Technical University of Denmark *Page 1 of 42 pages.*

Written examination: 14. August 2016

Course name and number: **Introduction to Statistics (02323, 02402 og 02593)**

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 11 exercises. To answer the questions you need to fill in the prepared 30-question multiple choice form (on three 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	I.2	I.3	II.1	II.2	II.3	II.4	II.5	III.1	IV.1
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer	3	3	2	4	2	5	2	1	4	5

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

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

The questionnaire contains 42 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 power company has developed an app that can help their consumers to analyze and reduce their electricity consumption. It must now be tested whether users of the app have reduced their electricity consumption after they have installed it. The electricity consumption is measured in kWh. Let X denote the difference in electricity consumption between the month before they installed the app (X_{before}) and electricity consumption the month after they installed the app (X_{after}), such that

$$X = X_{\text{after}} - X_{\text{before}}$$

The difference is registered for 40 randomly selected users who have installed the app at different times of the year. The sample average and sample standard deviation are calculated to

$$\bar{x} = -22.6$$
$$s_X = 45.5$$

Question I.1 (1)

Calculate a 95% confidence interval for difference in mean μ_X in electricity consumption from before to after the app was installed:

1 ☐ $-11.3 \pm 2.02 \cdot \frac{45.5}{39} = [-13.7, -8.94]$

2 ☐ $-22.6 \pm 2.02 \cdot \frac{45.5}{39} = [-25.0, -20.2]$

3* ☐ $-22.6 \pm 2.02 \cdot \frac{45.5}{6.32} = [-37.1, -8.06]$

4 ☐ $-22.6 \pm 2.02 \cdot \frac{2070}{6.32} = [-684, 639]$

5 ☐ $-22.6 \pm 2.02 \cdot \frac{2070}{39} = [-130, 84.6]$

———— FACIT-BEGIN ————

We simply use Method 3.9 for calculating the one sample confidence interval of the mean for a single sample and insert the correct numbers

$$-22.6 \pm 2.02 \cdot \frac{45.5}{6.32} = [-37.1, -8.06]$$

where 2.02 is the 0.975 t-quantile found in R with $40 - 1 = 39$ degrees of freedom

```r
qt(0.975,39)

## [1] 2.022691
```

———— FACIT-END ————

Question I.2 (2)

Is there a significant decrease in electricity consumption from from the month before to the month after the installation of the app at 5% significance level (both the conclusion and reasoning (p-value) must be correct)?

1 ☐ Yes, a significant decrease can be detected, since the p-value for the obvious two-sided test is 0.027

2 ☐ No, a significant decrease cannot be detected, since the p-value for the obvious two-sided test is 0.027

3* ☐ Yes, a significant decrease can be detected, since the p-value for the obvious two-sided test is 0.0032

4 ☐ No, a significant decrease cannot be detected, since the p-value for the obvious two-sided test is 0.0032

5 ☐ No, a significant decrease cannot be detected, since the p-value for the obvious two-sided test is 0.21

———— FACIT-BEGIN ————

We follow Method 3.23. We want to test the hypothesis

$$\begin{aligned} \mathrm{H}_0 &: \mu_X = 0 \\ \mathrm{H}_1 &: \mu_X \neq 0 \end{aligned}$$

First we calculate t_{obs} using the equation:

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{-22.6 - 0}{45.5/\sqrt{40}} = -3.141$$

We can then take the absolute value and find the p-value using a t-distribution with $40-1 = 39$ degrees of freedom

```r
2 * (1-pt(3.141427, 39))

## [1] 0.003204177
```

which is way below 0.05 and hence very strong evidence against the null hypothesis, which is rejected and a significant decrease (since $\bar{x} < 0$) in electricity consumption is found.

———— FACIT-END ————

Question I.3 (3)

It has been found that some consumers who install the app don't start to use the app right away after installation. Therefore, the onboarding (the process the user must go through the first time the app is opened after installation) has been redesigned to get users started faster. If the probability that a user doesn't get started right away is set to $p = 0.20$ and 100 new users are registered, and X denotes the number of those who doesn't get started right away, then find the one of the following R expressions, which calculates the probability of getting less than 10 new users who doesn't get started away, i.e. $P(X < 10)$?

1 ☐ `phyper(q=1, m=20, n=80, k=10)`

2* ☐ `pbinom(q=9, size=100, prob=0.2)`

3 ☐ `dbinom(x=10, size=100, prob=0.2)`

4 ☐ `1 - pbinom(q=10, size=100, prob=0.2)`

5 ☐ `dbinom(x=10, size=100, prob=0.2)`

———— FACIT-BEGIN ————

We are sampling a binomial distributed variable which counts the number of successes in 100 draws and by defining a success to: the user do not start right after installation, the probability of success is 0.2. In R `pbinom()` calculates the probability of `q` or less successes. Therefore, to calculate $P(X < 10) = P(X \leq 9)$ then `q=9`. See Section 2.3.1 about binomial distribution.

———— FACIT-END ————

Continues on page 5

Exercise II

In a series of experiments it has been investigated how the compressive strength of concrete depends on the composition of the concrete. The registered explanatory variables are the amount of cement, water and sand measured in kg/m^3. The compressive strength is measured in MPa. A summary of the data is given in the following table:

	Cement	Water	Fine	Strength
Min.	200.0	146.0	594.0	12.25
1st Qu.	289.0	185.0	754.0	22.49
Median	339.0	186.0	781.0	31.35
Mean	344.9	188.5	776.4	31.83
3rd Qu.	393.0	192.0	809.0	37.42
Max.	540.0	228.0	945.0	74.99

Question II.1 (4)

First a simple linear regression model with the amount of cement as an explanatory variable is fitted and the following output from R is obtained (where a few numbers are replaced by letters):

```
## Call:
## lm(formula = Strength ~ Cement, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.5512  -0.6858   0.6280   1.4791  20.8302
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.736438   3.389030     A   1.96e-05 ***
## Cement        0.137930   0.009567     B   < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.05 on 59 degrees of freedom
## Multiple R-squared:  0.7789,Adjusted R-squared:  0.7752
## F-statistic: 207.9 on 1 and 59 DF,  p-value: < 2.2e-16
```

The relevant test statistic for testing if the amount of cement has a significant effect on the compressive strength is found to?

1 ☐ $0.009567/0.137930 \approx 0.06936$

2 ☐ $1.96e^{-5}$

3 ☐ $-15.736438/3.389030 \approx -4.643$

4* ☐ $0.137930/0.009567 \approx 14.42$

5 ☐ 6.05

———— FACIT-BEGIN ————

We use Theorem 5.12 to find the test-statistic, where the null hypothesis is that the cement has no effect, thus $\beta_1 = 0$

$$T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}} = \frac{0.137930 - 0}{0.009567} = 14.41727$$

———— FACIT-END ————

Question II.2 (5)

The following model has been fitted

$$Strength_i = \beta_0 + \beta_1 \cdot Cement_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

and it is wanted to use the model fit for predicting the compressive strength given the amount of cement.

The prediction has the lowest variance when the amount of cement is:

1 ☐ 339.0 kg/m^3

2* ☐ 344.9 kg/m^3

3 ☐ 540.0 kg/m^3

4 ☐ 0.0 kg/m^3

5 ☐ Cannot be answered based on the provided information

———— FACIT-BEGIN ————

The lowest variance is found when the predictor (cement) takes on its mean value. This can be read from the table in the beginning of the exercise to be 344.9 kg/m^3. This can be explained by looking at Method 5.18, where the last term becomes 0 when $x_{\text{new}} = \bar{x}$, and hence this will have the least variance for the prediction.

———— FACIT-END ————

Question II.3 (6)

An equivalent analysis for the dependence between the amount of water and the compressive strength is carried out. The following R output is obtained:

```
## Call:
## lm(formula = Strength ~ Water, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.361  -8.593  -1.161   5.239  28.568
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 116.1345    23.6644   4.908 7.62e-06 ***
## Water        -0.4474     0.1253  -3.570 0.000718 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.67 on 59 degrees of freedom
## Multiple R-squared:  0.1776,Adjusted R-squared:  0.1637
## F-statistic: 12.74 on 1 and 59 DF,  p-value: 0.0007184
```

If the residuals meet the usual assumptions, then the conclusion of the analysis is:

1 ☐ Water doesn't have a significant effect on the compressive strength. More water results in stronger concrete

2 ☐ Water doesn't have a significant effect on the compressive strength. Less water results in stronger concrete

3 ☐ Water has a significant effect on the compressive strength. More water results in stronger concrete

4 ☐ Water doesn't have a significant effect on the compressive strength. Therefore it cannot be determined how the amount of water affects the compressive strength

5* ☐ Water has a significant effect on the compressive strength. Less water results in stronger concrete

———— FACIT-BEGIN ————

The p-value for the slope is 0.000718 which is lower than 5% so it is very significant. As the sign of the estimated coefficient is negative then the compressive strength increases when the amount of water is reduced.

———— FACIT-END ————

Question II.4 (7)

It is chosen to estimate a multiple linear regression model using the square root of the compressive strength as response. The other three variables are used as explanatory variables:

```
## Call:
## lm(formula = sqrt(Strength) ~ Cement + Water + Fine, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31854 -0.12044  0.06208  0.18913  0.73313
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.0623090  1.3125406   3.857 0.000295 ***
## Cement       0.0120327  0.0007024  17.131  < 2e-16 ***
## Water       -0.0244132  0.0037730  -6.470 2.41e-08 ***
## Fine         0.0012022  0.0009038   1.330 0.188774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3293 on 57 degrees of freedom
## Multiple R-squared:  0.9069,Adjusted R-squared:  0.902
## F-statistic: 185.1 on 3 and 57 DF,  p-value: < 2.2e-16
```

Which of the following conclusions from the above output is most correct:

1 ☐ A model which only has "Cement" as explanatory variable should be fitted

2* ☐ The variable "Fine" should be removed and the reduced model should be fitted

3 ☐ The variable "(Intercept)" should be removed and the reduced model should be fitted

4 ☐ The variable "Water" should be removed and the reduced model should be fitted

5 ☐ A model which only has "Fine" as explanatory variable should be fitted

———— FACIT-BEGIN ————

The parameter "Fine" has a non-significant p-value and therefore the model should be updated without that variable (Backward selection as explained in Section 6.3). That includes estimating the parameters in the reduced model.

———— FACIT-END ————

Question II.5 (8)

Based on the presented R-outputs in the exercise the number of observations used in the analyses is found to:

1* ☐ 61

2 ☐ 60

3 ☐ 59

4 ☐ 2

5 ☐ 1

———— FACIT-BEGIN ————

For the first two analysis a model with 2 parameters is estimated and there are 59 degrees of freedom for the residuals. Since we know from Theorem 6.2 that $DF = n - (p+1)$, we can use the R output above where there is 3 parameters and 57 degrees of freedom to conclude that $n = 57 + 3 + 1 = 61$.

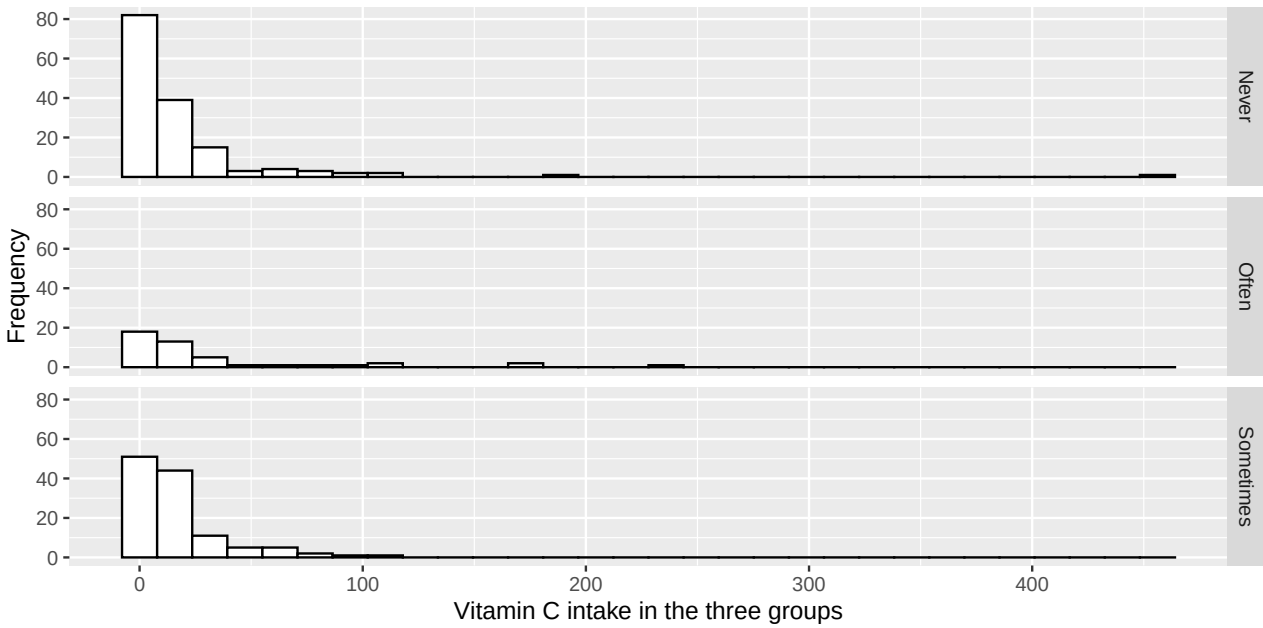
———— FACIT-END ————

Continues on page 10

Exercise III

A study has been conducted of the relation between intake of vitamin C and frequency of exercise. Below is a table of descriptive statistics for the intake of vitamin C in the three exercise groups as well as histograms of the observations from the experiments:

	Exercise		
	Never	Sometimes	Often
Mean	20.11	16.71	33.26
Median	7.23	9.27	11.27
SD	44.05	20.55	52.42
n	152	120	45



We want to study whether there is a dependence between the intake of vitamin C and the frequency of exercise. To do this a one-way analysis of variance is conducted. Where `z` is the vitamin C intake and `Group` is a factor with three levels (Never exercise, Sometimes exercise, Often exercise).

Two analyses have been conducted by running the following R-code (the reading of data is not shown):

```
anova(lm(z ~ Group, data = dat))

## Analysis of Variance Table
##
## Response: z
##            Df Sum Sq Mean Sq F value Pr(>F)
## Group       2   9060  4529.8   3.064 0.0481 *
```

```
## Residuals 314 464227  1478.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(log(z) ~ Group, data = dat))
```

```
## Analysis of Variance Table
##
## Response: log(z)
##            Df Sum Sq Mean Sq F value Pr(>F)
## Group       2   2.94  1.4693   0.918 0.4004
## Residuals 314 502.53  1.6004
```

Question III.1 (9)

Which of the following statements is correct?

