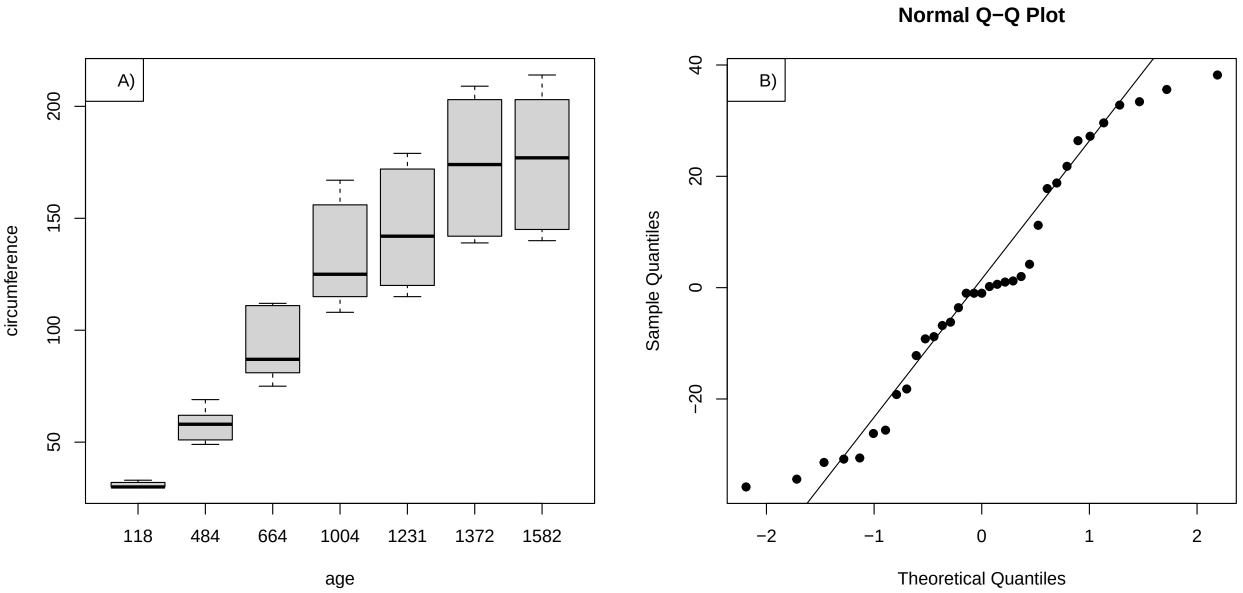
----------------------------------- FACIT-BEGIN -----------------------------------

The number we are looking for is $\sqrt{C}$, which is given by $D = 16008.4/C$ or $C = 16008.4/D$, the values of D can be found as the inverse of the F-distribution with 6 and 28 degrees of freedom at $1 - 1.661 \cdot 10^{-10}$, hence

```
(D <- qf(1-1.661e-10,df1=6,df2=28))

## [1] 27.47238

(C <- 16008.4 / D)

## [1] 582.7089

sqrt(C)

## [1] 24.13936
```

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

As part of model control the figure below, showing a box-plot of the circumference as a function of age and a qq-normal plot of the residuals from the model, is created.



Continue on page 26

Question VII.3 (20)

Which of the following statements is most appropriate based on the figure above (including reference to plot)?

1 □ The independence assumption is clearly violated (plot B).

2 □ All the usual assumptions seems to be fulfilled (based on both plots).

3 □ The normality assumption on the residuals is clearly violated (plot A).

4* □ The assumption on variance-homogeneity is clearly violated (plot A).

5 □ The independence assumption is clearly fulfilled (plot B).

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

None of the plots can be used to assess independence and hence 1 and 5 are not correct. Plot A does not concern the normalty assumption, i.e. 3 is not correct. From plot A it is clear that the wariance is increasing with age and hence the variance-homogeity is violated so 4 is correct and then 2 is not correct.

```
------------------------------------ FACIT-END ------------------------------------
```

Regardless of the answer to the previous question, it is decided to test if there is a significant difference between trees using the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \tag{2}$$

where α_i denote the effect of time and β_j denote the effect of tree. The residual variance is estimated at $\hat{\sigma}^2 = 186.4$, also note that all trees are measured once at every age (i.e. 5 trees).

Question VII.4 (21)

On significance level $\alpha = 0.05$ and based on the 2-way anova analysis, which of the following statements is correct (all parts of the answer should be correct)?

