Written examination: 21 June 2018

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

Aids and facilities allowed: All

The questions were answered by

____________________ ____________________ __________

(student number) (signature) (table number)

There are 30 questions of the "multiple choice" type included in this exam divided on 15 exercises. To answer the questions you need to fill in the prepared 30-question multiple choice form (on 6 seperate pages) in CampusNet.

5 points are given for a correct 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 another answer is given, then it does not count (i.e. "0 points"). Hence, if more than one answer option is given to a single question, which in fact is technically possible in the online system, it will not count (i.e. "0 points"). The number of points corresponding to specific marks or needed to pass the examination is ultimately determined during censoring.

The final answers should be given in the exam module in CampusNet. The table sheet here is ONLY to be used as an "emergency" alternative (remember to provide your study number if you hand in the sheet).

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

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

Exercise	XI.3	XI.4	XI.5	XI.6	XII.1	XIII.1	XIII.2	XIV.1	XV.1	XV.2
Question	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Answer	4	5	4	4	4	4	3	4	2	3

The questionnaire contains 37 pages.

Continues on page 2

Multiple choice questions: *Note that not all the suggested answers are necessarily meaningful. In fact, some of them are very wrong but under all circumstances there is one and only one correct answer to each question.*

Exercise I

A dice is thrown on a garden table with a certain distance between the planks. The probability that the dice ends on an edge between two planks is 0.2. If the dice does not end up between two planks, it has equal probability of landing on any of the 6 sides. If the dice ends on the edge between 2 planks, then it is given the value 0.

Question I.1 (1)

What is the expected value of the distribution corresponding to a dice roll?

1 □ 2.5

2* □ 2.8

3 □ 3

4 □ 3.3

5 □ 3.5

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

The probability that the die lands on each side is 0.8/6, so we use Definition 2.13 and get that the mean is

$$0 \cdot 0.8 + 0.8 \cdot \sum_{i=1}^{6} i/6 = 0.8 \cdot 3.5 = 2.8$$

which is answer no. 2

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

Continues on page 3

Exercise II

10 individuals are divided into two groups of 5 individuals in each. Each individual has had a cell count, as part of a medical examination. The numbers are given in the table below

Group 1	351	276	94	293	364
Group 2	494	403	159	329	492

It can be assumed that the variances in the two groups are the same, and that the data originates from normal distributions.

Question II.1 (2)

Using the note's definition of quantiles, what is the 1st, 2nd and 3rd quartiles of the data in group 1?

1 □ (276, 94, 293)

2 □ (276, 351, 351)

3* □ (276, 293, 351)

4 □ (185, 293, 357.5)

5 □ (276, 351, 364)

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

```
quantile(c(351,276,94,293,364),type=2,prob=c(0.25,0.5,0.75))

## 25% 50% 75%
## 276 293 351
```

or with $n = 5$, we have $n/4 = 1.25$, $2n/4 = 2.5$, and $3n/4 = 3.75$, since these are all non-integers we get $Q_1 = x_{(2)}$, $Q_2 = x_{(3)}$, and $Q_3 = x_{(4)}$, and the sorted observations are $(94, 276, 293, 351, 364)$, and hence the answer is no. 3.

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

Question II.2 (3)

If we assume that the two groups are independent, what is then the p-value of a standard t-test for whether there is a difference in the counts between the two groups?

1 □ 0.44

2 □ 0.34

3 □ 0.14

4 □ 0.04

5* □ 0.24

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

Since the individuals in the two groups are different, we should use the standard two sample t-test, in R we find the answer by.

```
gr1 <- c(351, 276, 94, 293, 364)
gr2 <- c(494, 403, 159, 329, 492)
t.test(gr1,gr2,var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  gr1 and gr2
## t = -1.2665, df = 8, p-value = 0.241
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -281.51137   81.91137
## sample estimates:
## mean of x mean of y
##     275.6     375.4
```

The p-value can be read directly from the output as 0.24, which is answer no. 5.

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

Continues on page 5

Question II.3 (4)

If the two groups instead refer to two different methods for counting cells applied to 5 individuals, such that columns refer to individuals, what would the p-value for the test of a difference between the two groups, then be?

1 □ 0.299

2 □ 0.049

3 □ 0.199

4 □ 0.099

5* □ 0.009

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

In this case we are in a paired situation, i.e.

```
  gr1 <- c(351, 276, 94, 293, 364)
  gr2 <- c(494, 403, 159, 329, 492)
  t.test(gr1, gr2, paired = TRUE)

##
##  Paired t-test
##
## data:  gr1 and gr2
## t = -4.7898, df = 4, p-value = 0.008713
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -157.65008  -41.94992
## sample estimates:
## mean of the differences
##                   -99.8
```

and the p-value is $0.0087 \approx 0.009$ which is answer no. 5

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

Continues on page 6

Exercise III

One of the focus areas in the recent collective bargain (OK18) was the salary. The 98 municipalities in Denmark are grouped in five regions: Hovedstaden, Midtjylland, Nordjylland, Sjælland og Syddanmark. In the following the salary of teachers (in 1.000 kr) is analysed.