1 ☐ The analysis of `z` is the most correct since it is seen from the table that the intake of vitamin C is higher in the group who exercise often

2 ☐ The analysis of `z` is the most correct since we want to determine the effect on vitamin C and not the effect on the logarithm of vitamin C

3 ☐ The analysis of `log(z)` is the most correct since here the effect of Group was not statistically significant

4* ☐ The analysis of `log(z)` is the most correct since the histograms indicate that the assumption of normality is not be valid for `z`

5 ☐ The analysis of `z` is the most correct since here the effect of Group was statistically significant

———— FACIT-BEGIN ————

We want to compare the means in the three groups. However, from the table of summary statistics we can see that for each group its mean is much higher than its median and they also have high SD values. So the distributions are highly skewed, which is also seen in the histograms.

From Section 8.2.4, we know that this will give us problems with the assumption that the data should be normally distributed in each group.

Looking at the histograms we see that we have only positive values and highly skewed distributions, therefore taking the logarithm is the most correct to do in this case in order to transform towards normality as described in Section 3.1.9.

———— FACIT-END ————

Continues on page 13

Exercise IV

In a study of high blood pressure it is investigated whether two different diets influences the blood pressure. The study was conducted with 150 persons which were divided in 2 groups of 75 people. Group I is a control group receiving normal diet, while group II received healthy dietary supplements. The results are shown in the following table; a certain level of the blood pressure (or above) is defined as 'high' (not necessarily too high):

Group	Normal blod pressure	High blod pressure	Total
I	55	20	75
II	57	18	75
Total	112	38	150

Question IV.1 (10)

A 95% confidence interval for the proportion of persons on normal diet (Group I) with high blood pressure is:

1 ☐ $0.49 \pm 1.96\sqrt{\frac{0.49107 \cdot (1-0.49107)}{112}}$ giving 0.49 ± 0.09

2 ☐ $0.49 \pm 1.645\sqrt{\frac{0.49107 \cdot (1-0.49107)}{55}}$ giving 0.49 ± 0.11

3 ☐ $0.27 \pm 1.645\sqrt{\frac{0.26667 \cdot (1-0.26667)}{112}}$ giving 0.27 ± 0.07

4 ☐ $0.27 \pm 1.96\sqrt{\frac{0.49107 \cdot (1-0.49107)}{75}}$ giving 0.27 ± 0.09

5* ☐ $0.27 \pm 1.96\sqrt{\frac{0.26667 \cdot (1-0.26667)}{75}}$ giving 0.27 ± 0.10

———— FACIT-BEGIN ————

We will use Method 7.3 to find the confidence interval of a single proportion using the formula

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

where $z_{1-\alpha/2}$ is found in R as

```
qnorm(0.975)

## [1] 1.959964
```

And $\hat{p}$ can be found as

$$\hat{p} = 20/75 = 0.26667 = 0.27$$

We then insert into the formula:

$$0.27 \pm 1.96\sqrt{\frac{0.26667 \cdot (1 - 0.26667)}{75}} = 0.27 \pm 0.10$$

Or in R:

```
(ph <- 20/75)

## [1] 0.2666667

qnorm(0.975)

## [1] 1.959964

1.96*sqrt(ph*(1-ph)/75)

## [1] 0.100083

ph + c(-1, 1)*1.96*sqrt(ph*(1-ph)/75)

## [1] 0.1665836 0.3667497
```

———— FACIT-END ————

Question IV.2 (11)

Which of the following computed values is a usual test statistic for a test of the hypothesis that the proportions of people with high blood pressure in Groups I and II are the same?

1 ☐ $1.96^2 = 3.84$

2 ☐ $\frac{1}{56} + \frac{1}{19} = 0.07$

3* ☐ $\frac{1}{56} + \frac{1}{56} + \frac{1}{19} + \frac{1}{19} = 0.14$

4 ☐ $\frac{55-57}{\sqrt{150}} = 0.16$

5 ☐ $\frac{(112-38)^2}{150} = 36.5$

———— FACIT-BEGIN ————

The easiest approach is probably to use R to do a usual χ^2-test in a 2-by-2 table (see Example 7.21):

```r
X <- matrix(c(55, 20, 57, 18), nrow=2)
res <- chisq.test(X, correct=FALSE)

X
```

```
##      [,1] [,2]
## [1,]   55   57
## [2,]   20   18
```

```r
res
```

```
##
##  Pearson's Chi-squared test
##
## data:  X
## X-squared = 0.14098, df = 1, p-value = 0.7073
```

And from this we can also get the expected values in each cell, so we can use Method 7.20 to find the χ^2 test-statistic manually:

```r
res$expected
```

```
##      [,1] [,2]
## [1,]   56   56
## [2,]   19   19
```

So, the test-statistic becomes (note that the rows/columns have been switched around):

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \frac{1}{56}+\frac{1}{56}+\frac{1}{19}+\frac{1}{19} = 0.14$$

———— FACIT-END ————

Question IV.3 (12)

Suppose you had compared the proportions between 3 diet groups, that is, had a data table in the following form (e.g. x_{32} denotes the number of people in Group III, who had high blood pressure):

Group	Normal blod pressure	High blod pressure
I	x_{11}	x_{12}
II	x_{21}	x_{22}
III	x_{31}	x_{32}

Which distribution would be used in the usual hypothesis test for a comparison of the three proportions (i.e. the proportions of persons in each of the three groups with high blood pressure)?

1 ☐ A χ^2-distribution with 3 degrees of freedom

2 ☐ An F-distribution with 2 and 3 degrees of freedom

3 ☐ An F-distribution with 2 and 5 degrees of freedom

4* ☐ A χ^2-distribution with 2 degrees of freedom

5 ☐ A χ^2-distribution with 1 degree of freedom

———— FACIT-BEGIN ————

As stated in Method 7.22 a χ^2-test in an r-by-c frequency table is tested using the χ^2-distribution with $(r-1)\cdot(c-1) = 2$ degrees of freedom.

———— FACIT-END ————

Question IV.4 (13)

This question is based on the text and the table given in the previous question.

What is a pre-planned 95% confidence interval for the difference between the proportions with normal blood pressure in Group I and Group III?

1* ☐ $\frac{x_{11}}{x_{11}+x_{12}} - \frac{x_{31}}{x_{31}+x_{32}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{(x_{11}+x_{12})^3} + \frac{x_{31}\cdot x_{32}}{(x_{31}+x_{32})^3}}$

2 ☐ $\frac{x_{11}}{x_{12}} - \frac{x_{31}}{x_{32}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{(x_{11}+x_{12})^3} + \frac{x_{31}\cdot x_{32}}{(x_{31}+x_{32})^3}}$

3 ☐ $\frac{x_{11}}{x_{12}} - \frac{x_{31}}{x_{32}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{x_{11}+x_{12}} + \frac{x_{31}\cdot x_{32}}{x_{31}+x_{32}}}$

4 ☐ $\frac{x_{11}}{x_{11}+x_{21}+x_{31}} - \frac{x_{31}}{x_{11}+x_{21}+x_{31}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{(x_{11}+x_{21}+x_{31})^2} + \frac{x_{31}\cdot x_{32}}{(x_{11}+x_{21}+x_{31})^2}}$

5 ☐ $\frac{x_{11}}{x_{11}+x_{12}} - \frac{x_{31}}{x_{31}+x_{32}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{(x_{11}+x_{21}+x_{31})^2} + \frac{x_{31}\cdot x_{32}}{(x_{11}+x_{21}+x_{31})^2}}$

———— FACIT-BEGIN ————

We use Method 7.15 to find the confidence interval for the difference between two proportions. In this case $\hat{p}_1 = \frac{x_{11}}{x_{11}+x_{12}}$ and $\hat{p}_2 = \frac{x_{31}}{x_{31}+x_{32}}$ and we can find $z_{0.975}$ in R as

```r
qnorm(0.975)

## [1] 1.959964
```

We then need to calculate $\hat{\sigma}_{\hat{p}_1-\hat{p}_2}$. To simplify the expression a bit we notice that

$$\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} = \frac{x_{11}\cdot x_{12}}{(x_{11}+x_{12})\cdot(x_{11}+x_{12})\cdot(x_{11}+x_{12})} = \frac{x_{11}\cdot x_{12}}{(x_{11}+x_{12})^3}$$

and the same goes for the second term. Hence the final interval is it follows that the correct interval is:

$$\frac{x_{11}}{x_{11}+x_{12}} - \frac{x_{31}}{x_{31}+x_{32}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{(x_{11}+x_{12})^3} + \frac{x_{31}\cdot x_{32}}{(x_{31}+x_{32})^3}}$$

———— FACIT-END ————

Question IV.5 (14)

A new study is planned to explore a completely new diet. The required precision is that the 90% confidence interval for the proportion of people with normal blood pressure achieves a mean width of 0.1. The total cost for handling a single person is 100 kr, and there is assigned in total 25000 kr for this in the budget for the study. Can the requirement be fulfilled within the given budget (note, you know nothing about the value of the proportion)?

1 ☐ Yes, since $100 \cdot \left(\frac{1.96}{0.05}\right)^2 = 153664 < 25000$

2 ☐ Yes, since $100 \cdot 0.3 \cdot \cdot 0.7 \cdot \left(\frac{1.645}{0.05}\right)^2 = 22730.61 < 25000$

3 ☐ No, since $\frac{1}{4} \cdot \left(\frac{1.96}{0.05}\right)^2 \approx 384$

4 ☐ Yes, since $\frac{1}{4} \cdot \left(\frac{1.96}{0.05}\right)^2 \approx 384$

5* ☐ No, since $100 \cdot \frac{1}{4} \cdot \left(\frac{1.645}{0.05}\right)^2 = 27060.25 > 25000$

———— FACIT-BEGIN ————

First we want to know how many people we need for testing. We use Method 7.13. Since we have no information about the expected proportion p, we need to use Equation 7-25, with a ME of 0.05 (since the mean width of the confidence interval wanted is 0.1):

$$n = \frac{1}{4}\left(\frac{z_{0.95}}{0.05}\right)^2 = \frac{1}{4}\left(\frac{1.645}{0.05}\right)^2 = 270.60$$

Since we need 271 people at 100 DKK each our budget is too small

```r
100*0.25*(1.645/0.05)^2
```

```
## [1] 27060.25
```

———— FACIT-END ————

Continues on page 19

Exercise V

Question V.1 (15)

Let $X_i \sim N(\mu, \sigma^2)$, $i = 1, ..., n$. If $n = 10$ what is then

$$P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} > 2\right)$$

where S^2 and $\overline{X}$ is the empirical variance and average, respectively (random variables)?

1 ☐ $6.8 \cdot 10^{-5}$

2* ☐ 0.038

3 ☐ 0.012

4 ☐ 0.988

5 ☐ 0.962

———— FACIT-BEGIN ————

If we compare this with Theorem 2.89 we reconized it as a t-distribution with $n - 1$ degrees of freedom hence we can find the probability in R as

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

## [1] 0.03827641
```

———— FACIT-END ————

Continues on page 20

Exercise VI

Question VI.1 (16)

A multiple choice exam have 30 questions and 5 answer options for each question. There is one and only one correct answer for each question. What is the probability of answering correct on to at least 15 of the 30 questions, if you answer completely at random?

1 ☐ 0.000119

2* ☐ 0.00023

3 ☐ 0.835

4 ☐ 0.999

5 ☐ $4.5 \cdot 10^{-9}$

———— FACIT-BEGIN ————

We have 30 draws of success or non-success and each draw is independent of each other (i.e. we just put an answer at random for each question). Let X be the number of successes (correct answers) out of the 30 draws, then X is binomial distributed with $n = 30$ and with probability of a success $p = 1/5$ (see Section 2.3). We want to calculate

$$P(X \geq 15) = P(X > 14) = 1 - P(X < 14) \tag{1}$$

and this we can look in R by

```
1-pbinom(14,size = 30, prob=1/5)
```

```
## [1] 0.0002312256
```

———— FACIT-END ————

Question VI.2 (17)

Which of the following random variables should not be approximated by a normal distribution?

1 ☐ $\sum_{i=1}^{200} X_i$, $X_i \sim Pois(2)$

2 ☐ $\sum_{i=1}^{5} X_i$, $X_i \sim N(\mu, \sigma^2)$

3 ☐ $\sum_{i=1}^{10} X_i$, $X_i \sim Binom(1000, 0.5)$

4 ☐ $\sum_{i=1}^{100} X_i$, $X_i \sim Exp(1)$

5* ☐ $\sum_{i=1}^{7} X_i$, $X_i \sim Exp(1)$

———— FACIT-BEGIN ————

Using the central limit theorem described in Section 3.1.4 we see that 1 and 4 can be approximated by a normal distribution (they are sums of a large number (> 15) of random variables), 2 is a normal distribution (a sum of normal distributed variables is also normal distributed, Theorem 2.40), the binomial with n=1000 is well approximated by a normal distribution (Remark 7.4) and hence also the sum of such variables. This leaves only 5 as a relevant option and indeed the sum of 7 exponential random variables will not be well approximated by a normal distribution.

———— FACIT-END ————

Continues on page 22

Exercise VII

The lifetime of a particular type of electronic buttons is assumed to follow an exponential distribution with a mean of 5 years.

Question VII.1 (18)

What is the variance of the lifetime of such buttons?

1 ☐ 1 years

2 ☐ 1/5 years

3 ☐ 5 years

4* ☐ 25 years

5 ☐ 1/25 years

———— FACIT-BEGIN ————

According to Theorem 2.49 the mean and rate of an exponential distribution is related by $\lambda = \frac{1}{\mu}$. We know that $\mu = 5$, hence $\lambda = \frac{1}{5}$ and therefore we can again use Theorem 2.49 to find the variance

$$\sigma^2 = \frac{1}{\frac{1}{5^2}} = 25 \tag{2}$$

———— FACIT-END ————

Question VII.2 (19)

If 10 of such buttons are installed in different systems (without interaction), what is then the probability that none of these breaks down within the first year?

1* ☐ 0.1353

2 ☐ 0.4350

3 ☐ 0.1074

4 ☐ 0.3758

5 ☐ 0.6241

———— FACIT-BEGIN ————

First, we calculate the probability that one button will break down in the first year (as in the previous we have that the rate is $\lambda = \frac{1}{\mu} = \frac{1}{5}$ per year)

```
pexp(1, 1/5)

## [1] 0.1812692
```

and this we can use in the binomial distribution to calculate the probability that none of the 10 buttons will break down in the first year

```
dbinom(0, 10, 0.18127)