1 □ There is not a significant effect of age since $15.9 > 4.54$

2* □ There is a significant effect of age since $85.9 > 2.51$

3 □ There is a significant effect of age since $1.66 \cdot 10^{-10} < 0.05$

4 □ There is not a significant effect of age since $1.66 \cdot 10^{-6} < 0.05$

5 □ There is a significant effect of age since $15.9 > 2.78$

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

2 answers (3, and 4) are based on p-values, but the p-values is the one given in the 1-way analysis and that will change, so 3 and 4 are not correct. To find the correct one we need to fill in part of the 2-way anova table. First note that

$$\begin{aligned} df_{tree} =& 4 \tag{3}\\ df_{age} =& 6 \tag{4}\\ df_{res} =& df_{tree} df_{age} \tag{5} \end{aligned}$$

In order to find the correct answer we need $F_{age} = \frac{MS(age)}{MSE}$, $MS(age) = 16008.4$ is given in the 1-way anova table, and $MSE = \hat{\sigma}^2 = 186.4$, and we can find the test statistics and critical value as

```
df.tree <- 4
df.age <- 6
df.res <- df.tree * df.age
16008.4/186.4

## [1] 85.88197

qf(0.95,df.age,df.res)

## [1] 2.508189
```

which is answer no 2.

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

Continue on page 28

Question VII.5 (22)

What is the 95%, Bonferoni corrected, Least Significant Difference (LSD) for pairwise comparison of all age-categories?

1 □ 9.46

2 □ 49.2

3 □ 24.4

4 □ 10.90

5* □ 29.3

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

The number of pairwise comparisons is $7 \cdot 6/2 = 21$, and the corrected α-level is $\tilde{\alpha} = \frac{\alpha}{21}$, and the critical value is then $t_{1-\tilde{\alpha}/2}$, using 24 degress of freedom. Further $(MSE = \hat{\sigma}^2)$

$$LSD = t_{1-\tilde{\alpha}/2}\sqrt{MSE\left(\frac{1}{n}+\frac{1}{n}\right)} \tag{6}$$

with $n = 5$ (number of trees). This is calculated by

```
(alpha <- 0.05/(7*6/2))

## [1] 0.002380952

(cv <- qt(1-alpha/2,df=24))

## [1] 3.395988

(LSD <- cv * sqrt(2 * 186.4/5))

## [1] 29.32372
```