Below two analysis in R of the data are shown (`lldr` contains data for the salary of teachers as described above):

```
lm.lldr <- lm(loen ~ Region, lldr)
summary(lm.lldr)

##
## Call:
## lm(formula = loen ~ Region, data = lldr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.14456 -0.41405 -0.03269  0.49197  1.98813
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         42.6325     0.1530 278.603  < 2e-16 ***
## RegionMidtjylland  -0.6566     0.2432  -2.700  0.00825 **
## RegionNordjylland  -0.5423     0.2918  -1.859  0.06625 .
## RegionSjælland     -0.4222     0.2517  -1.677  0.09685 .
## RegionSyddanmark   -0.5352     0.2330  -2.297  0.02386 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.824 on 93 degrees of freedom
## Multiple R-squared:  0.0932,Adjusted R-squared:  0.0542
## F-statistic:  2.39 on 4 and 93 DF,  p-value: 0.0564

anova(lm.lldr)

## Analysis of Variance Table
##
## Response: loen
##           Df Sum Sq Mean Sq F value Pr(>F)
## Region     4  6.491 1.62270  2.3896 0.0564 .
## Residuals 93 63.152 0.67906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Continues on page 7

Question III.1 (5)

The estimated model is typically categorized as a

1 □ Simple linear regression model

2 □ Two way analysis of variance

3 □ T-test

4* □ One way analysis of variance

5 □ None of the above

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

A one way analysis of variance is the correct answer as Region is a factor.

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

Question III.2 (6)

Using the usual significance level ($\alpha = 0.05$), then the conclusion of the analysis is that

1 □ there are significant differrences between the regions as $2.39 > 1.96$

2* □ there are no significant differences between the regions as $0.0564 > 0.05$

3 □ there are very significant differences between the regions as $0.00825 < 0.05$

4 □ there are significant differences between the regions as $0.0564 > 0.05$

5 □ data is missing for one municipality as $93 + 4 \neq 98$.

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

We need the p-value, which is given in the anova table as 0.056 and hence there are no significant differences between the regions as $0.0564 > 0.05$

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

Continues on page 8

Exercise IV

The table below contains average salaries for persons with a medium-cycle higher education for a period of 6 years and grouped in the five regions.

Region	2011	2012	2013	2014	2015	2016
Hovedstaden	31.36	31.93	32.49	33.02	34.77	34.50
Sjælland	31.68	32.11	32.16	32.87	33.28	33.62
Syddanmark	30.84	31.36	31.40	31.91	32.57	32.76
Midtjylland	29.87	30.12	30.62	30.74	31.52	32.00
Nordjylland	29.28	29.63	29.82	30.10	30.42	30.65

The result of the relevant analysis is:

```
            Df Sum Sq Mean Sq F value   Pr(>F)
Region       4   3778   944.6       A       C
year         5   1734   346.8       B       D
Residuals   20    242    12.1
```

(Some number are left out and replaced by “A”, “B”, “C”, and “D”).

Question IV.1 (7)

The relevant test statistic to investigate the effect of year is

1 □ 1737 / 346.8 = 5.009

2* □ 346.8 / 12.1 = 28.661

3 □ 5 / 1734 = 0.003

4 □ 944.6 / 346.8 = 2.724

5 □ 1734 / 242 = 7.165

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

The test statistic to investigate the effect of year is

$$F_{year} = \frac{MS(Year)}{MSE} = 346.8/12.1 = 28.661$$

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

Continues on page 9

Question IV.2 (8)

The test statistic for the difference between regions is to be compared to a

1 □ χ^2 distribution with 20 degrees of freedom

2 □ t-distribution with 4 degrees of freedom

3 □ F(4, 5) distribution

4 □ χ^2 distribution with 4 degrees of freedom

5* □ F(4, 20) distribution

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

Since this is an anova table we will need a F-distribution, with $n = 25$ observations and $k = 5$ estimated efftects the number of degrees of freedom become $df_1 = k-1 = 4$ and $df_2 = n-k = 20$, i.e a F(4, 20) distribution (or answer no. 5).

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

Continues on page 10

Exercise V

Suppose that a one-sample test has been made with 15 observations.

Question V.1 (9)

What is the critical value of a standard t-test at significance level $\alpha = 0.01$ for whether the underlying distribution has mean value 0 against the alternative that it is positive?

1 □ 2.98=qt(0.995,df=14)

2* □ 2.62=qt(0.99,df=14)

3 □ 1.75=qt(0.95,df=15)

4 □ 2.60=qt(0.99,df=15)

5 □ 2.95=qt(0.995,df=15)

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

The alternative here is $H_1 : \mu > 0$, and hence this is a one-sided t-test, with the p-value given by (see eq. (3-38))

$$P(T > t_{obs})$$

hence the critical value for an level $\alpha = 0.01$ is $t_{1-\alpha} = t_{0.99}$, where the degrees of freedom is $n - 1 = 14$

```
qt(0.99,df=14)

## [1] 2.624494
```

which is answer no. 2.

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

Continues on page 11

Exercise VI

In an indoor climate study, conditions in buildings are investigated. There is a special focus on the experience that the users of the buildings have and therefore a number of questionnaire surveys are conducted. In the following statistics on the survey results are carried out.

Question VI.1 (10)

In an office building where a renovation has been carried out, a survey before and a survey after the renovation was carried out. One of the questions were: “Have you felt dry in your throat while being in the office during the last week?”

	Before	After
All the time	34	22
Part of the time	39	43
Not at all	42	34

The following R code has been run:

```
prop.test(x=c(34,22), n=c(34+39+42,22+43+34), correct=FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(34, 22) out of c(34 + 39 + 42, 22 + 43 + 34)
## X-squared = 1.4847, df = 1, p-value = 0.223
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.04345776  0.19031766
## sample estimates:
##    prop 1    prop 2
## 0.2956522 0.2222222

chisq.test(matrix(c(34, 22, 39, 43, 42, 34), ncol = 2, byrow = TRUE),
           correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  matrix(c(34, 22, 39, 43, 42, 34), ncol = 2, byrow = TRUE)
## X-squared = 2.426, df = 2, p-value = 0.2973
```

Continues on page 12

It is now decided to investigate whether the entire distribution of answers to the question is different before renovation and after renovation. What is the conclusion at significance level $\alpha = 0.05$ (both conclusion and argument must be correct)?

1 □ The null hypothesis that there is no difference between the distribution before and after is accepted, since the p-value for the relevant test is 0.223

2 □ The null hypothesis that there is no difference between the distribution before and after is rejected, since the p-value for the relevant test is 0.223

3* □ The null hypothesis that there is no difference between the distribution before and after is accepted, since the p-value for the relevant test is 0.2973

4 □ The null hypothesis that there is no difference between the distribution before and after is rejected, since the p-value for the relevant test is 0.2973

5 □ None of the above answers are correct.

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

Since it is the entirre distribution that is tested here, we will need the result from `chisq.test`, the p-value for the test of equal distributions is in this case 0.2973 which is greater the the decided significance level, and hence we accept the null hypothesis, that there is no difference between the distribution before and after. I.e. answer no 3 is correct.

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

Question VI.2 (11)

In a multi-employer office, an indoor climate investigation is desired at the individual office locations relative to each other. Therefore, employees are continuously asked over time, via an app, if they think the indoor climate is good. E.g. they get the question: “Has it been too hot in your office space today?”. During a week, the following answers to this question were collected for the office spaces near the window:

Yes	28
No	17

Which of the following answers is the correctly calculated 99% confidence interval for the proportion of employees at the windows seats that have been too hot?

1 □ $\frac{28}{45} \pm 1.96\sqrt{\frac{476}{2025}}$

2* □ $\frac{28}{45} \pm 2.58\sqrt{\frac{476}{91125}}$

3 □ $\frac{28}{45} \pm 2.58\sqrt{\frac{476}{2025}}$

4 □ $\frac{28}{45} \pm 1.96\sqrt{\frac{476}{91125}}$

5 □ $\frac{28}{45} \pm 2.58\sqrt{\frac{784}{2025}}$

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

The confidence interval is in this case given by

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

in our case we have $n = 45$, $\hat{p} = \frac{28}{45}$ and we can calculate $z_{1-\alpha/2} = z_{0.995}$ by

```
qnorm(0.995)

## [1] 2.575829
```

hence the correct answer is

$$\frac{28}{45} \pm 2.58\sqrt{\frac{\frac{28}{45}\frac{17}{45}}{45}} = \frac{28}{45} \pm 2.58\sqrt{\frac{28 \cdot 17}{45^3}}$$

with

```
 28*17

## [1] 476

 45^3

## [1] 91125
```

we get answer no. 2.

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

Continues on page 14

Question VI.3 (12)

The survey from the previous question is run continuously and an expected width of the confidence interval of 20% is required at the level of significance $\alpha = 0.05$.

The minimum number of responses to be collected depends on the scenario (ie, the assumed probability, p, for positive response). How many fewer answers should be collected in a scenario, where it is assumed that $p = 0.25$ relative to the most conservative scenario, ie. assuming $p = 0.5$?

1 □ 6 fewer answers

2* □ 24 fewer answers

3 □ 35 fewer answers

4 □ 73 fewer answers

5 □ 100 fewer answers

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

The general formula to use is

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

if the expected confidence interval width should be 20% the margin of error should be $ME = 0.2/2 = 0.1$, and hence whith an assumption of $p = 0.25$ we would need

```
0.25*0.75 * (qnorm(0.975)/0.1)^2

## [1] 72.02735
```

and with an assumption of $p = 0.5$ we would need

```
0.5^2 * (qnorm(0.975)/0.1)^2

## [1] 96.03647
```

hence the difference is $96 - 72 = 24$ observations, this is answer no. 2.

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

Continues on page 15

Exercise VII

In an investigation, it is wanted to clarify whether there is a correlation between customers' perception of how green their shopping habits are, relative to their actual purchases, based on the amount of organic groceries they buy. Customers have via a supermarket app answered the question: "I think a lot about the environment when I buy groceries" and the supermarket has registered customers' purchases and divided them into 3 categories according to how many organic products they actually bought.

Organic	*I think a lot about the environment when I buy groceries*		
products	Agree	Neither or	Disagree
Many	21.5%	6.8%	5.1%
Average	17.5%	13%	6.8%
Few	9.6%	11.9%	7.9%

A total of 177 responses was collected from customers (the cells do not sum up to exactly 100% due to regular rounding).

Question VII.1 (13)

Under the null hypothesis of independence in the customer's perception and action, what is the expected number of observations for customers, with organic purchases "Many" and which indicates "Agree" that they think about the environment when they buy groceries (ie, the upper left cell)?

1 □ $(0.068 + 0.13 + 0.119) \cdot 177 \cdot (0.175 + 0.13 + 0.068) \cdot 177/177 = 20.9$

2* □ $(0.215 + 0.068 + 0.051) \cdot 177 \cdot (0.215 + 0.175 + 0.096) \cdot 177/177 = 28.7$

3 □ $(0.215) \cdot 177 \cdot (0.175 + 0.13 + 0.068) \cdot 177/177 = 14.2$

4 □ $(0.215 + 0.068 + 0.051) \cdot 177 = 59.1$

5 □ $(0.175 + 0.13 + 0.068) \cdot 177 = 66$

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

Under the null hypothesis the expected number of observation in the upper left cell is

$$\frac{n_{many} \cdot n_{agree}}{n}$$

in our case the number can be calculated by (with $n = 177$)

$$n_{many} = (0.215 + 0.068 + 0.051) \cdot n n_{many} = (0.215 + 0.175 + 0.096) \cdot n$$

I.e. answer no 2, with the numerical value