## [1] 0.135334

# or simply directly by
(1-0.18127)^10

## [1] 0.135334
```

———— FACIT-END ————

Continues on page 24

Exercise VIII

An experiment is carried out to examine four different methods (A, B, C, D) for removing impurities in a chemical process. At the same time it is wanted to adjust for using three different reactors in the experiment. The data are shown in the table below:

	Reactor 1	Reactor 2	Reactor 3	Sum
Method A	23.97	29.54	37.91	91.42
Method B	12.67	17.48	20.28	50.43
Method C	25.85	40.09	38.00	103.94
Method D	21.29	23.58	20.19	65.06
Sum	83.78	110.69	116.38	310.85

The sums of squares have been calculated:

	SS
Reactor	151.61
Method	593.40
Residual	100.73
Total variation	845.74

Question VIII.1 (20)

We now want to investigate whether it is reasonable to assume that the four methods clean equally well. Using the above sums of squares the usual test statistic, here denoted by A, as well as the critical value (the level of significance is $\alpha = 0.05$), is found to:

1 ☐ $A = 5.89$ and $A < 8.94$ (i.e. no significant effect of Method)

2* ☐ $A = 11.78$ and $A > 4.76$ (i.e. a significant effect of Method)

3 ☐ $A = 441.79$ and $A > 4.76$ (i.e. a significant effect of Method)

4 ☐ $A = 0.70$ and $A < 8.94$ (i.e. no significant effect of Method)

5 ☐ $A = 3.91$ and $A < 4.76$ (i.e. no significant effect of Method)

FACIT-BEGIN

Since this is a two-way ANOVA we can use Theorem 8.22 to calculate the test statistic:

$$F = \frac{SS(Method)/(k-1)}{SSE/((k-1)(l-1))}$$

Where k is the number of methods (here $k = 4$) and l the number of reactors (here $l = 3$).

From the table we have

$$SS(Method) = 593.40$$
$$SSE = 100.73$$

So the test statistic becomes

$$F = \frac{593.40/3}{100.73/6} = 11.78$$

This test statistic follows an F-distribution with $(k-1)$ and $(k-1)(l-1)$ degrees of freedom. So the critical value is from an F(3,6) and can be found in R as

```
qf(0.95,3,6)

## [1] 4.757063
```

Since the observed test statistic $11.78 > 4.76$ then the hypothesis of all the methods being equally good is rejected.

Using R:

```
anova(lm(Y ~ Method + Reactor))

## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Method     3 593.40 197.800 11.7821 0.00631 **
## Reactor    2 151.61  75.804  4.5153 0.06361 .
## Residuals  6 100.73  16.788
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

——— FACIT-END ———

Question VIII.2 (21)

A model for two-way analysis of variance will often be formulated for this type of experiment by

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma^2),\ i = (1, 2, 3, 4),\ j = (1, 2, 3)$$

Here y_{ij} is the observed purity for method i reactor j, μ the overall mean, α the method-effect and β the reactor-effect.

What is the usual estimate of the effect of Method B (i.e. $\hat{\alpha}_2$) in the model?

1 ☐ $\hat{\alpha}_2 = 50.43$

2 ☐ $\hat{\alpha}_2 = -260.42$

3 ☐ $\hat{\alpha}_2 = 16.81$

4 ☐ $\hat{\alpha}_2 = 4.20$

5* ☐ $\hat{\alpha}_2 = -9.09$

——— FACIT-BEGIN ———

According to Equation 8-35 and Equation 8-36 we can calculate

$$\hat{\mu} = \frac{1}{4 \cdot 3} \sum_{i=1}^{4} \sum_{j=1}^{3} y_{ij}$$

$$\hat{\alpha}_2 = \left(\frac{1}{3} \sum_{j=1}^{3} y_{2j} \right) - \hat{\mu}$$

Inserting here we have

$$\hat{\mu} = \frac{1}{12} \cdot 310.85 = 25.90$$

$$\hat{\alpha}_2 = \left(\frac{1}{3} \right) \cdot 50.43 - 25.90 = -9.09$$

——— FACIT-END ———

Question VIII.3 (22)

Before the experiment was conducted it was decided to compare Method B and Method C.

What is the 95% confidence interval for the comparison of B and C, if this is the only post-hoc comparison to be carried out?

1 ☐ $(50.43 - 103.94)/3 \pm 2.3060\sqrt{\frac{100.73}{12-4}(2/3)} = [-24.52, -11.16]$

2 ☐ $(50.43 - 103.94) \pm 2.4469\sqrt{\frac{100.73}{6}(2/3)} = [-61.70, -45.32]$

3 ☐ $(50.43 - 103.94)/3 \pm 2.4469\sqrt{\frac{100.73}{6}(1/3 + 1/4)} = [-25.49, -10.18]$

4* ☐ $(50.43 - 103.94)/3 \pm 2.4469\sqrt{\frac{100.73}{6}(2/3)} = [-26.02, -9.65]$

5 ☐ $(50.43 - 103.94)/3 \pm 2.3060\sqrt{\frac{100.73}{12-4}(1/3 + 1/4)} = [-24.09, -11.59]$

———— FACIT-BEGIN ————

We will use Method 8.9 and use SSE from the two-way analysis and $(n - k)$ replaced by $(k - 1)(l - 1)$ to find MSE. Method 8.9 now says:

$$\bar{y}_2 - \bar{y}_3 \pm t_{1-\alpha/2}\sqrt{\frac{SSE}{(4-1)(3-1)}(1/3 + 1/3)}$$

From the table we get

$$\begin{aligned} \bar{y}_2 &= 50.43/3 = 16.81 \\ \bar{y}_3 &= 103.94/3 = 34.65 \\ SSE &= 100.73 \end{aligned}$$

We have $(k - 1)(l - 1) = (4 - 1)(3 - 1) = 6$ degrees of freedom so $t_{1-\alpha/2}$

```
qt(0.975, df=6)
```

```
## [1] 2.446912
```

All in all we have

$$(50.43 - 103.94)/3 \pm 2.4469\sqrt{\frac{100.73}{6}(2/3)} = -17.8367 \qquad \pm 8.1861 = [-26.02; -9.65]$$

———— FACIT-END ————

Question VIII.4 (23)

If it was decided to make all pairwise comparisons between the methods, what is then the Bonferroni corrected "least significant difference (LSD)"?

1* ☐ $3.8630\sqrt{2 \cdot 16.788/3} = 12.92$

2 ☐ $3.4789\sqrt{2 \cdot 12.591/3} = 10.08$

3 ☐ $2.4469\sqrt{2 \cdot 16.788/6} = 5.79$

4 ☐ $3.8630\sqrt{2 \cdot 12.591/3} = 11.19$

5 ☐ $2.4469\sqrt{2 \cdot 16.788/3} = 8.19$

———— FACIT-BEGIN ————

We want to find $LSD_{Bonferroni}$. The standard formula for LSD is found in Remark 8.13, with the Bonferroni correction explained in Remark 8.14. We notive that we have $m = 3$ observations for each Method.

$$LSD_{Bonferroni} = t_{1-\alpha_{Bonferroni}/2}\sqrt{2 \cdot MSE/m}$$

Now the number of tests is $M = k(k-1)/2 = (4 \cdot 3)/2 = 6$ so $\alpha_{Bonferroni} = 0.05/6 = 0.008333$ and the quantile in the t-distribution with $(k-1)(l-1) = (4-1)(3-1) = 6$ degrees of freedom is

```
alphaBonf<- 0.05/6
qt(1-alphaBonf/2, df=6)

## [1] 3.862991
```

We also need MSE

$$MSE = \frac{SSE}{(k-1)(l-1)} = \frac{100.73}{6} = 16.788$$

So now

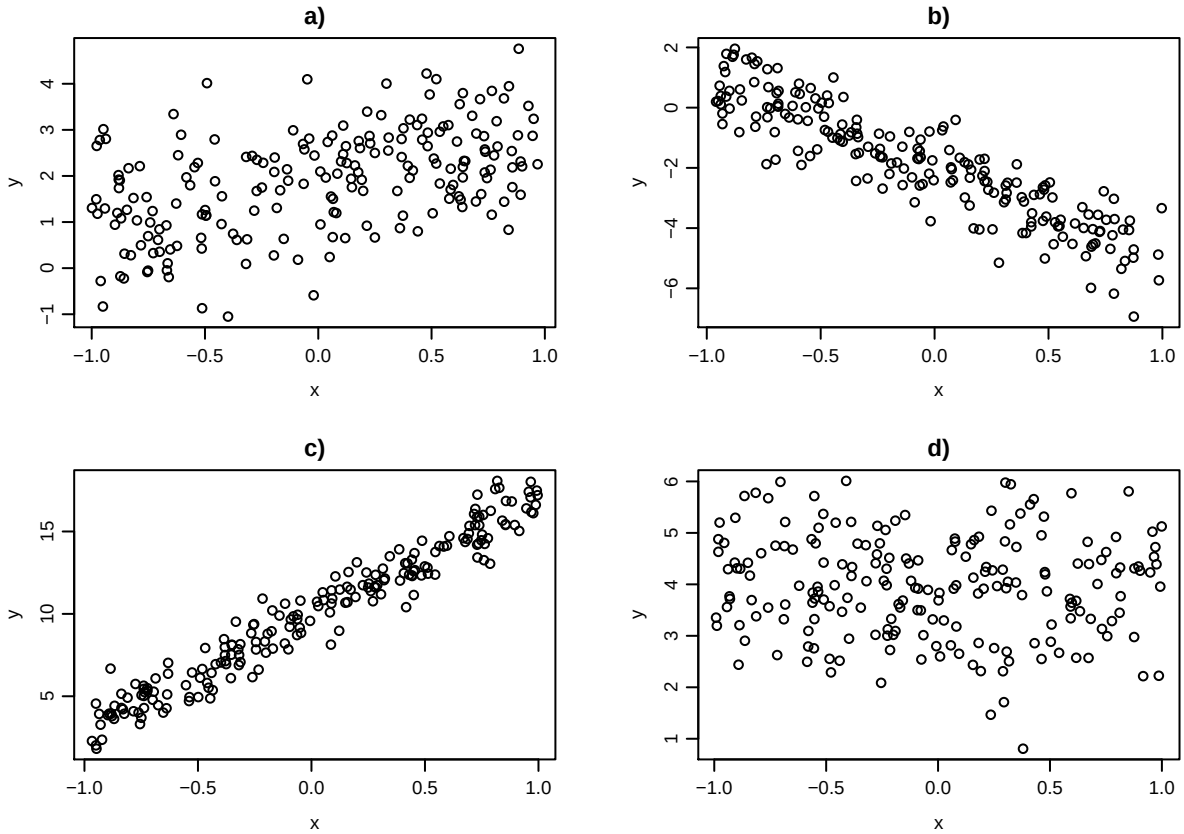
$$LSD_{Bonferroni} = 3.8630\sqrt{2 \cdot 16.788/3} = 12.92$$

———— FACIT-END ————

Continues on page 29

Exercise IX

Below are four scatter plots of y and x observations:



Question IX.1 (24)

Which four correlation coefficients (in the order a), b), c), d)) fits best with the observations in the figure?

1 ☐ 0.9, -0.5, 0.65, 0

2* ☐ 0.5, -0.9, 0.97, 0

3 ☐ 0.5, -0.9, 0.65, 0

4 ☐ 0.5, 0.97, 0, -0.9

5 ☐ 0.97, -0.9, 0, 0.5

———— FACIT-BEGIN ————

Lets go through options:

- 1: It cannot be that a) have 0.9 and c) have 0.65, since the correlation in c) is clearly higher than in a)
- 2: The correlations fits well with the plots
- 3: That a) and c) should have more or less same correlation is not the case, and that b) should have stronger correlation than c) (opposite sign, but stronger) is not the case
- 4: The correlation of b) is not positive
- 5: The correlation of c) is not zero

Hence, only one of the answers fits nicely.

———— FACIT-END ————

Question IX.2 (25)

From inspection of plot `c)` in the figure, which estimates of the parameters in the usual simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$ fits best to those observations in plot `c)`?

1 ☐ $\hat{\beta}_0 = 10$, $\hat{\beta}_1 = 5$ and $\hat{\sigma} = 10$

2 ☐ $\hat{\beta}_0 = 16$, $\hat{\beta}_1 = -5$ and $\hat{\sigma} = 1$

3 ☐ $\hat{\beta}_0 = 10$, $\hat{\beta}_1 = 7$ and $\hat{\sigma} = 10$

4 ☐ $\hat{\beta}_0 = -10$, $\hat{\beta}_1 = -7$ and $\hat{\sigma} = 5$

5* ☐ $\hat{\beta}_0 = 10$, $\hat{\beta}_1 = 7$ and $\hat{\sigma} = 1$

———— FACIT-BEGIN ————

From plot c) it can be seen that the estimates should be

- From the y-values around $x = 0$, it is clear that the intercept estimate $\hat{\beta}_0$ should be around 10
- From the y-values around $x = 0$ (i.e. ≈ 10) and the y-values around $x = 1$ (i.e. slightly above 15, hence ≈ 17), it the most reasonable slope estimate $\hat{\beta}_1$ is around 7 (hence 5 can not directly be ruled out)
- The spread of the y-values seems constant around a fitted straight line. Thus, by considering a normal distribution around this straight line, then we know that around 95% of all points should be within ± 2 standard deviations (σ), hence a $\hat{\sigma} = 1$ seems very reasonable (e.g. if $\hat{\sigma} = 5$ then around $x = 0$ the y-values should be in the range from 0 to 20, and the spread of the points would be much higher!)

Only one options fits to all three estimates.

———— FACIT-END ————

Continues on page 32

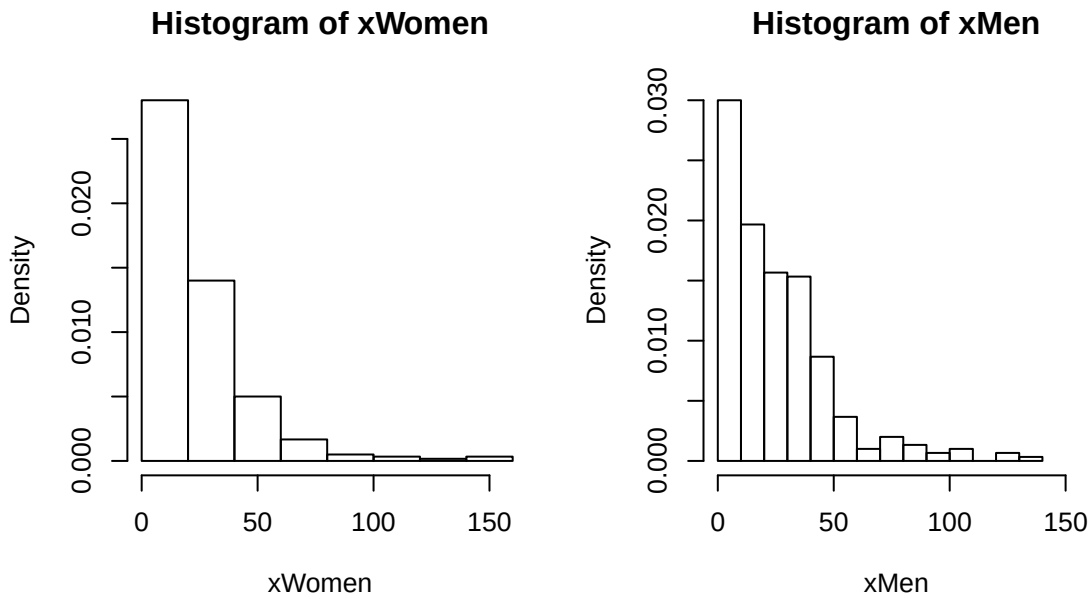
Exercise X

DTU Food monitors and provides analyses of Danish dietary habits. They have done so since 1985 at the request of the Ministry of Environment and Food of Denmark, and the results are obtained on the basis of surveys. The results are used as input for adjustments to the official dietary recommendations and the basis for a political priority of prevention efforts.