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

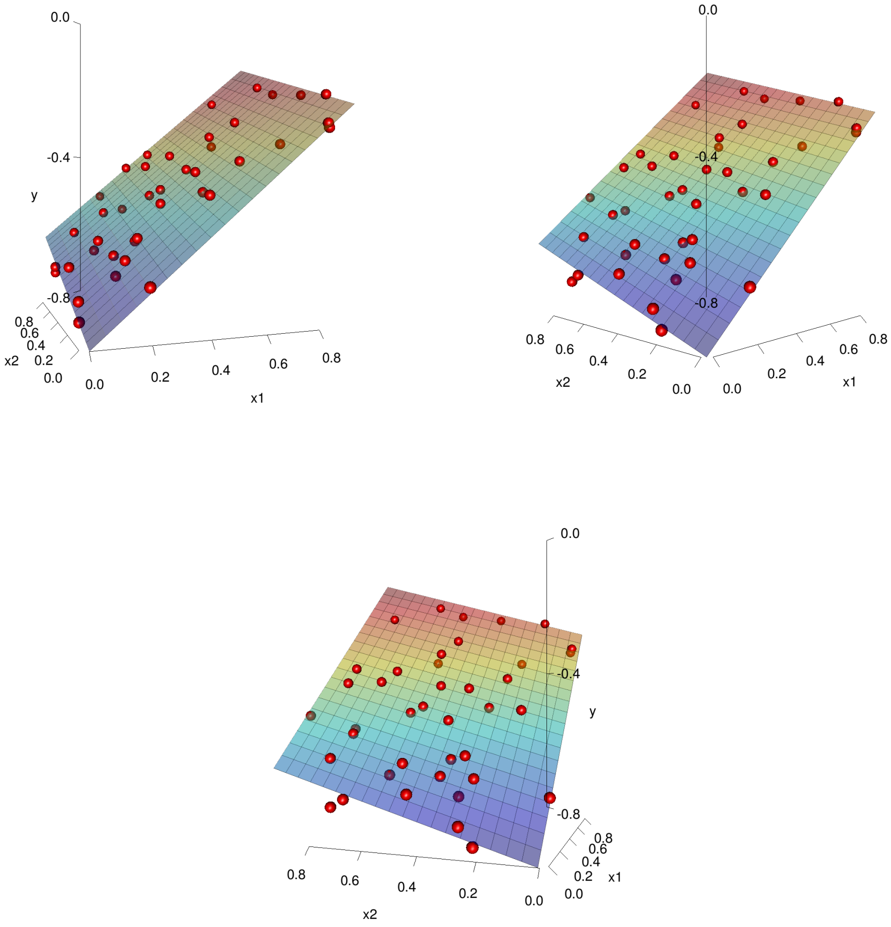
Continue on page 29

Exercise VIII

The multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i \text{ where } \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.}$$

is fitted to data with $i = 1, 2, \ldots, n$ and the result is visualized below. The three plots below show the points and the estimated plane, using different rotations, such that it can be seen from different angles. Note, that the plane is coloured red for high and blue for low y values and remember to check the axis labels for the direction of the plot axis:



Continue on page 30

Question VIII.1 (23)

What is the correct ordering of the estimated coefficients?

1 □ $\hat{\beta}_0 < \hat{\beta}_1 < \hat{\beta}_2$

2* □ $\hat{\beta}_0 < \hat{\beta}_2 < \hat{\beta}_1$

3 □ $\hat{\beta}_1 < \hat{\beta}_0 < \hat{\beta}_2$

4 □ $\hat{\beta}_1 < \hat{\beta}_2 < \hat{\beta}_0$

5 □ $\hat{\beta}_2 < \hat{\beta}_1 < \hat{\beta}_0$

`----------------------------------- FACIT-BEGIN -----------------------------------`

We can see the estimated slopes (β_1 and β_2) on the slope of the estimated plane in the `x1` and `x2` dimensions. They are both positive and it's clear that for β_1 the slope something like 1 and for β_2 is something like 0.3.

The intercept (β_0) is clearly negative and around -1.

`------------------------------------ FACIT-END ------------------------------------`

Question VIII.2 (24)

From the plots it is clear that the estimated standard deviation of the error, i.e. $\hat{\sigma}$, can only be one of the following values, which one?

1* □ $\hat{\sigma} = 0.03$

2 □ $\hat{\sigma} = 2.1$

3 □ $\hat{\sigma} = 4.7$

4 □ $\hat{\sigma} = 357$

5 □ $\hat{\sigma} = 10028$

`----------------------------------- FACIT-BEGIN -----------------------------------`

We see in the model that $\varepsilon \sim N(0, \sigma)$, so σ is the standard deviation of the errors, so the estimate of it is the average distance. We look at the distances between the points and the plane in the `y` dimension and see that the points are "close" to the plane (i.e. residuals).

If $\hat{\sigma} = 2.1$ then the points would be much further away from the plance and even the range of the y-axis would have been much higher, so we can deduce that it can only be the answer $\hat{\sigma} = 0.03$. points would be much further

`------------------------------------ FACIT-END ------------------------------------`

The multiple linear regression problem can be written in matrix-vector notation as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]^T$.

Continue on page 32

Question VIII.3 (25)

In the R-code below `X` denote $\boldsymbol{X}$, `beta` denote $\hat{\boldsymbol{\beta}}$, `sigma` denote $\hat{\sigma}$, and `n` denote the number of observations. Which of the following pieces of R-code calculate the upper limit in a 95% prediction interval for a new observation at $x_1 = x_2 = 0$?

1 □

```
x <- matrix(c(0, 0, 0),ncol=3)
x %*% beta + sigma * qt(0.975, df = n - 3) *
   sqrt(1 + x %*% solve(t(X) %*% X) %*% t(x))
```

2* □

```
x <- matrix(c(1, 0, 0),ncol=3)
x %*% beta + sigma * qt(0.975, df = n - 3) *
   sqrt(1 + x %*% solve(t(X) %*% X) %*% t(x))
```

3 □

```
x <- matrix(c(1, 0, 0),ncol=3)
X %*% beta + sigma * qt(0.975, df = n) *
   sqrt(x %*% solve(t(X) %*% X) %*% t(x))
```

4 □

```
x <- matrix(c(0, 0, 0),ncol=3)
x %*% beta + sigma / sqrt(n) * qt(0.975, df = n - 3) *
   sqrt(1 + x %*% solve(t(X) %*% X) %*% t(x))
```

5 □

```
x <- matrix(c(1, 0, 0),ncol=3)
x %*% beta + sigma / sqrt(n) * qt(0.975, df = n) *
   sqrt(1 + x %*% solve(t(X) %*% X) %*% t(x))
```

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

With $x_1 = x_2 = 0$, the CI apply for $\boldsymbol{x} = [1, 0, 0]$, the leaves answer 2,3, and 5. The critical value should be taken frrom a t-distribution with $n - 3$ degrees of freedom, this exclude 3, and 5. Hence 2 must be correct.