```
(0.215+0.068+0.051)*177 * (0.215+0.175+0.096)*177 / 177

## [1] 28.7
```

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

Question VII.2 (14)

The number of degrees of freedom (df) and the critical value (q_{crit}) for the relevant test at significance level $\alpha = 0.05$ becomes:

1 □ $df = 177$ and $q_{\text{crit}} = 1.96$

2 □ $df = 168$ and $q_{\text{crit}} = 3.84$

3 □ $df = 2$ and $q_{\text{crit}} = 5.99$

4* □ $df = 4$ and $q_{\text{crit}} = 9.49$

5 □ $df = 3$ and $q_{\text{crit}} = 7.82$

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

The critical value should be calculated form a χ^2 distribution with $(r-1)(c-1)$, where r and c is the number af rows and columns respectively in the table, i.e. $df = 2 \cdot 2 = 4$, and the critical value can be calculated by

```
qchisq(0.95,df=4)

## [1] 9.49
```

this is answer no. 4.

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

Continues on page 17

Exercise VIII

In a lottery, 7 out of 36 are winning numbers. A lottery ticket consists of 7 different randomly selected (of the 36) numbers. If a ticket contains at least 4 winning numbers a prize is given.

Question VIII.1 (15)

What is the probability of winning a prize if you buy a lottery ticket?

1 □ 0.00062%

2 □ 0.031%

3 □ 0.54%

4* □ 1.64%

5 □ 2.5%

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

This is typical examplpe of a hyper geometric random variable, i.e. we have

$$P(X = x) = f(x, n, a, N)$$

where n is the number of draws ($n = 7$), a is the number of winning number ($a = 7$), and finally N is the number total numbers that might be drawn (here 36). The probabilyty can be found as

$$P(X \geq 4) = \sum_{i=4}^{7} P(X = i)$$

In R we can calculate his number by

```
sum(dhyper(4:7,7,36-7,7))

## [1] 0.01636622
```

i.e. 1.64% or answer no. 4.

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

Continues on page 18

Exercise IX

Below are observations of the return time for breast cancer (measured in the natural logarithm of months, i.e. log (months)).

Patient	1	2	3	4	5
log(month)	3.3	4.34	4.34	3.58	2.30

It can be assumed that the observations in the table are independent and normally distributed ($N(\mu, \sigma^2)$), and that the average and the empirical variance are $\hat{\mu} = 3.57$, and $s^2 = 0.72$, respectively.

Question IX.1 (16)

Based on the above estimates, what is the expected return time in months for the breast cancer?

1 □ 73

2 □ 54

3* □ 51

4 □ 36

5 □ 25

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

Since *log* of the return time is assumed normal the return time can be assumed *log*-normal $X \sim LN(\mu, \sigma^2)$ and the expected value of a *log*-normal random variable is

$$E[X] = e^{\mu+\sigma^2/2}$$

using the estimates given above we get

```
exp(3.57+0.72/2)

## [1] 50.90698
```

this is answer no. 3.

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

Question IX.2 (17)

What is a 95% confidence interval for the variance parameter (σ^2)?

1 □ $[0.32; 7.43]$

2* □ $[0.26; 5.95]$

3 □ $[3.07; 4.06]$

4 □ $[0.31; 4.05]$

5 □ $[2.52; 4.62]$

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

The general formula for a confidence interval for the variance is

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right]$$

Incserting $n = 5$ and $s^2 = 0.72$ we get

```
4*0.72/c(qchisq(0.975,df=4),qchisq(0.025,df=4))

## [1] 0.2584516 5.9452718
```

which is answer no. 2.

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

Continues on page 20

Exercise X

Two groups of observations of writing speed (measured in words per minute) are given in the table below

Group 1	35	50	55	60	65	60	70	55	45	55	60	45	65	55	50	60
Group 2	55	60	75	65	60	70	75	70	65	72	73	65	80	50	55	70

It is also given that the sample variance in the two groups are $s_1^2 = 78.23$ and $s_2^2 = 70.87$ respectively, and that the number of observations in each group is $n_1 = n_2 = 16$.

Question X.1 (18)

What is the number of degrees of freedom for a Welch t-test for whether the two groups have the same writing speed?

1 □ 16

2 □ 18.9

3 □ 23.2

4* □ 29.9

5 □ 32

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

We can either use the formula for adjested degrees of freedom, or we could type the number into R

```
s1 <- 78.73; s2 <- 70.87
n <- 16
(s1/n+s2/n)^2/((s1/n)^2/(n-1)+(s2/n)^2/(n-1))

## [1] 29.91741

Group1 <-c(35, 50, 55, 60, 65, 60, 70, 55, 45, 55, 60, 45, 65, 55, 50, 60)
Group2 <-c(55, 60, 75, 65, 60, 70, 75, 70, 65, 72, 73, 65, 80, 50, 55, 70)
t.test(Group1,Group2)