In the surveys the respondents were asked how much fish they eat in grams of fish per day. A random sample of 300 female respondents is taken from the dietary survey from 2005 to 2008, and equivalently also a random sample of 300 male respondents from the same dietary survey.

The observations are read into a vector for women `xWomen` and a vector for men `xMen`. The empirical density (density histogram) for the observations for each gender is plotted:

```
par(mfrow=c(1,2), cex=0.8)
hist(xWomen, prob=TRUE, xlim=range(xWomen,xMen))
hist(xMen, prob=TRUE, xlim=range(xWomen,xMen))
```



It is found that the assumption of normal distribution is not met. Therefore no assumption of distribution should be made in the analysis.

Question X.1 (26)

To determine the 95% confidence interval for the mean intake of fish per day for women in the population on the basis of the sample from the 2005-2008 dietary survey, the following R-code is run (of which everything might not necessarily be meaningful):

```r
## Number of simulated samples
k <- 10000

simsamples <- replicate(k, sample(xWomen, replace=TRUE))
simprms <- apply(simsamples, 2, mean)
quantile(simprms, c(0.025,0.975))

##  2.5% 97.5%
##  19.0  24.3

simsamples <- replicate(k, rnorm(length(xWomen),
                                 mean=log(mean(xWomen)), sd=sd(xWomen)))
simprms <- apply(simsamples, 2, mean)
quantile(simprms, c(0.025,0.975))

##  2.5% 97.5%
## 0.348 5.777

simsamples <- replicate(k, sample(xWomen, replace=TRUE))
simprms <- apply(simsamples, 2, median)
quantile(simprms, c(0.005,0.995))

##  0.5% 99.5%
##  11.1  20.6

simsamples <- replicate(k, sample(xMen, replace=TRUE))
simprms <- apply(simsamples, 2, median)
quantile(simprms, c(0.005,0.995))

##  0.5% 99.5%
##  17.0  25.2

simsamples <- replicate(k, runif(100,0,40))
simprms <- apply(simsamples, 2, mean)
quantile(simprms, c(0.025,0.975))

##  2.5% 97.5%
##  17.7  22.3
```

Based on the above outputs, what is then a correct 95% confidence interval for the mean intake of fish per day for women in the population?

1* ☐ $[19.0, 24.3]$

2 ☐ $[0.348, 5.777]$

3 ☐ $[11.1, 20.6]$

4 ☐ $[17.0, 25.2]$

5 ☐ $[17.7, 22.3]$

———— FACIT-BEGIN ————

We have to find the result from the simulation which is correct according to the information we have. Lets check the answer options:

- 1: The `xWomen` sample is re-sampled randomly `k` times (i.e. non-parametric bootstrapping, hence no assumption about the distribution) and for each re-sample the sample mean is calculated, and finally all the correct quantiles of these are found to get the 95% confidence interval
- 2: The re-sampling is from a normal distribution with parameters calculated from the `xWomen` sample (i.e. parametric bootstrapping). This is not in accordance with the information: "the assumption normal distribution is not met"
- 3: The re-sampling is correct, but instead of the sample mean, then the sample median is calculated
- 4: It is the `xMen` sample which is re-sampled, not the `xWomen` sample
- 5: It is just a uniform distribution which is re-sampled, hence it has nothing to do with the intake of fish per day for women

———— FACIT-END ————

Question X.2 (27)

Note, that an assumption of normal distribution of the intake of fish is not met and no transformations are found to be reasonable to meet such an assumption.

In the following, we want to investigate whether men and women eat significant different amounts of fish per day, corresponding to the following hypothesis:

$$H_0 : q_{0.5,\text{male}} = q_{0.5,\text{female}}$$
$$H_1 : q_{0.5,\text{male}} \neq q_{0.5,\text{female}}$$

where $q_{0.5,gender}$ denotes the 50% quantile for the specified gender.

In order to test the hypothesis the following R-code has been run (note that everything not necessarily is meaningful):

```r
t.test(xMen-xWomen)
```

```
##
##  One Sample t-test
##
## data:  xMen - xWomen
## t = 2.0508, df = 299, p-value = 0.04115
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.1553667 7.5300188
## sample estimates:
## mean of x
##  3.842693
```

```r
t.test(xMen, xWomen)
```

```
##
##  Welch Two Sample t-test
##
## data:  xMen and xWomen
## t = 1.9908, df = 597.95, p-value = 0.04695
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.05193859 7.63344693
## sample estimates:
## mean of x mean of y
##  25.37784  21.53515
```

```r
simprms <- replicate(k, median(sample(xMen, replace=TRUE))-
                     median(sample(xWomen, replace=TRUE)))
quantile(simprms, c(0.025, 0.975))
```

```
##   2.5%  97.5%
## 0.0462 9.9185
```

```r
simprms <- replicate(k,
                     mean(rnorm(length(xMen),mean(xMen),sd(xMen))) -
                     mean(rnorm(length(xWomen),mean(xWomen),sd(xWomen))))
quantile(simprms, c(0.025, 0.975))
```

```
##   2.5%  97.5%
## 0.0866 7.6462
```

```r
median(xMen) > median(xWomen)
```

```
## [1] TRUE
```

At a significance level $\alpha = 0.05$, what is then the conclusion of the hypothesis test (both the conclusion and argumentation must be correct)?

1 ☐ Since p-value $= 0.1191 > 0.05$ it cannot be rejected that $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

2 ☐ Since $0 \in [-0.764, 7.191]$ it cannot be rejected that $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

3* ☐ Since $0 \notin [0.564, 7.714]$ it can be rejected that $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

4 ☐ Since $0 \in [-0.776, 7.225]$ it cannot be rejected that $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

5 ☐ Since $\hat{q}_{0.5,\text{male}} > \hat{q}_{0.5,\text{female}}$ it can be rejected that $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

———— FACIT-BEGIN ————

The t-tests are not correct, since they deal with the difference in mean, not in median (i.e. 50% quantile). We have to use the result from the non-parametric bootstrap simulation, where the two samples are re-sampled and the difference in median is calculated for each re-sample:

```
simprms <- replicate(k, median(sample(xMen, replace=TRUE))-
                        median(sample(xWomen, replace=TRUE)))
quantile(simprms, c(0.025, 0.975))
```

which result is a simulated 95% confidence interval. We know that a hypothesis for a value which is not in the confidence interval (with same significance level) will be rejected. Therefore, the hypothesis of no difference in median between the two groups is rejected, since 0 is not in the confidence interval.

———— FACIT-END ————

Question X.3 (28)

Among the 10 official dietary guidelines from the Food Authority is one which states that a person should eat plenty of fish and preferably 350 g per week. This applies to both men and women.

In the following we will examine whether the sample from the dietary survey 2005-2008 provides evidence that dietary recommendation is met. Since dietary survey measured the daily intake of fish, it corresponds (a little simplified) to the following hypothesis:

$$H_0 : \mu = 50$$
$$H_1 : \mu \neq 50$$

Data is not normally distributed, mainly because of the many 0-observations, i.e. respondents who do not eat fish. To test the hypothesis the following R-code has been run (note that all of it may not be meaningful):

```r
## Number of simulated samples
k <- 10000

simsamples <- replicate(k, sample(xMen, replace=TRUE))
simprms <- apply(simsamples, 2, mean)
quantile(simprms, c(0.005, 0.025, 0.975, 0.995))

##  0.5%  2.5% 97.5% 99.5%
##  22.0  22.8  28.1  29.0

simsamples <- replicate(k, sample(xWomen, replace=TRUE))
simprms <- apply(simsamples, 2, sd) - 50
quantile(simprms, c(0.005, 0.025, 0.975, 0.995))

##  0.5%  2.5% 97.5% 99.5%
## -31.7 -30.5 -22.0 -20.7

simsamples <- replicate(k, sample(c(xMen,xWomen), replace=TRUE))
simprms <- apply(simsamples, 2, mean) - 50
quantile(simprms, c(0.005, 0.025, 0.975, 0.995))

##  0.5%  2.5% 97.5% 99.5%
## -29.0 -28.4 -24.6 -23.9

simsamples <- replicate(k, sample(c(xMen,xWomen), replace=TRUE))
simprms <- apply(simsamples, 2, quantile, probs=0.90)
quantile(simprms, c(0.005, 0.025, 0.975, 0.995))

##  0.5%  2.5% 97.5% 99.5%
##  44.7  45.8  54.5  57.4
```

Using a significance level of $\alpha = 0.05$ what is the conclusion on the hypothesis that the dietary recommendation regarding intake of fish is met (both the conclusion and argumentation must be correct)?

1* ☐ Since $0 \notin [-27.6, -23.6]$ the null hypothesis is rejected, hence there is not a significant difference (i.e. $\mu \neq 50$) hence the recommendation is not met

2 ☐ Since $-50 \notin [-28.2, -23.0]$ the null hypothesis is rejected, hence there is not a significant difference (i.e. $\mu \neq 50$) hence the recommendation is not met

3 ☐ Since $50 \in [45.8, 55.0]$ it cannot be rejected that $\mu \neq 50$, hence the recommendation is met

4 ☐ Since $50 \notin [22.5, 29.6]$ the null hypothesis is rejected, hence there is not a significant difference (i.e. $\mu \neq 50$) hence the recommendation is not met

5 ☐ Since $44.9 - 50 = -5.1 < 0$ the null hypothesis is rejected, hence there is not a significant difference (i.e. $\mu \neq 50$) hence the recommendation is not met

———— FACIT-BEGIN ————

We have to check that we take the result which is calculated the right way, the following points should be correct:

- The correct samples: the re-sampling should include both `xMen` and `xWomen`
- The correct statistic is calculated on the re-samples according to the null hypothesis: the difference between the sample mean and 50
- The correct quantiles of the re-sampled test statistic values according to the significance level of $\alpha = 0.05$: the 95% confidence interval, which is formed by the 2.5% and 97.5% quantiles

This leaves option 1 to be answer where the conclusion and arguments are correct according to all three points above.

———— FACIT-END ————

Continues on page 39

Exercise XI

A new scanner for measuring the mass of the muscles in the body has been developed. It is much easier and faster to use compared to the otherwise available scanners. It is tested in an experiment to find out if it gets the similar results as the normally used scanner. For the experiment 20 randomly selected women aged 20 to 40 years have been scanned with both scanners.

The measured muscle mass in kg are read into R, such that the order of the women is the same in each vector:

```
## Sample from the new scanner
x1 <- c(37.6, 31.3, 22.9, 27.1, 41.8, 23.3, 24.5, 24.6, 32.1, 23.8, 33.9, 37.7,
        22.5, 38.6, 31.8, 21.0, 32.2, 17.1, 32.6, 15.5)
## Sample from the old scanner
x2 <- c(35.9, 28.7, 27.9, 29.8, 46.8, 24.2, 28.0, 23.7, 35.2, 26.4, 36.0, 40.9,
        24.8, 42.1, 32.5, 23.7, 36.7, 19.2, 37.7, 16.3)
```

and the following R code is run:

```
(mean(x1)-mean(x2)) + c(-1,1) * qt(0.975, df=38) * sd(x2-x1)/sqrt(40)

## [1] -2.9228 -1.5372

t.test(x1, x2)

##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -0.916, df = 37.8, p-value = 0.37
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.1591  2.6991
## sample estimates:
## mean of x mean of y
##    28.595    30.825

t.test(x1, x2, paired=TRUE)

##
##  Paired t-test
##
## data:  x1 and x2
## t = -4.61, df = 19, p-value = 0.00019
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -3.2429 -1.2171
## sample estimates:
## mean of the differences
##                   -2.23
```

```
t.test(x2-mean(x1))
```

```
##
##  One Sample t-test
##
## data:  x2 - mean(x1)
## t = 1.25, df = 19, p-value = 0.23
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.5043  5.9643
## sample estimates:
## mean of x
##      2.23
```

```
t.test(x1-mean(x2))
```

```
##
##  One Sample t-test
##
## data:  x1 - mean(x2)
## t = -1.35, df = 19, p-value = 0.19
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -5.6965  1.2365
## sample estimates:
## mean of x
##     -2.23
```

Question XI.1 (29)

What is the correct 95% confidence interval for the mean difference in measurement of muscle mass between the old and the new scanner (here rounded to three significant digits)?

1 ☐ [-2.92, -1.54]

2 ☐ [-7.16, 2.70]

3* ☐ [-3.24, -1.22]

4 ☐ [-1.50, 5.96]

5 ☐ [-5.70, 1.24]

———— FACIT-BEGIN ————

This is a paired setup since each women is measure first with the one scanner and then the other. Therefore, we could take the difference and calculate the confidence interval using the differences directly

```
t.test(x1-x2)

##
##  One Sample t-test
##
## data:  x1 - x2
## t = -4.61, df = 19, p-value = 0.00019
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -3.2429 -1.2171
## sample estimates:
## mean of x
##     -2.23
```

which is exactly the same as the we get with the option `paired=TRUE`.

———— FACIT-END ————

Question XI.2 (30)

The accuracy of a scan depends on how much the person to be scanned is moving. In the new scanner the person is not fastened, so it is of interest to find out how large the variation of the measurements is. To investigate this the 20 women were scanned two times with the new scanner each and the difference in muscle mass (X_Δ) for each woman between the 2 scans were measured to:

i	1	2	3	4	5	6	7	8	9	10
$x_{\Delta,i}$	-0.62	1.12	0.24	2.07	-2.91	2.02	0.36	0.43	-1.77	-0.18
i	11	12	13	14	15	16	17	18	19	20
$x_{\Delta,i}$	1.02	0.85	0.43	1.39	2.82	-4.03	2.84	2.8	1.36	-0.07

The sample mean and standard deviation were calculated to

$$\bar{x}_\Delta = 0.509$$
$$s_{x_\Delta} = 1.82$$

Which of the following is a correct 99% confidence interval for the standard deviation of the measured muscle mass of the new scanner?