Lets just state the correct answer. The confidence interval is given by

$$\boldsymbol{x}\boldsymbol{\beta} \pm \hat{\sigma} t_{1-\alpha/2}\sqrt{1 + \boldsymbol{x}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^T} \tag{7}$$

which is answer no. 2.

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

Continue on page 33

Exercise IX

When modeling probabilities the probit transformation is often used. For a specific problem for modeling the probability of opening windows, the following model is proposed

$$p(U) = \Phi\left(\beta_0 + \beta_1 \cos\left(\frac{2\pi h}{24} + U + \psi\right)\right) \tag{8}$$

where Φ is the CDF for standard normal random variable, h is time of day (measured in hours), and β_0, β_1, and ψ are fixed parameters. Further U is a uniform random variable on the interval $[0, 2\pi]$.

Question IX.1 (26)

What is the error propagation approximation of the mean and variance of p (ϕ is the probability density function (pdf) for a standard normal random variable)?

1 □ $\mathrm{E}[p] \approx p(\pi)$, and $\mathrm{V}[p] \approx \Phi^2\left(\beta_0 + \beta_1 \cos\left(\frac{2\pi h}{24} + \pi + \psi\right)\right) \beta_1^2 \sin^2\left(\frac{2\pi h}{24} + \pi + \psi\right) \frac{\pi^2}{12}$

2 □ $\mathrm{E}[p] \approx \Phi(\pi)$, and $\mathrm{V}[p] \approx \phi^2\left(\beta_0 + \beta_1 \cos\left(\frac{2\pi h}{24} + \pi + \psi\right)\right) \beta_1^2 \sin^2\left(\frac{2\pi h}{24} + \pi + \psi\right) \frac{\pi^2}{12}$

3 □ $\mathrm{E}[p] \approx 0$, and $\mathrm{V}[p] \approx \phi^2(\pi) \frac{\pi^2}{3}$

4 □ $\mathrm{E}[p] \approx \Phi(\pi)$, and $\mathrm{V}[p] \approx \Phi^2\left(\beta_0 + \beta_1 \cos\left(\frac{2\pi h}{24} + \pi + \psi\right)\right) \beta_1 \sin\left(\frac{2\pi h}{24} + \pi + \psi\right) \frac{\pi^2}{3}$

5* □ $\mathrm{E}[p] \approx p(\pi)$, and $\mathrm{V}[p] \approx \phi^2\left(\beta_0 + \beta_1 \cos\left(\frac{2\pi h}{24} + \pi + \psi\right)\right) \beta_1^2 \sin^2\left(\frac{2\pi h}{24} + \pi + \psi\right) \frac{\pi^2}{3}$

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

We need the derivative of $p(u)$ which is given by

$$p'(U) = \phi\left(\beta_0 + \beta_1 cos\left(\frac{2\pi h}{24} + U + \psi\right)\right) \beta_1 sin\left(\frac{2\pi h}{24} + U + \psi\right) \tag{9}$$

also from the distribution of U we have $E[U] = \pi$, and $\mathrm{V}[U] = \frac{\pi^2}{3}$. Hence

$$\begin{aligned} \mathrm{E}[p] &\approx p(\mathrm{E}[U]) = p(\pi) && (10)\\ \mathrm{V}[p] &\approx (p'(\mathrm{E}[U]))^2 V[U] && (11)\\ &= \phi^2\left(\beta_0 + \beta_1 \cos\left(\frac{2\pi h}{24} + \pi + \psi\right)\right) \beta_1^2 \sin^2\left(\frac{2\pi h}{24} + \pi + \psi\right) \frac{\pi^2}{3} && (12) \end{aligned}$$

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

Continue on page 34

Question IX.2 (27)

With $h = 12$, $\beta_0 = \psi = 0$, and $\beta_1 = 1$, which R-code approximate the mean and variance of p?