##
##  Welch Two Sample t-test
##
## data:  Group1 and Group2
## t = -3.583, df = 29.927, p-value = 0.001187
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -17.17242  -4.70258
## sample estimates:
## mean of x mean of y
##   55.3125   66.2500
```

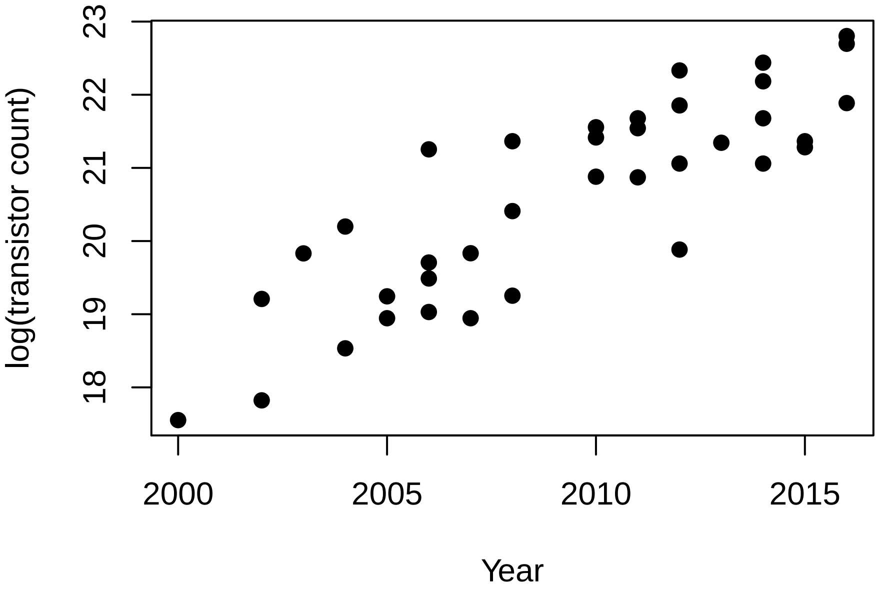
In eiter case we get $df = 29.9$.

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

Continues on page 22

Exercise XI

The number of transistors ("transistor count") on an integrated circuit is the most common measure of the complexity of integrated circuits. The figure below shows the natural logarithm of "transistor count", as a function of the year since 2000 for a selection of transistors that came on the market each year.



The figure suggests that a log-linear model for the number of "counts" could be appropriate. To investigate the hypothesis, the following R-code (where the variable `year` = Year-2000), have been run:

```
summary(fit <- lm(log(count) ~ year))

##
## Call:
## lm(formula = log(count) ~ year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4239 -0.6099 -0.1532  0.5468  1.5184
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.16313    0.27627  65.743  < 2e-16 ***
## year         0.26206    0.02676   9.792 1.47e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7212 on 35 degrees of freedom
## Multiple R-squared:  0.7326,Adjusted R-squared:  0.7249
## F-statistic: 95.88 on 1 and 35 DF,  p-value: 1.465e-11
```

Question XI.1 (19)

If the significance level $\alpha = 0.05$ is used, which of the following statements about the model is then correct (both conclusion and argument must be correct)?

1 □ The intercept with the y-axis is not significantly different from 0 since $0.28 > 0.05$

2 □ The intercept with the y-axis is significantly different from 0 since $0.027 < 0.05$

3* □ The intercept with the y-axis is significantly different from 0 since $2 \cdot 10^{-16} < 0.05$

4 □ The intercept with the y-axis is not significantly different from 0 since $2 \cdot 10^{-16} < 0.05$

5 □ The intercept with the y-axis is significantly different from 0 since $1.47 \cdot 10^{-11} < 0.05$

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

All statements are about the intercept, and from the summary we get that the intercept is significantly different from 0 since the p-value is $< 2 \cdot 10^{-16}$, which is answer no. 3

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

Question XI.2 (20)

A frequently discussed topic is the relative increase in transistor counts per year, i.e. $\gamma = \frac{\hat{Y}_{year+1}}{\hat{Y}_{year}}$ where $\hat{Y}_{year}$ is the expected number in year. What is a 95% confidence interval for the constant γ?

1 □ $[0.21; 0.32]$

2 □ $[18.37; 18.47]$

3* □ $[1.23; 1.37]$

4 □ $[17.31; 19.54]$

5 □ This can not be answered with the given information

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

First note that the number γ is given by

$$e^{\hat{\beta}_0+\hat{\beta}_1(year+1)}/e^{\hat{\beta}_0+\hat{\beta}_1(year)} = e^{\hat{\beta}_0+\hat{\beta}_1(year+1)-\hat{\beta}_0-\hat{\beta}_1(year)} = e^{\hat{\beta}_1}$$

Hence we can calculate a 95% confidence interval for γ by

$$e^{CI_{\beta_1}}$$

the confiden interval for β_1 is given by

$$CI_{\beta_1} = \hat{\beta}_1 \pm t_{1-\alpha/2} se_{\beta_1}$$

with $\hat{\beta}_1 = 0.2662$, $se_{\beta_1} = 0.02676$, and the degrees of freedom fir the t-distribution being 35, hence we can calculate the interval by

```
exp(0.2662 + c(-1, 1) * 0.02676*qt(0.975,df=35))

## [1] 1.235993 1.377852
```

which is answer no. 3.

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

For use in the further calculations, the following R-code has been run

```
c(mean(year), var(year))