1 ☐ $0.509 \pm 2.86 \cdot \frac{1.35}{\sqrt{20}} = [-0.35, 1.37]$

2 ☐ $\left[\frac{20 \cdot 1.82^2}{38.6}, \frac{20 \cdot 1.82^2}{6.84}\right] = [1.72, 9.69]$

3 ☐ $0.509 \pm 2.09 \cdot \frac{1.35}{\sqrt{20}} = [-0.12, 1.14]$

4 ☐ $0.509 \pm 2.86 \cdot \frac{1.82}{\sqrt{19}} = [-0.69, 1.70]$

5* ☐ $\left[\sqrt{\frac{19 \cdot 1.82^2}{38.6}}, \sqrt{\frac{19 \cdot 1.82^2}{6.84}}\right] = [1.28, 3.03]$

———— FACIT-BEGIN ————

The confidence interval for the standard deviation is given in Method 3.19 with the formula

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

and then we just need to insert the values and find the quantiles in the χ^2 distribution by

```
qchisq(0.005, 19)

## [1] 6.844

qchisq(0.995, 19)

## [1] 38.582
```

to find that

$$\left[\sqrt{\frac{19 \cdot 1.82^2}{38.6}}, \sqrt{\frac{19 \cdot 1.82^2}{6.84}}\right] = [1.28, 3.03]$$

is the correct interval.

———— FACIT-END ————

THE EXAM IS FINISHED. Enjoy the late summer!

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

Written examination: 13. December 2016

Course name and number: **Introduction to Statistics (02323 and 02402)**

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 18 exercises. To answer the questions you need to fill in the prepared 30-question multiple choice form (on three 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	III.1	III.2	IV.1	IV.2	V.1	VI.1	VII.1	VIII.1
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer	2	3	2	4	4	4	3	4		4

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

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

The questionnaire contains 41 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

Archaeopteryx is a genus of bird-like dinosaurs that is transitional between non-avian feathered dinosaurs and modern birds. Assume that we have data from 6 fossils of Archaeopteryx including measurements of the length of the thigh bone (femur) and the upper arm bone (humerus) as shown in the table below.

Femur	38	46	56	59	64	74
Humerus	41	50	63	70	71	76

Data can be loaded into R by:

```
femur = c(38,46,56,59,64,74)
humerus = c(41,50,63,70,71,76)
```

Question I.1 (1)

Archaeologists have long believed that there should be a linear relationship between the length of the femur and length of humerus of extinct animals such as Archaeopteryx. What conclusion can be made by analyzing the above data when the significance level $\alpha = 0.05$ is used?

1 ☐ There is reason to assume a linear relationship, as the length of bones in animals are always positively correlated.

*2 ☐ There is reason to assume a linear relationship, with p-value for the relevant test being 0.0013.

3 ☐ There is reason to assume a linear relationship, with p-value for the relevant test being 0.0780.

4 ☐ There is no reason to assume a linear relationship, with p-value for the relevant test being 0.0033.

5 ☐ There is no reason to assume a linear relationship, with p-value for the relevant test being 0.0780.

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

We need to test if there is a significant correlation, which we can do by typing in the values in R and fit a simple linear regression model. This we can do by testing if the slope (β_1) is significantly different from zero, i.e.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

which is equivalent to the hypothesis

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

where ρ is the correlation, see Section 5.6. We type in the numbers in R and fit a simple linear regression model

```
femur <- c(38,46,56,59,64,74)
humerus <- c(41,50,63,70,71,76)
summary(lm(humerus ~ femur))

##
## Call:
## lm(formula = humerus ~ femur)
##
## Residuals:
##      1      2      3      4      5      6
## -2.106 -1.353  1.338  5.246  1.092 -4.217
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9332     7.3951   0.532  0.62299
## femur         1.0309     0.1289   7.998  0.00133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.693 on 4 degrees of freedom
## Multiple R-squared:  0.9411,Adjusted R-squared:  0.9264
## F-statistic: 63.96 on 1 and 4 DF,  p-value: 0.001325
```

We find that the p-value for the test is 0.00133, which is much below $\alpha = 0.05$. Thus the null hypothesis is rejected and we conclude that there is a linear relationship.

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

Continues on page 4

Exercise II

A study aims to investigate whether intake of a natural product affects weight. The study should include 10 subjects (men with similar weight). The weight change D_i $(i = 1, \ldots, 10)$ after one month of use of the natural product is recorded. It is of interest to test if the weight change can be assumed to be zero, i.e. to test the hypothesis $H_0 : \mu_D = 0$ against the alternative $H_1 : \mu_D \neq 0$. It is decided to apply the significance level $\alpha = 0.05$.

Question II.1 (2)

Assuming that the standard deviation of weight change is $\sigma = 1$ kg, what is the power for detecting an actual weight change of at least 1 kg? (Hint: the function `power.t.test` in R can be useful here.)

1 ☐ 50.0%

2 ☐ 69.3%

*3 ☐ 80.3%

4 ☐ 89.7%

5 ☐ 99.3%

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

In order to find the power of the test to detect an actual change of at least 1 kg, then the recommended R function can be used, we just need to give it the four parameters, see Example 3.67:

- Sample size $n = 10$
- Change to detect $\delta_0 = 1$
- Assumed standard deviation of the population $\sigma = 1$
- Significance level $\alpha = 0.05$

Further, we have to tell it that it is a one-sample test and the alternative is two-sided.

Then the function calculates the power

```
power.t.test(n=10, delta=1, sd=1, sig.level=0.05, type="one.sample", alternative="two.si
```

```
##
##      One-sample t test power calculation
##
##               n = 10
##           delta = 1
##              sd = 1
##       sig.level = 0.05
##           power = 0.8030962
##     alternative = two.sided
```

which is 80.3%.

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

Continues on page 6

Exercise III

It is believed that the amount of cholesterol in chicken eggs, X, is normally distributed with mean $\mu = 200$ mg and standard deviation $\sigma = 15$ mg, i.e. $X \sim N(200, 15^2)$.

Question III.1 (3)

What is the proportion of chicken eggs having an amount of cholesterol higher than 205 mg?

1 ☐ $P(X > 205) = 0.631$

*2 ☐ $P(X > 205) = 0.369$

3 ☐ $P(X > 205) = 0.491$

4 ☐ $P(X > 205) = 0.394$

5 ☐ $P(X > 205) = 0.605$

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

We need to calculate the proportion (or the probability of drawing a random egg from the population) above 205. Remember

$$P(X > 205) = 1 - P(X \leq 205)$$

where $P(X \leq 205)$ is the cumulated distribution function (cdf) for a normal distribution and that we can get from R by

```
1 - pnorm(q=205, mean=200, sd=15)
```

```
## [1] 0.3694413
```

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

Question III.2 (4)

Industrial kitchens may buy cartons of eggs, where the content, Y, in a carton corresponds to the combined content of 100 eggs, i.e. the total content of cholesterol in a carton is $Y = \sum_{i=1}^{100} X_i$. The content of the 100 eggs can be assumed independent from each other.

You buy a carton of eggs, corresponding to buying 100 eggs. Which of the following R commands gives the probability that the total cholesterol, Y, is higher than 20.5 g (note that 200 mg is 0.2 g)?

1 ☐ `pnorm(q=100*0.205, mean=100*0.200, sd=100*0.015)`

2 ☐ `1-pnorm(q=100*0.200, mean=100*0.205, sd=sqrt(100*0.015*0.015))`

3 ☐ `pnorm(q=100*0.205, mean=100*0.200, sd=100*100*0.015)`

*4 ☐ `1-pnorm(q=100*0.205, mean=100*0.200, sd=sqrt(100*0.015*0.015))`

5 ☐ `pnorm(q=100*0.205, mean=100*0.200, sd=sqrt(100*0.015*0.015))`

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

We need to use the Theorem 2.54: Mean and variance of linear combinations. We need to find the mean

$$\begin{aligned}\mathrm{E}(Y) &= \mathrm{E}\left(\sum_{i=1}^{100} X_i\right) \\ &= \mathrm{E}(X_1 + X_2 + \cdots + X_n) \\ &= \mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n) \\ &= 100 \cdot \mathrm{E}(X) = 100 \cdot 200 \text{ mg} = 100 \cdot 0.200 \text{ g}\end{aligned}$$

and the variance

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{Var}\left(\sum_{i=1}^{100} X_i\right) \\ &= \mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) \\ &= 100 \cdot \mathrm{Var}(X) = 100(\cdot 15 \text{ mg})^2 = 100 \cdot (0.015 \text{ g})^2\end{aligned}$$

which then gives the standard deviation

$$\sigma_Y = \sqrt{100 \cdot (0.015 \text{ g})^2} = \sqrt{100 \cdot 0.015 \cdot 0.015} \text{ g}$$

Finally, the probability we need to calculate is

$$P(Y > 20.5) = 1 - P(Y \leq 20.5)$$

which in R is written as: `1-pnorm(q=100*0.205, mean=100*0.200, sd=sqrt(100*0.015*0.015))`.

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

Continues on page 8

Exercise IV

We consider a binomial random variable Y where $n = 100$ and $p = 0.45$.

Question IV.1 (5)

Calculate $P(Y > 40)$:

1 ☐ 0.183

2 ☐ 0.971

3 ☐ 0.420

*4 ☐ 0.817

5 ☐ 0.866

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

We need first to remember

$$P(Y > 40) = 1 - P(Y \leq 40)$$

Since we know the parameters of the distribution, we can find this probability in R as

```
1 - pbinom(q=40, size=100, prob=0.45)

## [1] 0.8169431
```

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

Question IV.2 (6)

We define a new random variable X so that $X = k \cdot Y$, where the constant k is given by $k = 2$ and Y is binomial distributed random variable with $n = 100$ and $p = 0.45$. Please state the variance of the random variable X:

1 ☐ $\text{Var}(X) = \text{Var}(k \cdot Y) = k + n \cdot p(1-p) = 26.75$

2 ☐ $\text{Var}(X) = \text{Var}(k \cdot Y) = k^2 \cdot n^2 \cdot p^2(1-p)^2 = 49.50^2$

3 ☐ $\text{Var}(X) = \text{Var}(k \cdot Y) = k^2 \cdot n^2 \cdot p(1-p) = 9900$

*4 ☐ $\text{Var}(X) = \text{Var}(k \cdot Y) = k^2 \cdot n \cdot p(1-p) = 99.00$

5 ☐ $\text{Var}(X) = \text{Var}(k \cdot Y) = k \cdot n \cdot p(1-p) = 49.50$

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

We need to use the Theorem 2.54, combined with Theorem 2.21. This gives us a formula for variance of a linear function of a random variable and the variance of a binomial distributed random variable

$$\text{Var}(X) = \text{Var}(k \cdot Y) = k^2\text{Var}(Y) = k^2 \cdot n \cdot p(1-p) = 2^2 \cdot 100 \cdot 0.45 \cdot (1-0.45) = 99.$$

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

Continues on page 10

Exercise V

We consider an exponentially distributed random variable X with parameter β. The distribution function is given by $F(X \leq x) = 1 - e^{-x/\beta}$, where $x > 0$ and $\beta > 0$. Please note that the mean value of X equals β.

Question V.1 (7)

Please state the median of X:

1 ☐ The median of X becomes $0.5 \cdot 2 \cdot \beta$

2 ☐ The median of X becomes $0.5^2 \cdot \beta$.

*3 ☐ The median of X becomes $\log(2) \cdot \beta$ (where log is the natural logarithm)

4 ☐ The median of X becomes $\log(\frac{1}{2}) \cdot \beta^2$ (where log is the natural logarithm)

5 ☐ The median of X becomes $2 \cdot \beta$

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

The median is the quantile where exactly half of the probability mass is below, so we can set the cdf equal to 0.5 and then solve for x

$$\begin{aligned} P(X \leq x) &= 0.5 \Leftrightarrow \\ 1 - e^{-x/\beta} &= 0.5 \Leftrightarrow \\ e^{-x/\beta} &= 0.5 \Leftrightarrow \\ \frac{1}{e^{x/\beta}} &= 0.5 \Leftrightarrow \\ e^{x/\beta} &= \frac{1}{0.5} \Leftrightarrow \\ \frac{x}{\beta} &= \log 2 \Leftrightarrow \\ x &= \log 2 \cdot \beta. \end{aligned}$$

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

Continues on page 11

Exercise VI

A biologist is interested in examining the effects of three different diets (A, B, C) for cultivating tiger shrimps. She purchases 24 uniform larvae from a hatchery for the experiment. Each larva is placed in its own container, and it is determined by random which diet that should be given, such that each diet is tested on 8 different larvae. The larvae grow to become tiger shrimps, and after completing the study period the weight of the shrimps is measured, Y_{ij} (in grams). Since the weight can be assumed to follow a normal distribution, the following model is applied

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}.$$

In the model α_i describes the effect of diet i ($i = 1, 2, 3$). Finally, μ is the average and ε_{ij} is the model residuals which are assumed normally distributed with mean 0 and standard deviation σ_ε. An ANOVA of the above model is given below, and it is seen that diet is statistically significant.

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
Diet       2  44.67 22.3350  8.9221 0.001568 **
Residuals 21  52.57  2.5033
```

Question VI.1 (8)

Beforehand there was an interest in comparing the mean value of diet A and diet C. Their estimated mean values are $\hat{\mu}_A = 12.7251$ and $\hat{\mu}_C = 15.7251$, respectively. Please provide a 95 % confidence interval for the mean difference in weight between diet A and diet C.

1 ☐ $-3.000 \pm 2.119\sqrt{52.488^2(\frac{1}{12} + \frac{1}{12})}$

2 ☐ $-3.000 \pm 1.960\sqrt{64.624(\frac{1}{12} + \frac{1}{12})}$

3 ☐ $-3.000 \pm 1.960\sqrt{44.670(\frac{1}{3} + \frac{1}{3})}$

*4 ☐ $-3.000 \pm 2.080\sqrt{2.503(\frac{1}{8} + \frac{1}{8})}$

5 ☐ $-3.000 \pm 2.080\sqrt{52.570(\frac{1}{4} + \frac{1}{4})}$

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

We must calculate a pre-planned confidence interval as described in Method 8.9:

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)},$$

We can read most values from the ANOVA R output, and the t-quantile can be found as

```
qt(p=0.975, df=21)

## [1] 2.079614
```

When inserting values into the formula it becomes the one shown in answer 4.

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

Continues on page 13

Exercise VII

The exercise is no longer a part of the curriculum

Question VII.1 (9)

The question is no longer a part of the curriculum

Continues on page 14

Exercise VIII

A study aims at comparing the wear resistance of two different kinds of rubber (A and B) used as material for shoe soles. The study includes 100 school children aged 8-10 years. Each child receives a pair of shoes where the sole of one shoe is made of material A, while the sole of the other shoe is made of material B. For each pair of shoes, it is decided by randomization whether material A should be on shoe to the right or to the left. The children use the shoes every day for 3 months, and after the experiment has been completed the wear (in mm) on each shoe is measured.

Question VIII.1 (10)

If it can be assumed that the measured wear is continuous and normally distributed for each kind of rubber, please specify which of the following statistical tests should be applied, if you want to test whether the materials A and B are equal with respect to wear:

1 ☐ A contingency table analysis

2 ☐ An F test comparing two variances

3 ☐ A usual (non-paired) t-test

*4 ☐ A paired t-test

5 ☐ A one-way ANOVA

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

We must test for a difference in mean between two groups, hence a two-sample t-test. Now the question is if it is a paired setup or not. Since each children have both a sole in Material A and a sole in Material B, and each pair is exposed to the same wear (although there could be a difference for each child between right and left, however this is compensated by randomizing the left and right material). Thus, this makes a paired setup and we can take the difference between the soles for each child and use a one-sample test. Hence, we should use a paired t-test. See Section 3.2.3 for more information.

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

Question VIII.2 (11)

It turns out that the null hypothesis is accepted, i.e. it is concluded that the two materials wear out equally. Instead the researchers calculate for each child in the study the average wear for the pair of shoes. It is of interest to analyze, using a standard t-test, if boys and girls wear the shoes equally, or alternatively, if there is a difference in wear between gender (two-sided test). A total of 50 girls and 50 boys were included in the experiment.

As the wear can be assumed normally distributed within each gender with equal variance, we get the usual test statistics $t_{obs} = 2.23$ with 98 degrees of freedom. The p-value becomes:

1 ☐ The p-value becomes 0.014

*2 ☐ The p-value becomes $2 \cdot 0.014$

3 ☐ The p-value becomes $2 \cdot 0.05$

4 ☐ The p-value becomes 0.23

5 ☐ The p-value becomes $1 - 0.23$

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

We have the observed statistic t_{obs}, which under the null hypothesis is t-distributed, and the degrees of freedom are given, so we can simply calculate the p-value by