1 □

```
u <- runif(10000)
c(mean(pnorm(cos(pi + u))), var(pnorm(cos(pi + u))))
```

2 □

```
u <- rnorm(10000, 0, 2 * pi)
c(mean(dnorm(cos(pi + u))), var(dnorm(cos(pi + u))))
```

3* □

```
u <- runif(10000, 0, 2 * pi)
c(mean(pnorm(cos(pi + u))), var(pnorm(cos(pi + u))))
```

4 □

```
u <- runif(10000, 0, 2 * pi)
c(mean(rnorm(cos(pi + u))), var(rnorm(cos(pi + u))))
```

5 □

```
u <- rnorm(10000)
c(mean(pnorm(cos(pi + u), 0, 2 * pi)), var(pnorm(cos(pi + u), 0, 2 * pi)))
```

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

With $h = 12$, $\beta_0 = \psi = 0$, and $\beta_1 = 1$, we have

$$p(U) = \Phi\left(cos\left(\pi + U\right)\right) \tag{13}$$

where U should be sampled from a uniform on $[0, 2\pi]$. 1, and 5 use the wrong distribution of U. 2 calculate the density rather than the distribution functions and 4 sample for a normal distribution rather than finding the distribution function, and 3 sample from the right distribution and calculate the distribution function.

```
------------------------------------ FACIT-END ------------------------------------
```

Continue on page 35

Exercise X

The A-series of paper is defined by, long edge = $\sqrt{2}k$ times the short edge. A machine is cutting an A-series of paper. Assume that the accuracy of the machine can be expressed as

$$X \sim N(k, \sigma^2)$$
$$Y \sim N(\sqrt{2}k, \sigma^2)$$

where X is the short edge and Y is the long edge, it can further be assumed the X and Y are independent.

Question X.1 (28)

With X and Y as defined above, what is $E[X^2 + Y^2]$?

1 □ $3k^2$

2 □ $3k^2 + 2k\sigma^2$

3 □ $3k^2 + k\sigma^2$

4* □ $3k^2 + 2\sigma^2$

5 □ $3k^2 + 2k^2\sigma^2$

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

The expectation can be written as

$$E[X^2+Y^2] = E\left[\sigma^2\left(\frac{(X-k)^2 - k^2 + 2kX}{\sigma^2} + \frac{(Y-\sqrt{2}k)^2 - 2k^2 + 2\sqrt{2}kY}{\sigma^2}\right)\right] \tag{14}$$

$$= \sigma^2\left(E\left[\frac{(X-k)^2}{\sigma^2}\right] + E\left[\frac{(Y-\sqrt{2}k)^2}{\sigma^2}\right]\right) - k^2 + 2kE[X] - 2k^2 + 2\sqrt{2}kE[Y] \tag{15}$$

$\frac{(X-k)^2}{\sigma^2}$ and $\frac{(Y-\sqrt{2}k)^2}{\sigma^2}$ both follow χ^2 distributions with 1 degrees of fredom and hence both have expectation equal 1. Hence the result can be written as

$$= 2\sigma^2 - 3k^2 + 2k^2 + 4k^2 \tag{16}$$
$$= 3k^2 + 2\sigma^2 \tag{17}$$

which is answer 4.

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

Question X.2 (29)

Again with X and Y as defined above, what is $P\left(\frac{(X-k)^2}{(Y-\sqrt{2}k)^2} < 2\right)$?

1 □ 0.39

2 □ 0.50

3 □ 0.55

4 □ 0.67

5* □ 0.61

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

$\frac{(X-k)^2}{(Y-\sqrt{2}k)^2} = \frac{(X-k)^2/\sigma^2}{(Y-\sqrt{2}k)^2/\sigma^2} \sim F(1,1)$, and hence the probability can be calculated as

```
pf(2,1,1)

## [1] 0.6081734
```

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

Continue on page 37

Question X.3 (30)

What is $P\left(\frac{X-k}{|Y-\sqrt{2}k|} < -1\right)$? (the answer should apply for all choices of σ and k).

1 □ 0.33

2 □ $P(T < -\sqrt{2}k)$, where $T \sim t(1)$

3 □ $P(T < -\sigma^2)$, where $T \sim t(1)$

4* □ 0.25

5 □ $P(T < -\frac{1}{\sigma^2})$, where $T \sim t(1)$

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

$\frac{X-k}{|Y-\sqrt{2}k|} = \frac{(X-k)/\sigma}{\sqrt{(Y-\sqrt{2}k)^2/\sigma^2}} \sim t(1)$, hence the probability can be calculated as

```
pt(-1,df=1)

## [1] 0.25
```

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

The exam is finished. Enjoy the summer!