## [1]  9.324324 20.169670
```

Continues on page 25

Question XI.3 (21)

What is, based on the model, the usual 95% confidence interval for $\log(count)$ in year 2020?

1 □ $[21.94; 24.87]$

2 □ $[438.27; 656.78]$

3 □ $[19.62; 27.19]$

4* □ $[22.78; 24.03]$

5 □ $[529.06; 565.99]$

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

We nee the formula

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot x_{new} \pm t_{1-\alpha/2}\sigma\sqrt{\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$

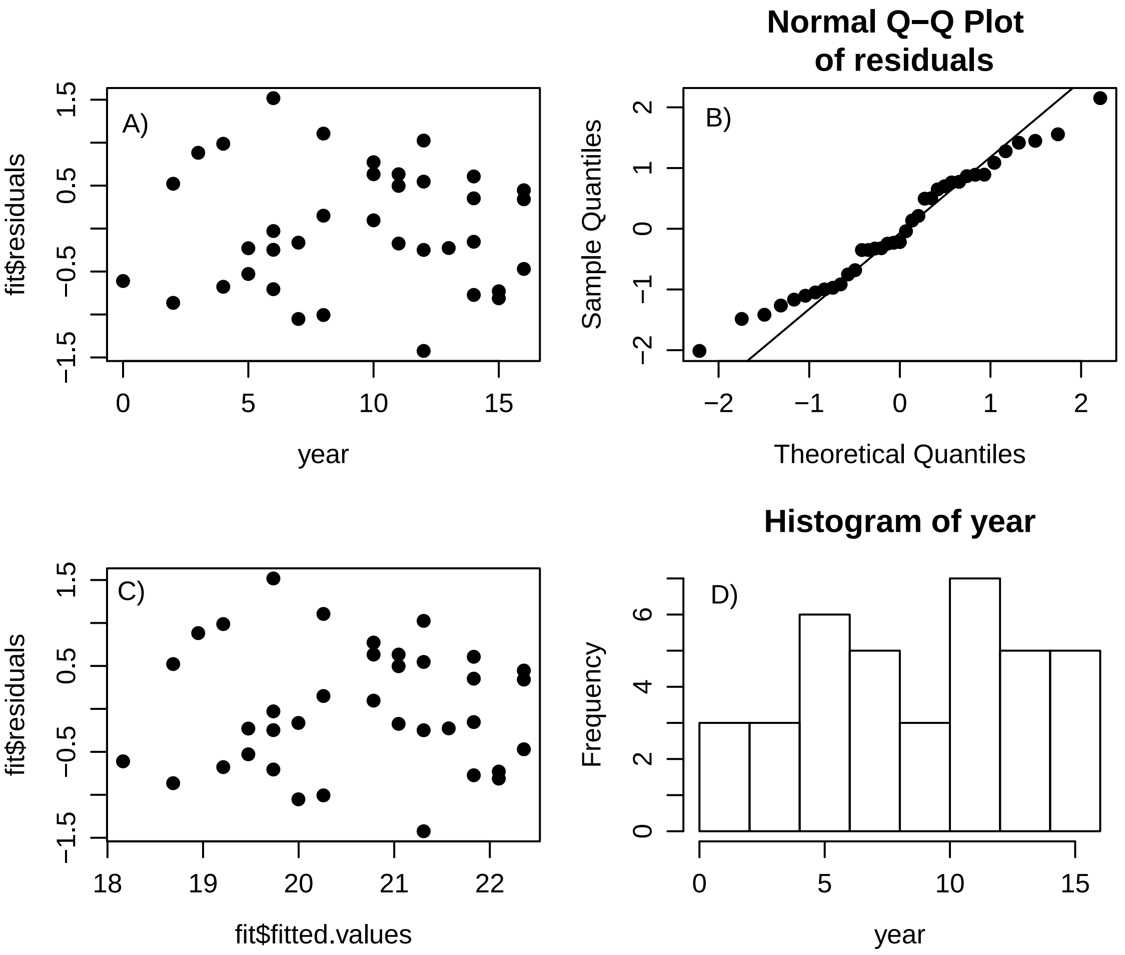
with $S_{xx} = (n-1)s_x^2$, inserting the observed values we get

```
m <- 9.324
Sxx <- 36 * 20.1697
18.16313 + 0.26206 * 20 + c(-1, 1) * qt(0.975,df=35)* 0.7212 * sqrt(1/37+(20-m)^2/Sxx)

## [1] 22.77630 24.03236
```

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

For the purpose of examining the model's assumptions, the following figure has been made.



Continues on page 27

Question XI.4 (22)

Which of the following statements about the model can, based on the figure, be correct (both conclusion, figure reference, and argument must be correct)?

1 □ Since the residuals as a function of `year` are not on a straight line (Plot A), the assumptions of the model are clearly not met.

2 □ Since the variable `year` does not follow a normal distribution (Plot D), the assumptions of the model are not met.

3 □ The assumption of independence is clearly not met (Plot B)

4 □ As the variable `year` approximately follows a normal distribution (Plot D), the assumptions of the model are met.

5* □ The assumption of variance homogeneity appears to be met (Plot C)

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

The residuals as a function of year should appear ramdom, hence 1is not correct. The regressor is not required to follow a normal, hence 2 is not correct. Plot B is about the normal assumption, not independence hence 3 is not correct. Answer no 4 is not correct with the same argument as answer no. 2. Plot C can be used for checking variance homogeneity, and the assumption appears to be fulfilled hence answer no. 5 is correct.

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

As a first study of whether the slope changes over time, the following model has been run in R (where `year2` = $\texttt{year}^2$)

```r
fit <- lm(log(count)~year+year2)
```

The model 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 $\boldsymbol{X}$ is the so-called design matrix.

Question XI.5 (23)

Which of the following statements, about the design matrix corresponding to the model in R above, is correct?

1 □ The first column contains log-counts

2 □ The second column consists of the vector `year2`

3 □ The first column consists of the vector `year`

4* □ The first column is a vector of ones

5 □ The design matrix has 3 rows

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

The i'th row of the design matrix will be

$$\boldsymbol{X}_{i,\cdot} = [1 \quad year_i \quad year_i^2]$$

hence the forst column will be a vector of ones and 4 is correct (and on the same ground answer 1,2 and 3 is not correct). The design matrix have the same number of rows as the number of observations (37), hence 5 is not correct.

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

Continues on page 29

For use in investigating the parameters in the model, the following R code has been run (where `X` is the design matrix)