```
2 * (1 - pt(2.23, df=98))

## [1] 0.02802943
```

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

Continues on page 16

Exercise IX

A course at a university is offered each semester typically with more than 300 students taking the exam. Examination results for 280 students who have passed the course at the previous exam is given in the table below. For example, the tables shows that 24 students got the grade 12. The distribution of the 280 grades is considered in the next 4 questions.

Grade	02	4	7	10	12	In total
Count	22	78	84	72	24	280

The data (grades) can be loaded into R by:

```
grades = rep(x=c(2,4,7,10,12), times=c(22,78,84,72,24))
```

Question IX.1 (12)

Use the central limit theorem to determine a 95% confidence interval for the mean grade based on the students who have passed the exam. (It is important in this question that the grades are perceived numerically, eg. 02 corresponds to the number 2, etc.).

1 ☐ [6.51 ; 7.43]

*2 ☐ [6.62 ; 7.32]

3 ☐ [4 ; 10]

4 ☐ [5.12 ; 8.67]

5 ☐ [5.99 ; 8.72]

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

According to the central limit theorem (Theorem 3.14), we know that even if this data is not normal distributed, then if we have $n > 30$ observations, the standardized sample mean follows a standard normal distribution and we can use the t-distribution to calculate a confidence interval. Therefore we load the sample into R and either use the in-built function or calculate the confidence interval using the formula

```
## Read the data
grades = rep(x=c(2,4,7,10,12), times=c(22,78,84,72,24))
## Use the inbuilt function
t.test(grades)
```

```
##
##  One Sample t-test
##
## data:  grades
## t = 38.972, df = 279, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  6.619297 7.323560
## sample estimates:
## mean of x
##  6.971429

## Use the formula
mean(grades) + c(-1,1) * qt(0.975, df=280-1) * sqrt(sd(grades)^2/280)

## [1] 6.619297 7.323560
```

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

Continues on page 18

Question IX.2 (13)

You now want to test whether the proportion of students who have passed the exam with a grade of '7' or higher can be assumed to be 65%, which has been an objective in designing the grading scale. We denote this proportion p_{7+}. From the table in the previous question it is seen that 180 students out of 280 got the grade '7' or higher.

Determine the p-value when we wish to test $H_0 : p_{7+} = 0.65$ against $H_1 : p_{7+} \neq 0.65$:

*1 ☐ 0.8021

2 ☐ $1.745 \cdot 10^{-6}$

3 ☐ $8.725 \cdot 10^{-7}$

4 ☐ 0.5989

5 ☐ 0.4011

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

This is a single proportion hypothesis test. Using Theorem 7.10

$$z_{\text{obs}} = \frac{180 - 280 \cdot 0.65}{\sqrt{280 \cdot 0.65 \cdot 0.35}} = -0.2506,$$

with which we can calculate the p-value by

$$p\text{-value} = 2P(Z > |-0.2506|) = 2 \cdot (1 - P(Z \leq 0.2506)),$$

using R

```
2 * (1 - pnorm(0.2506))

## [1] 0.8021234
```

Or we could simply calculate it directly in R

```
prop.test(x=180, n=280, p=0.65, alternative="two.sided", correct=FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  180 out of 280, null probability 0.65
## X-squared = 0.062794, df = 1, p-value = 0.8021
## alternative hypothesis: true p is not equal to 0.65
```

```
## 95 percent confidence interval:
##  0.5851475 0.6967000
## sample estimates:
##         p
## 0.6428571
```

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

Question IX.3 (14)

A student is interested in analyzing the data in more detail, and run the following code using the 280 grades stored in the vector `grades`

```
k = 100000
samples = replicate(k, sample(grades, replace = TRUE))
simval = apply(samples, 2, sd)
resultater = quantile(simval, c(0.025,0.975))
```

Please state which numerical result that has been calculated in the vector `resultater`:

*1 ☐ A 95% confidence interval for the standard deviation of the grades (non-parametric bootstrap)

2 ☐ A 95% confidence interval for the distribution of the grades (parametric bootstrap)

3 ☐ A 95% prediction of the median of the grades (non-parametric bootstrap)

4 ☐ A 95% prediction for the standard error of the grades (parametric bootstrap)

5 ☐ A 95% confidence interval for 75% percentile of the grades (parametric bootstrap)

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

It is clear that it is a simulation results, namely found using a bootstrapping method. First, the sample is re-sampled simply by drawing randomly from the sample with replacement, thus there is no assumption about the distribution (i.e. non-parametric). Second, the standard deviation is calculated for all the resampled samples, and from these the 2.5% and 97.5% quantiles are found, therefore: A non-parametric 95% confidence interval for the standard deviation has been bootstrapped.

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

Continues on page 20

Question IX.4 (15)

You now want to examine if the distribution of the grades is the same for men and women. The distribution of grades by sex is shown in the table below.

Grades	02	4	7	10	12	In total
Men	14	47	59	47	18	185
Women	8	31	25	25	6	95

Please calculate the expected number of men with the grade '7' in the case where the grade distribution is assumed equal for men and women (i.e. assuming the null hypothesis):

*1 ☐ 55.5

2 ☐ 59

3 ☐ 47.57

4 ☐ 28.5

5 ☐ 42

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

We must calculate the expected value in cell (1,3) under the null hypothesis that the distribution is the same between the genders. According to Method 7.3, the expected proportion of men is

$$\frac{x}{n} = \frac{185}{185+95} = 0.6607,$$

which we multiply with the total number of observations with the grade 7

$$(59+25) \cdot \frac{185}{185+95} = 55.5.$$

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

Continues on page 21

Exercise X

A discrete random variable X, is used to describe the number of events during a time interval. X has the density function on the familiar form: $P(X = x) = \frac{2^x}{x!}e^{-2}$, for $x \geq 0$.

Question X.1 (16)

What is the mean of X?

1 ☐ $\frac{1}{2}$

2 ☐ $\log(2)$ (where log is the natural logarithm)

*3 ☐ 2

4 ☐ π

5 ☐ 2^2

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

We recognize that the distribution used for characterizing number of events per time interval is the Poisson distribution, and we recognize the pdf from Definition 2.27. Then we can see that $\lambda = 2$ and we are asked about the mean, which (see Theorem 2.28) is simply equal to λ.

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

Continues on page 22

Exercise XI

A study aims at comparing cognitive abilities of 3 groups of children. The groups consist of a) children with Tourette's Syndrome (TS), b) children with ADHD and c) children without any of these diagnoses (Control).

In the study, each child is asked to solve a sequence of tasks on a computer and the average reaction time, R_i (milliseconds) is recorded for each child. The study included 17 children with TS, 13 with ADHD and 20 controls, i.e. a total of $n = 50$ children.

When analyzing the data from the experiment it has been assumed that the reaction time R_i is normally distributed for each group with constant variance, σ_E^2. In order to compare if the mean reaction time is the same for the three groups (TS, ADHD and controls) the following ANOVA table is provided

```
Analysis of Variance Table

Response: reactiontime
          Df Sum Sq Mean Sq F value    Pr(>F)
group      A 485848  242924 D          .976e-07 ***
Residuals  B 542563  C
---
```

It is seen, however, that not all numbers are given in the ANOVA table, but some are only shown by the symbols A, B, C and D. These 4 symbols are part of the solution to the next question.

Question XI.1 (17)

Which distribution does the value D follow if the mean reaction time is the same for all three groups (TS, ADHD and Control)?

1 ☐ $F(A, A+B)$ i.e. an F-distribution with degrees of freedom A and $A+B$ from the ANOVA table

2 ☐ $F(C, B)$ i.e. an F-distribution with degrees of freedom C and B from the ANOVA table

*3 ☐ $F(A, B)$ i.e. an F-distribution with degrees of freedom A and B from the ANOVA table

4 ☐ $F(A, C)$ i.e. an F-distribution with degrees of freedom A and C from the ANOVA table

5 ☐ $F(B, A)$ i.e. an F-distribution with degrees of freedom B and A from the ANOVA table

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

It is recognized as a one-way ANOVA, since there is one factor `group`. According to Theorem 8.6 we know that the test-statistic follows an F-distribution under the null hypothesis with the degrees of freedoms in the R output.

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

Continues on page 24

Question XI.2 (18)

What is the conclusion from the ANOVA table if the significance level $\alpha = 0.05$ is applied?

1 ☐ We must reject the hypothesis that the mean reaction time of the control children equals the mean reaction time for children with TS or ADHD.

2 ☐ We must reject the hypothesis that the variance of the reaction time of the control children equals the variance of the reaction time for children with TS or ADHD.

3 ☐ We can prove that the variance of the reaction time is the same for all three groups since the p-value equals $0.976 \cdot 10^{-7}$.

*4 ☐ We must reject the hypothesis that the average reaction time is the same for all three groups since the p-value equals $0.976 \cdot 10^{-7}$.

5 ☐ We can not reject the hypothesis that the average reaction time is the same for all three groups.

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

The null hypothesis is that the mean in equal in the three groups, hence that the mean reaction time is equal in the three groups. The p-value is $0.976 \cdot 10^{-7}$, which is way under $\alpha = 0.05$, and therefore the null hypothesis is rejected.

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

Question XI.3 (19)

Subsequently it is decided to examine whether the age of the children $x_{1,i}$ influences the reaction time, Y_i. Another experiment was conducted in which $n = 12$ children with no diagnosis (control), but with different ages solved the sequence of tasks and the average reaction time was measured. The model $Y_i = \beta_0 + \beta_1 \cdot x_{1,i} + \varepsilon_i$ is applied in order to examine the association between age and reaction time. In the model the residuals ε_i are assumed i.i.d. normally distributed with constant variance, hence $\varepsilon_i \sim N(0, \sigma^2)$. You get the following output for the new experiment:

```
Call:
lm(Reactiontime ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max
-54.520 -35.522   4.268  27.160  51.949
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   933.15     153.23   6.090 0.000117 ***
Age           -41.05      15.36  -2.672 0.023400 *
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 37.72 on 10 degrees of freedom
Multiple R-squared:  0.4166,Adjusted R-squared:  0.3583
F-statistic: 7.141 on 1 and 10 DF,  p-value: 0.0234
```

Using the numbers in the output from the model please provide the test statistics related to the hypothesis $H_0 : \beta_1 = 0$:

*1 ☐ -2.672

2 ☐ 153.23

3 ☐ -41.05

4 ☐ 37.72

5 ☐ 0.4166

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

The observed test statistic t_{obs} for the test if slope β_1 is zero, can be found under `t value` in the summary output in the row of the explanatory variable, here `Age`.

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

Question XI.4 (20)

Using the analysis result from the previous question find the estimate of the correlation coefficient, $\hat{\rho}$, between Reaction time (Y_i) and Age ($x_{1,i}$):

1 ☐ $\hat{\rho} = -\sqrt{0.3583}$

*2 ☐ $\hat{\rho} = -\sqrt{0.4166}$

3 ☐ $\hat{\rho} = \sqrt{0.3583}$

4 ☐ $\hat{\rho} = 0.4166$

5 ☐ $\hat{\rho} = -\sqrt{0.3583/0.4166}$

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

See Section 5.6. We know the relation between the proportion of explained variance and the sample correlation coefficient.

We take the root of the $\hat{r}^2$ value (explained variance) and the sign of the estimated slope, and get

$$\hat{\rho} = \text{sign}(\hat{\beta}_1)\sqrt{\hat{r}^2} = -\sqrt{0.4166}.$$

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

Continues on page 27

Question XI.5 (21)

A critique of the model $Y_i = \beta_0 + \beta_1 \cdot x_{1,i} + \varepsilon_i$ used in the previous question is that it does not account for whether the answer is correct or not on the individual questions, but simply the reaction time is recorded.

It is decided to expanded model to the following: $Y_i = \beta_0 + \beta_1 \cdot x_{1,i} + \beta_2 \cdot x_{2,i} + \varepsilon_i$, where $x_{2,i}$ is the number of correct answers in the sequence of questions (the remaining variables are defined as they were in the previous question). Based on a new study including 12 different children we obtain the results:

```
Call:
lm(Reactiontime ~ Age + Correct)

Residuals:
    Min      1Q  Median      3Q     Max
-39.958 -24.407   6.917  12.897  42.297

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  892.633    123.203   7.245 4.84e-05 ***
Age          -48.104     15.081  -3.190   0.0110 *
Correct        5.310      1.765   3.009   0.0147 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 28.73 on 9 degrees of freedom
Multiple R-squared:  0.5908,Adjusted R-squared:  0.4999
F-statistic: 6.498 on 2 and 9 DF,  p-value: 0.01793
```

Provide the estimates $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\sigma}^2$:

1 ☐ $\hat{\beta}_1 = -48.104$, $\hat{\beta}_2 = -5.310$ and $\hat{\sigma}^2 = 28.73$

2 ☐ $\hat{\beta}_1 = 892.633$, $\hat{\beta}_2 = -48.104$ and $\hat{\sigma}^2 = 28.73$

3 ☐ $\hat{\beta}_1 = 892.633$, $\hat{\beta}_2 = -48.104$ and $\hat{\sigma}^2 = 5.310$

4 ☐ $\hat{\beta}_1 = -48.104$, $\hat{\beta}_2 = 5.310$ and $\hat{\sigma}^2 = 892.633$

*5 ☐ $\hat{\beta}_1 = -48.104$, $\hat{\beta}_2 = 5.310$ and $\hat{\sigma}^2 = 28.73^2$

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

We find the two parameter estimates $\hat{\beta}_1$ (`Age`) and $\hat{\beta}_2$ `Correct` under `Estimate` in the printed results. The estimate of the variance of the error ($\varepsilon_i \sim N(0, \sigma^2)$) at `Residual standard error`.

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

Continues on page 29

Exercise XII

An engineer is studying a process Y which can be expressed by $Y = U/B$. It can be assumed that U and B are independent random variables. The engineer has 20 pairwise measurements of U and B stored as vectors in the statistical program R and these are referred to as `uobs` and `bobs`,respectively.

Question XII.1 (22)

The engineer would like to calculate a 95% confidence interval for the variance of Y, i.e. σ_Y^2 using non-parametric bootstrap. Which of the following chunks of code in R is most appropriate?

1 ☐
```
samples = replicate(10000,sample(uobs/bobs,replace=FALSE))
results = apply(samples,1, var)
quantile(results, c(0.025, 0.975))
```

*2 ☐
```
samples = replicate(10000,sample(uobs/bobs,replace=TRUE))
results = apply(samples,2, var)
quantile(results, c(0.025, 0.975))
```

3 ☐
```
samples = replicate(10000,sample(var(uobs)/var(bobs),replace=TRUE))
results = apply(samples,2, var)
quantile(results, c(0.025, 0.975))
```

4 ☐
```
samples = replicate(10000,sample(uobs/bobs,replace=TRUE))
results = apply(samples,1, var)
quantile(results, c(0.95))
```

5 ☐
```
samples = replicate(10000,sample(uobs/bobs,replace=FALSE))
results = apply(samples,2, var)
quantile(results, c(0.025, 0.975))
```

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

We want to find the code which is right and first we see that they are all non-parametric (since they all use the `sample` command instead of specifying a distribution).

Then we check if `replace=TRUE`, if not then it is not a useful bootstrapping (see Section 4.3): 2, 3 and 4 is fine.

Then we see that 3 is some weird expression with the ratio of variances: so we are left with 2 and 4.

We check 4, and find two problems: the `apply` function is used on dimension 1 (such that it applies the function on the rows of the generated data and not the columns) and only the 95% quantile is calculate on the bootstrapped values.

Finally, we check 2 and find that it calculates the confidence interval correct.

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

Continues on page 31

Question XII.2 (23)

We continue with the problem from the previous question, i.e. we analyze a process Y that can be expressed by $Y = U/B$.

If we assume that $U \sim N(\mu = 35, \sigma^2 = 10^2)$ and $B \sim N(\mu = 50, \sigma^2 = 10^2)$, what is the probability that Y exceeds 1, i.e. please calculate the probability $P(Y > 1)$:

1 ☐ < 0.001

*2 ☐ 0.1444

3 ☐ 0.4701

4 ☐ 0.5298

5 ☐ 0.8556

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

We can write up

$$P(Y > 1) = P(\frac{U}{B} > 1) = P(U > B) = P(U - B > 0) = 1 - P(U - B \leq 0).$$

We know from Theorem 2.40 that a linear function of normal distributed random variables is also normal distributed. With Theorem 2.56 we can calculate the mean of $U - B$

$$\mu_{U-B} = E(U - B) = E(U) - E(B) = 35 - 50 = -15,$$

and the variance

$$\sigma^2_{U-B} = \text{Var}(U - B) = \text{Var}(U) + \text{Var}(B) = 100 + 100 = 200.$$

Hence $U - B \sim N(-15, 200)$ and now we can calculate

$$P(Y > 1) = 1 - P(U - B \leq 0),$$

in R by

```
1 - pnorm(q=0, mean=-15, sd=sqrt(200))

## [1] 0.1444222
```

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

Continues on page 32

Exercise XIII

A research institute wants to estimate a 95% confidence interval for the true proportion p of consumers who are consciously trying to purchase organic foods when shopping. The research institute plans to ask n consumers the question: "Do you consciously chose to buy organic foods when you do your grocery shopping?". Possible answers to this must be "Yes" or "No".

Question XIII.1 (24)

How many independent consumers n must respond to the survey for at 95% confidence interval for the true proportion, p, to not be wider than 0.04 (Hint: As a starting point for the calculations it can be assumed that 50% of consumers will answer 'Yes' to the question)?

1 ☐ $n = \frac{1.96 \cdot \frac{1}{2} \cdot \frac{1}{2}}{0.01} = 49$

2 ☐ $n = (\frac{1.96 \cdot \frac{1}{2} \cdot \frac{1}{2}}{0.02})^2 = 600.25$ i.e. at least 601

*3 ☐ $n = (\frac{1.96^2 \cdot \frac{1}{2} \cdot \frac{1}{2}}{0.02^2}) = 2401$

4 ☐ $n = (\frac{2 \cdot 1.96 \cdot \frac{1}{2} \cdot \frac{1}{2}}{0.02})^2 = 1250.50$ i.e. at least 1251

5 ☐ $n = (\frac{1.96 \cdot \sqrt{\frac{1}{2} \cdot \frac{1}{2}}}{0.01})^2 = 9604$

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

Since we are trying to determine the sample size for a one-proportion test, we can use Method 7.13

$$n = p(1-p)\left(\frac{z_{1-\alpha/2}}{ME}\right)^2 = \frac{1}{2} \cdot \frac{1}{2} \cdot \left(\frac{1.96}{0.02}\right)^2 = 2401,$$

and see that the answer is simply this formula modified with some parts shifted around.

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

Question XIII.2 (25)

The question is no longer a part of the curriculum

Continues on page 33

Exercise XIV

Assume that the number of attempts for the driving test (before it is passed) in a particular municipality can be described by the model $Y = X + 1$, where X is a Poisson distributed random variable with mean $\lambda = 0.4$, i.e. $X \sim Pois(\lambda = 0.4)$.

Question XIV.1 (26)

We now consider the number of attempts among 100 randomly selected individuals who must pass the driving test. What will be the mean μ and variance σ^2 of the total number of attempts $\sum_{i=1}^{100} Y_i$ for everyone passing the exam?

1 ☐ $\mu = 140$ and $\sigma^2 = \sqrt{40}$

2 ☐ $\mu = 140$ and $\sigma^2 = \sqrt{140}$

3 ☐ $\mu = 140$ and $\sigma^2 = 140$

*4 ☐ $\mu = 140$ and $\sigma^2 = 40$

5 ☐ $\mu = 140$ and $\sigma^2 = 40^2$

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

First we calculate the mean of X, which we from Theorem 2.28 is λ. We can then use Theorem 2.54 to find the mean and variance of Y.

$$\mathrm{E}(Y) = \mathrm{E}(X + 1) = \mathrm{E}(X) + 1 = \lambda + 1 = 1.4,$$

and similarly the variance

$$\mathrm{Var}(Y) = \mathrm{Var}(X + 1) = \mathrm{Var}(X) = \lambda = 0.4.$$

Then we use Theorem 2.56 to find the mean and variance of the linear combination of the 100 drivers

$$\mu = \mathrm{E}(\sum_{i=1}^{100} Y_i) = \mathrm{E}(Y_1) + \mathrm{E}(Y_2) + \cdots + \mathrm{E}(Y_{100}) = 100 \cdot 1.4 = 140,$$

and

$$\sigma^2 = \mathrm{Var}(\sum_{i=1}^{100} Y_i) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + \cdots + \mathrm{Var}(Y_{100}) = 100 \cdot 0.4 = 40.$$

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

Continues on page 34

Exercise XV

Assume that the number of traffic accidents per day, X, follows a Poisson distribution. From 200 independent observations, the rate λ has been estimated to $\hat{\lambda} = 1.2$.

Question XV.1 (27)

Please provide a 95% confidence interval for the true rate λ:

1 ☐ $[1.2; 4.8]$

2 ☐ $[0; 4]$

*3 ☐ $1.2 \pm 1.96 \cdot \sqrt{\frac{1.2}{200}}$

4 ☐ $1.2 \pm 1.96 \cdot \frac{1.2}{200}$

5 ☐ $1.2 \pm 1.96 \cdot \frac{1.2^2}{200^2}$

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

Since we have $n = 200 > 30$ then, according to the central limit theorem (Theorem 3.14), we can use the usual confidence interval based on the t-distribution (or standard normal distribution). So we need the sample mean and sample variance, which we using Theorem 2.28 find simply equal to the estimate of the rate for a Poisson distributed random variable

$$\hat{\mu} = \bar{x} = \hat{\lambda} = 1.2,$$
$$\hat{\sigma}^2 = s^2 = \hat{\lambda} = 1.2,$$

which we plug in the formula from Method 3.9

$$\begin{aligned} \bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} &= \bar{x} \pm t_{\alpha/2} \cdot \sqrt{\frac{s^2}{n}} \\ &= 1.2 \pm 1.96 \cdot \sqrt{\frac{1.2}{200}}, \end{aligned}$$

where $t_{\alpha/2} = t_{0.975} \approx z_{0.975}$ is found by

```
qt(p=0.975, df=199)
```

```
## [1] 1.971957
```

```
qnorm(p=0.975)
```

```
## [1] 1.959964
```

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

Continues on page 35

Exercise XVI

Two types of prescription drugs (A and B) to lower blood cholesterol, are compared in a clinical study. In analyzing the data, it was estimated how much drug A reduces cholesterol, denoted Δ_A, and correspondingly how much drug B is reducing cholesterol, denoted Δ_B (both drugs were found to reduce cholesterol and in the following positive values of Δ indicate reduction).

A 95% confidence interval for the difference in reduction ($\Delta_A - \Delta_B$) has been estimated. This interval is $[0.24; 0.50]$ mmol/L.

Question XVI.1 (28)

Which of the following is a reasonable conclusion to the survey?

1 ☐ Drug A reduces cholesterol by 0.24 mmol/L while drug B reduces cholesterol by 0.50 mmol/L

2 ☐ There is 95% probability that drug A is better to lower the cholesterol than drug B for any person

3 ☐ There is 95% probability that drug A will lower cholesterol with at least 39 mmol/L compared to drug B for any person

*4 ☐ There is at least 95% confidence that drug A reduces cholestrol better than drug B

5 ☐ None of the above

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

Lets go through the answers one by one:

- 1: We have a confidence interval for the difference in mean for the two drugs, but we don't know nothing about how much each of drug reduces cholesterol
- 2: We don't know exactly the probability that drug A is better than drug B. Actually, we can only talk about the probability like this before the experiment, i.e. not using the data. We could maybe have an estimate of a probability, but not like this state "there is 95% probability of ..." from values calculated from data
- 3: Same as 2
- 4: This formulation is correct. If we tested the null hypothesis $H_0 : \Delta_A - \Delta_B = 0$, we would find that zero is outside the confidence interval. Therefore we know that the p-value would be below 5% and thus the formulation "There is at least 95% confidence ..." is appropriate (we could also have used 'certainty' instead of 'confidence')
- 5: Since 4 is reasonable conclusion, then this is not correct

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

Continues on page 38

Exercise XVII

An engineer is planning to take a sample from a population. We consider the following three statements:

I. If the sample has variance zero, then the variance in the population is also zero.

II. If the population has variance zero, then the variance in the sample is also zero.

III. If the sample has zero variance, then the mean and median is the same in the sample.

Question XVII.1 (29)

Which of the three above statements are correct?

1 ☐ Only I. and II.

2 ☐ Only I. and III.

*3 ☐ Only II. and III.

4 ☐ I., II., and III. are all correct

5 ☐ None of the above

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

Lets try to verify the statements

- I.: If we take a sample from a discrete variable with multiple outcomes (e.g. a dice rool), then we can easily imagine that we could get a sample with e.g. 6 equal values (a Yatzy!), which would then have a sample variance of zero. However, the population would in this case not have a variance of zero. Hence Statement I. is not correct
- II.: If the population variance is zero, then there is only a single possible outcome value (e.g. a dice with 5 marked on each side), and every sample would be with only that value (e.g. we would roll a 5 every time). In this case the sample variance will always be zero. Hence Statement II. is correct.
- III.: If the sample has zero variance, then all the values in the sample are equal. In this case we can see that the sample mean will also be equal to this value, and also the median, since it is the value in the middle when the sample is ordered. As an example: the sample is (5,5,5,5,5,5), then the sample mean is 5, and the value in the middle (median) is 5. Hence Statement III. is correct.

So only Statement II. and III. are correct.

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

Continues on page 40

Exercise XVIII

It is well known that cuckoos lay their eggs in another bird species nests, and thus leaves the task to raise their offspring to the host bird. Furthermore, it is a theory that cuckoos are able to adapt the size of their eggs depending on the size of the host bird.

To investigate this theory an ornithologists has over a period measured the size (length of the egg) of 10 eggs in each of two different host bird nests, here called the host bird A and B, that is, a total 20 eggs are measured. She gets the estimates 2 mm for the standard deviation of the size for both the host bird A and B.

Question XVIII.1 (30)

It now turns out that the observed difference in size, $\bar{x}_A - \bar{x}_B$, of the eggs is 1 mm. What conclusion does one arrive at when you want to test the hypothesis $H_0 : \mu_A = \mu_B$ against $H_1 : \mu_A \neq \mu_B$, using an ordinary t-test and significance level $\alpha = 5\%$?

1 ☐ The difference in the size of the eggs is statistically significant

*2 ☐ The difference in the size of the eggs is statistically non-significant

3 ☐ One can not conclude anything without stating the actual size of the eggs for A and B

4 ☐ One can not conclude anything without the knowledge of the population size

5 ☐ It is not appropriate to use an ordinary t-test for this analysis

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

It is a two-sample test for the difference in mean, so we use Method 3.49. First we calculate the test statistic

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{1}{\sqrt{4/10 + 4/10}} = 1.25,$$

which we use to calculate the p-value

$$p\text{-value} = 2 \cdot P(T > 1.25)$$

where the degrees of freedom in the t-distribution is

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{\left(\frac{4}{10} + \frac{4}{10}\right)^2}{\frac{(4/10)^2}{9} + \frac{(4/10)^2}{9}} = 18.$$

Thus the p-value is

```r
2*(1 - pt(q=1.25, df=18))