```
solve(t(X) %*% X)

##              [,1]         [,2]          [,3]
## [1,]  0.405562066 -0.091123952  4.420991e-03
## [2,] -0.091123952  0.025056573 -1.337289e-03
## [3,]  0.004420991 -0.001337289  7.552324e-05
```

Question XI.6 (24)

If we denote the mean value parameters of the model $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]$, what is then the estimate of the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$?

1 □ -0.0911

2 □ -0.904

3 □ -0.00133

4* □ -0.972

5 □ 0.00442

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

The variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ is

$$V[\hat{\boldsymbol{\beta}}] = \sigma^2 \boldsymbol{X}^T \boldsymbol{X}$$

Hence the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ is

$$\frac{\sigma^2 (\boldsymbol{X}^T \boldsymbol{X})_{3,2}}{\sqrt{\sigma^2 (\boldsymbol{X}^T \boldsymbol{X})_{2,2} \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})_{3,3}}} = \frac{(\boldsymbol{X}^T \boldsymbol{X})_{3,2}}{\sqrt{(\boldsymbol{X}^T \boldsymbol{X})_{2,2} (\boldsymbol{X}^T \boldsymbol{X})_{3,3}}}$$

Inserting the observation from the R-code we get

```
-0.0013373/sqrt(0.025056*7.552e-5)

## [1] -0.9721695
```

which is answer no 4

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

Continues on page 30

Exercise XII

Let X and Y be random variables with variances σ_X^2 and σ_Y^2, respectively. Also, let the correlation between X and Y be ρ. Construct a random variable $Z = X - aY$ where $a \in \mathbb{R}$.

Question XII.1 (25)

Which value should a have for Z and Y to be uncorrelated?

1 □ ρ

2 □ 0

3 □ $\rho\sigma_x\sigma_y$

4* □ $\rho\sigma_x/\sigma_y$

5 □ $\rho\sigma_y^2$

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

The covariance between Z ad Y is

$$cov(Y, Z) = cov(Y, X - aY) = cov(Y, X) - aV(Y) = \sigma_x\sigma_y\rho - a\sigma_y^2$$

solving for a we get

$$a = \frac{\sigma_x\sigma_y\rho}{\sigma_y^2} = \frac{\sigma_x\rho}{\sigma_y}$$

This is answer no. 4.

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

Continues on page 31

Exercise XIII

As a part of a biological experiment, the growth of a certain type of bacteria was investigated in a Petri dish. The Petri dish was divided into 20 areas of equal size, and the number of bacterial colonies in each area was counted. This resulted in the following 20 observations, which have been read into R in the vector `bacteria`:

```
bacteria

##  [1] 7 6 4 6 2 2 7 1 2 5 7 6 3 3 6 6 2 6 4 5
```

The observations are assumed to be independent and Poisson distributed with mean λ.

Question XIII.1 (26)

Initially, it was decided to analyse the data without making use of the Poisson distribution assumption. Which of the following sequences of R code computes a 95% confidence interval for the standard deviation of the distribution of the number of bacterial colonies in each area using non-parametric bootstrap?

1 □

```
counts <- replicate(10000, sample(bacteria, replace = TRUE))
sd_count <- apply(counts, 2, sd)
quantile(sd_count, c(0.05, 0.95))
```

2 □

```
counts <- replicate(10000, rpois(20, mean(bacteria)))
sd_count <- apply(counts, 2, sd)
quantile(sd_count, c(0.025, 0.975))
```

3 □

```
counts <- replicate(10000, rpois(20, mean(bacteria)))
sd_count <- apply(counts, 2, sd)
quantile(sd_count, c(0.05, 0.95))
```

4* □

```
counts <- replicate(10000, sample(bacteria, replace = TRUE))
var_count <- apply(counts, 2, var)
quantile(sqrt(var_count), c(0.025, 0.975))
```

5 □

```
counts <- replicate(10000, rnorm(20, mean(bacteria), sd(bacteria)))
sd_count <- apply(counts, 2, sd)
quantile(sd_count, c(0.025, 0.975))
```

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

Since we do not use the distribution assumption this is a nonparametric bootstrap. Hence bootstrap samples are generated by resampling from the data set with replacement, and a 95% confidence interval is computed for the square root of the variance. This is answer no. 4.

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

Question XIII.2 (27)

What is computed in the R code below?

```
counts <- replicate(10000, rpois(20, mean(bacteria)))
y <- apply(counts, 2, sd)/apply(counts, 2, mean)
quantile(y, c(1-0.95, 0.95))
```

1 □ A 95% confidence interval for the coefficient of variaton of the distribution of the number of bacterial colonies in each area, computed using parametric bootstrap.

2 □ The ratio between the 90% confidence intervals for, respectively, the standard deviation and the mean of the distribution of the number of bacterial colonies in each area, computed using parametric bootstrap.

3* □ A 90% confidence interval for the coefficient of variation of the distribution of the number of bacterial colonies in each area, computed using parametric bootstrap.

4 □ 90% confidence intervals for the standard deviation and the mean of the distribution of the number of bacterial colonies in each area, computed using parametric bootstrap.

5 □ A 90% confidence interval for the coefficient of variation of the distribution of the number of bacterial colonies in each area, computed using non-parametric bootstrap.

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

Bootstrap samples are generated from a Poisson distribution with mean estimated from the data. The coefficient of variation is computed for each sample, and the 0.05 and 0.95 quantiles of the distribution of these coefficients of variation are computed.

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

Continues on page 33

Exercise XIV

As a part of a biological experiment, the growth of two different types of bacteria was investigated in a Petri dish. The Petri dish was divided into 15 areas of equal size, and the number of bacterial colonies in each area was counted. This resulted in the following 15 observations for each of the two bacteria types, which have been read into R in the vectors `bact1` and `bact2`:

```
bact1

##  [1] 8 5 5 4 7 7 1 4 5 4 6 0 1 4 5

bact2

##  [1]  5  5  7 10  8  7  5 10  7  4  9  6  5  3  4
```

The observations are assumed to be independent. The number of bacterial colonies in each area is assumed to be Poisson distributed with unknown mean λ_i for bacteria type i (here, i is either 1 or 2).

Question XIV.1 (28)

Which sequence of R code computes a 99% parametric bootstrap confidence interval for the difference in the expected number of bacterial colonies in each area between the two types of bacteria? Use the R output to conclude whether this difference is significant at significance level $\alpha = 0.01$. (Both parts of the answer must be correct).

1 □

```
z1 <- replicate(10000, sample(bact1, replace = TRUE))
z2 <- replicate(10000, sample(bact2, replace = TRUE))
d <- apply(z1, 2, mean) - apply(z2, 2, mean)
quantile(d, c(0.005, 0.995))

##        0.5%       99.5%
## -4.00000000 -0.06666667
```

As 0 is not included in the confidence interval, there is a significant difference in the number of bacterial colonies in each area between the two types of bacteria.

2 □

```
z1 <- replicate(10000, sample(bact1, replace = TRUE))
z2 <- replicate(10000, sample(bact2, replace = TRUE))
d <- apply(z1, 2, mean) - apply(z2, 2, mean)
quantile(d, c(0.025, 0.975))

##       2.5%      97.5%
## -3.4666667 -0.4666667
```

As 0 is not included in the confidence interval, there is a significant difference in the number of bacterial colonies in each area between the two types of bacteria.

3 □

```
z1 <- replicate(10000, rnorm(15, mean(bact1), sd(bact1)))
z2 <- replicate(10000, rnorm(15, mean(bact2), sd(bact2)))
d <- apply(z1, 2, mean) - apply(z2, 2, mean)
quantile(d, c(0.005, 0.995))

##       0.5%      99.5%
## -4.0125761  0.1580288
```

As 0 is included in the confidence interval, there is no significant difference in the number of bacterial colonies in each area between the two types of bacteria.

4* □

```
z1 <- replicate(10000, rpois(15, mean(bact1)))
z2 <- replicate(10000, rpois(15, mean(bact2)))
d <- apply(z1, 2, mean) - apply(z2, 2, mean)
quantile(d, c(0.005, 0.995))

##       0.5%      99.5%
## -4.1333333  0.2003333
```

As 0 is included in the confidence interval, there is no significant difference in the number of bacterial colonies in each area between the two bacteria types.

5 □

```
z1 <- replicate(10000, rpois(15, mean(bact2)-mean(bact1)))
z2 <- replicate(10000, rpois(15, mean(bact2)-mean(bact1)))
d <- apply(z1, 2, mean) - apply(z2, 2, mean)
quantile(d, c(0.005, 0.995))

##      0.5%     99.5%
## -1.266667  1.266667
```

As 0 is included in the confidence interval, there is no significant difference in the number of bacterial colonies in each area between the two types of bacteria.

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

Pairs of bootstrap samples are generated from two Poisson distribution with means estimated from the data. The difference between sample means is computed for each pair of bootstrap samples, and the 0.005 and 0.995 quantiles of the differences are computed.

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

Continues on page 36

Exercise XV

Bob throws a dart arrow against a dartboard located on a wall. Let $[X_1, X_2]$ be a bivariate normal random variable that measures the distance from the center of the disc to where the arrow hits in vertical and horizontal direction, respectively. Assume that the mean of both X_1 and X_2 is 0, that the correlation between X_1 and X_2 is 0, and that $\sigma_X^2 = \sigma_Y^2 = 1$

Question XV.1 (29)

If the diameter of the disc is 4, what is then the probability that Bob hits the wall (Hint: start by writing down the square distance to the center)?

1 □ 0.02

2* □ 0.14

3 □ 0.06

4 □ 0.10

5 □ 0.18

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

The radius of the disc is 2 and hence the probability of missing in

$$P(\text{missing disc}) = P(\sqrt{X_1^2 + X_2^2} > 2) = P(X_1^2 + X_2^2 > 4)$$

Since X_1 and X_1 follow a bivariate normal distribution with zero correlation, they are also independent, and hence $X_1^2 + X_2^2 \sim \chi^2(2)$. The numerical value of the probability can be calculated by

```r
1-pchisq(4,df=2)

## [1] 0.1353353
```

using usual rounding we get 0.14, which is answer number 2.

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

Question XV.2 (30)

Lise also throws a dart arrow against the dartboard. Lise's throw also follows a bivarite normal distribution, but with the variance parameters $\sigma_X^2 = \sigma_Y^2 = \frac{1}{2}$, while the other parameters are as Bobs. What is the probability that Lise's throw hits closer to the center than Bob's?

1 □ 0.8

2 □ 0.75

3* □ 0.667

4 □ 0.5

5 □ 0.9

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

The required probality is

$$P(D_{Dise} > D_{Bob}) = P\left(\frac{D^2_{Lise}}{D^2_{Bob}} > 1\right)$$

where D_i is the distance for person i. The squared distance for each of the two will follow a χ^2 distribution with 2 degrees of freedom, hence tha ration will follow a F-distribution with 2 and 2 degrees of freedom, and we can calculate the probability by

```
pf(2,df1=2,df2=2)

## [1] 0.6666667
```

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

SÆTTET ER SLUT. God sommer!