## [1] 0.2273077
```

which the only conclusion fitting this is answer 2.

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

The exam is over. Have a good Christmas vacation!

Technical University of Denmark *Page 1 of 45 sider.*

Written examination: 28. May 2016

Course name and number: **Introduction to Statistics (02323, 02402 and 02593)**

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 12 exercises. To answer the questions you need to fill in the prepared 30-question multiple choice form (on three 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"). Also, if more answers are 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 answer of the exercises should be given by filling in and submitting via the exam module in CampusNet. The table sheet here is ONLY to be used as an "emergency" alternative.

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

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

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

Remember to provide your **study number**. The questionnaire contains 45 pages. Please check that your questionnaire contains them all.

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

In an airport the security check screens exactly 10000 passengers each day. Based on data from a long period it was found that 8 out of 10000 passengers bring sharp objects in their carry on luggage. Let X be a random variable denoting the number of passengers with sharp objects on a day (based on exactly 10000 checks). X is assumed to follow a Binomial distribution.

Question I.1 (1)

What is the expected number of passengers with sharp objects on a day, and what is the variance of X?

1 ☐ $E[X] = 0.0008$ and $V[X] = 10000 \cdot 0.8 \cdot 0.2 = 1600$

2 ☐ $E[X] = 0.0008 \cdot 10000 = 8$ and $V[X] = 10000 \cdot 0.0008 \cdot 0.0002 = 0.0016$

3* ☐ $E[X] = 0.0008 \cdot 10000 = 8$ and $V[X] = 10000 \cdot 0.0008 \cdot 0.9992 = 7.994$

4 ☐ $E[X] = 0.0008 \cdot 10000 = 8$ and $V[X] = 10000 \cdot 0.0008 = 8$

5 ☐ $E[X] = 0.8 \cdot 10 = 8$ and $V[X] = 10 \cdot 0.8 \cdot 0.2 = 1.6$

———— FACIT-BEGIN ————

$E[X] = n * p = 10000 * 0.0008 = 8$ and $V[X] = n * p * (1 - p) = 10000 * 0.0008 * 0.9992 = 7.994$

———— FACIT-END ————

Question I.2 (2)

What is the probability of finding more than 10 passengers with sharp objects on a given day?

1 ☐ `qbinom(0.9, 10000, 0.0008)`

2 ☐ `1-dbinom(9990, 10000, 0.0008)`

3 ☐ `dbinom(10, 10000, 0.0008)`

4 ☐ `1-pbinom(9990, 10000, 0.0008)`

5* ☐ `1-pbinom(10, 10000, 0.0008)`

———— FACIT-BEGIN ————

The result is found by calculating 1 minus the probability of finding 10 or less sharp objects.

———— FACIT-END ————

Continues on page 4

Exercise II

A pharmaceutical company made a study in which 300 persons were randomly divided into 3 treatment groups of 100 patients each. One group was assigned to a placebo treatment, one group received the company's own product, and the last group got a competitor's product. For each patient the weight change over a period of time was measured and the final data set consists of 300 observations of weight changes. The focus is on comparing the average weight change in each group.

Question II.1 (3)

What kind of statistical analysis is most suitable for this?

1 ☐ Multiple linear regression analysis

2 ☐ Test for independence in a $r \times c$ frequency table (Contingency table)

3 ☐ Paired t-test

4 ☐ Toway analysis of variance

5* ☐ Oneway analysis of variance

———— FACIT-BEGIN ————

With the description this is clearly 3 independent samples of quantitative data, so the oneway anova is the right choice, so answer 5).

———— FACIT-END ————

Continues on page 5

Exercise III

A random variable X follows a uniform distribution on the interval $[0; 1]$.

Question III.1 (4)

The expected value and the variance of $(X + 2) \cdot 4$ is

1 ☐ $\mu = \frac{5}{2}$ and $\sigma^2 = 4^2$

2 ☐ $\mu = 10$ and $\sigma^2 = 4^2$

3 ☐ $\mu = 8$ and $\sigma^2 = 4^2$

4 ☐ $\mu = 8$ and $\sigma^2 = \frac{1}{3}$

5* ☐ $\mu = 10$ and $\sigma^2 = \frac{4}{3}$

——— FACIT-BEGIN ———

The transformed variable is uniformly distributed on $[8, 12]$ so by Eq. 2-52 and 2-53 the answer is: $\mu = \frac{1}{2}(12 - 8) = 10$ and $\sigma^2 = \frac{1}{12}(12 - 8)^2 = \frac{16}{12} = \frac{4}{3}$.

——— FACIT-END ———

Continues on page 6

Exercise IV

A drone manufacturer is focusing on the feasible flight time between recharges. The flight time depends among other things on the weight of the drone. The drone basically consists of a battery(B), a skeleton (S) and four engines with propellers ($M_1, \ldots, M_4$). It is assumed that the weights of the individual parts are independent and in the following all weights are in grams. The weights of the three types of components are given by the following Normal distributions: Battery: $B \sim N(100, 10^2)$, skeleton: $S \sim N(40, 5^2)$ and engines with propellers: $M_i \sim N(15, 2^2),\ i = 1, \ldots, 4$. (Each distribution is given on the usual form: $N(\mu, \sigma^2)$)

Question IV.1 (5)

The expected value and variance for the weight of the assembled drones are found to be

1 ☐ $\mu = 200$ and $\sigma^2 = 189$

2* ☐ $\mu = 200$ and $\sigma^2 = 141$

3 ☐ $\mu = 155$ and $\sigma^2 = 189$

4 ☐ $\mu = 155$ and $\sigma^2 = 129$

5 ☐ $\mu = 170$ and $\sigma^2 = 141$

———— FACIT-BEGIN ————

$\mu = 100 + 40 + 4 * 15 = 200$ and $\sigma^2 = 10^2 + 5^2 + 4 * 2^2 = 141$

———— FACIT-END ————

Continues on page 7

Exercise V

There is a recommendation to eat 600 grams of fruit and vegetables each day. Regularly surveys of Danish dietary habits are made to see if the recommendation is met.

The results of the daily intake of fruits and vegetables (in grams) for the last four of this kind of dietary studies (conducted in the years 1995, 2000-2002, 2003-2004 and 2005-2008) can be summarized by the following output from R.

Survey	n	median	mean	var	std
1995	1564	259.82	290.887	28861.55	169.887
2000-2002	3043	386.057	433.817	62029.21	169.887
2003-2004	1310	404.936	453.279	74159.29	272.322
2005-2008	1983	429.132	479.285	77166.51	277.789

Survey	2.5%	5.0%	Q1	Q3	95.0%	97.5%
1995	66.102	87.062	171.209	374.303	606.609	686.361
2000-2002	98.613	129.574	257.224	555.168	928.673	1055.419
2003-2004	83.48	127.528	256.286	583.723	974.246	1180.891
2005-2008	105.348	141.81	279.359	617.371	991.09	1189.367

In all the questions in this exercise one can assume that the data from each of the four studies are normally distributed.

Question V.1 (6)

The question is no longer part of the curriculum.

Continues on page 8

Question V.2 (7)

A new dietary study is planned on the basis of the observed variation in dietary survey 2005-2008. What should the sample size be if the 90% confidence interval for the mean intake of fruits and vegetables is aimed to have a width of 20 grams?

1 ☐ $n \approx 77166.51/\left(\frac{20}{1.96}\right)^2 = 741.1$

2 ☐ $n \approx \left(\frac{479.285 \cdot 1.6449}{10}\right)^2 = 6215.4$

3 ☐ $n \approx \frac{77166.51}{1.96 \cdot 20} = 1968.5$

4* ☐ $n \approx \left(\frac{1.6449 \cdot 277.789}{10}\right)^2 = 2087.9$

5 ☐ $n \approx \left(\frac{1.6449 \cdot \sqrt{1983}}{1.6456}\right)^2 = 1981.3$

———— FACIT-BEGIN ————

Cf. Model in question V.1 (6), we are still in the same normal distribution model.

To determine the sample size, with 90% confidence interval for the mean intake of fruits and vegetables (μ), so it not exceed a width of 20 grams based on the dietary survey from 2005 to 2008, used Method 3.45 in eNote 3 page 44:

$$n = \left(\frac{z_{1-\alpha/2} \cdot \sigma}{ME}\right)^2 = \left(\frac{1.6449 \cdot 277.789}{10}\right)^2 = 2087.9$$

Since $ME = 0.5 \cdot 20$ (ME is half the width of the confidence interval), as variance we use the estimate from 2005-2008 dietary survey, ie $\sigma = 277.789$ and since it is 90% confidence interval, which is under consideration, we have that $1 - \alpha/2 = 0.95$, $z_{0.95} = 1.6449$. Thus we see that the correct answer is 4.

```
qnorm(0.95)

## [1] 1.644854
```

———— FACIT-END ————

Question V.3 (8)

Determine the 95% confidence interval for the mean intake of fruit and vegetables in the 2003-2004 survey.

1 ☐ $404.936 \pm 1.9618 \cdot \sqrt{\frac{74159.29}{1310}} = [390.176; 419.697]$

2 ☐ $[83.48; 1180.891]$

3* ☐ $453.279 \pm 1.9618 \cdot 7.524 = [438.518; 468.040]$

4 ☐ $453.279 \pm 1.6460 \cdot \frac{272.322}{\sqrt{1310}} = [440.894; 465.664]$

5 ☐ $453.279 \pm 1.96 \cdot 272.322 = [-80.472; 987.030]$

———— FACIT-BEGIN ————

Consider 2003-2004 dietary survey.

Let X_i be a random variable denoting the ith respondent's intake of fruits and vegetables per. day in 2003-2004 dietary survey. Assume X_i is normally distributed N(μ ,σ^2), where the model parameters are estimated at: $\hat{\mu} = 453.279$ and $\hat{\sigma}^2 = 74159.29 = (272.322)^2$

To determine the $1 - \alpha$ confidence interval for the mean intake of fruits and vegetables in 2003-2004 dietary survey (μ) used Method 3.8 eNote 3 page 12

$\bar{x} \pm t_{1-\alpha/2} \cdot s/\sqrt{n} = 453.279 \pm 1.9618 \cdot 7.524 = [438.518; 468.040]$

Since it is 95% confidence interval, we must determine, we have $\alpha = 0.05$, $t_{0.975} = 1.9618$, as it is 97.5% percentile of the t-distribution with 1309 degrees of freedom we should use. In addition, $s/\sqrt{n} = 272.322/\sqrt{1310} = 7.524$

```
qt(0.975,1309)
```

```
## [1] 1.961778
```

Thus we see that the correct answer is 3

———— FACIT-END ————

Continues on page 10

Exercise VI

A major company took a random sample of 20 employees and determined their daily intake of fruits and vegetables, and registered the following observations of the daily intake (in grams):

740.59	262.28
667.96	730.55
809.33	324.19
1138.12	421.93
489.42	561.23
352.78	552.96
1309.66	130.96
259.86	440.82
896.01	955.03
481.00	257.80

In all the questions in this exercise one can assume that the data is normally distributed.

Summary from R gives the following results for the intake of fruits and vegetables:

```
   n      median       mean        variance        Std. dev.
  20    521.1898    589.1245      98996.08        314.6364

    2.5%        5.0%         Q1          Q3         95.0%      97.5%
 191.2095    251.4552    345.635    757.777    1146.697    1228.178
```

Question VI.1 (9)

Determine the 90% confidence interval for the variance σ^2 for the daily intake of fruits and vegetables of employees in the company

1 ☐ $[\frac{20 \cdot 314.636}{32.852}; \frac{20 \cdot 314.636}{8.907}] = [191.548; 706.492]$

2 ☐ $98996.08 \pm 30.144 \cdot \frac{314.636}{\sqrt{20}} = [96875.31; 101116.90] = [311.248^2; 317.989^2]$

3 ☐ $98996.08 \pm 1.7959 \cdot \frac{314.636^2}{\sqrt{20}} = [59241.79; 138750.40] = [243.396^2; 372.492^2]$

4 ☐ $[314.636^2 - 10.117 \cdot 314.636; 314.636^2 + 30.144 \cdot 314.636] = [309.537^2; 329.364^2]$

5* ☐ $[\frac{19 \cdot 314.636^2}{30.144}; \frac{19 \cdot 314.636^2}{10.117}] = [249.796^2; 431.181^2]$

———— FACIT-BEGIN ————

Consider the sample comprising 20 employees.

Let X_i be a random variable, denoting the ith employee's daily intake of fruits and vegetables in this random sample. Assume X_i is normally distributed N(μ ,σ^2), where the model parameters are estimated at: $\hat{\mu} = 589.1245$ and $\hat{\sigma}^2 = 98996.08 = 314.636^2$

To determine the $1-\alpha$ confidence interval for the variance (σ^2) for the daily intake of fruits and vegetables in the random sample used Method 3.18 eNote 3 page 24

$$\left[\frac{(n-1)\cdot s^2}{\chi^2_{1-\alpha/2}}; \frac{(n-1)\cdot s^2}{\chi^2_{\alpha/2}}\right] = \left[\frac{19\cdot 314.636^2}{\chi^2_{0.95}}; \frac{19\cdot 314.636^2}{\chi^2_{0.05}}\right] = \left[\frac{19\cdot 314.636^2}{30.144}; \frac{19\cdot 314.636^2}{10.117}\right] = [249.796^2; 431.181^2]$$

Since we should determine the 90% confidence interval, it is clear that $\alpha = 0.10$. $\chi^2_{\alpha/2}$, $\chi^2_{1-\alpha/2}$ are percentiles of the chi-square / χ^2-distribution with $\nu = n - 1 = 19$ degrees of freedom. It follows that: $\chi^2_{0.05} = 10.117$, $\chi^2_{0.95} = 30.144$ from

```
qchisq(0.05,19)
```

```
## [1] 10.11701
```

```
qchisq(0.95,19)
```

```
## [1] 30.14353
```

Thus we see that the correct answer is 5

———— FACIT-END ————

Continues on page 12

Question VI.2 (10)

Actually, the above data consist of 2 random samples, where the left column indicate the intakes of 10 men and the right column intakes for 10 women. One wants to investigate whether there are differences in men's and women's mean intake of fruits and vegetables.

The following R code is executed (not all necessarily sensible):

```
m <- c(740.59, 667.96, 809.33, 1138.12, 489.42, 352.78,
       1309.66, 259.86, 896.01, 481.00)
f <- c(262.28, 730.55, 324.19, 421.93, 561.23, 552.96,
       130.96, 440.82, 955.03, 257.80)
t.test(m, f, paired = TRUE)

##
##  Paired t-test
##
## data:  m and f
## t = 1.7378, df = 9, p-value = 0.1163
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -75.65101 577.04701
## sample estimates:
## mean of the differences
##                 250.698

mean(f) - mean(m)

## [1] -250.698

t.test(m, f)

##
##  Welch Two Sample t-test
##
## data:  m and f
## t = 1.9001, df = 16.481, p-value = 0.07506
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -28.33599 529.73199
## sample estimates:
## mean of x mean of y
##   714.473   463.775
```

Continues on page 13

```r
t.test(m, mu = median(f))
```

```
##
##  One Sample t-test
##
## data:  m
## t = 2.6577, df = 9, p-value = 0.02614
## alternative hypothesis: true mean is not equal to 431.375
## 95 percent confidence interval:
##  473.5091 955.4369
## sample estimates:
## mean of x
##   714.473
```

```r
t.test(f)
```

```
##
##  One Sample t-test
##
## data:  f
## t = 5.957, df = 9, p-value = 0.0002135
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  287.6581 639.8919
## sample estimates:
## mean of x
##   463.775
```

Continues on page 14

What is the conclusion of the test of the hypothesis (at the level $\alpha = 0.05$)

$$H_0 : \mu_f = \mu_m$$
$$H_1 : \mu_f \neq \mu_m$$

corresponding to the examination of whether there are differences in men's and women's mean intake of fruits and vegetables.

1 ☐ Yes, there is a significant difference between men's and women's intake of fruits and vegetables per day, given that the relevant p-value is 0.1163

2 ☐ It is apparent that there is a significant difference between men's and women's intake of fruits and vegetables per day, with $\hat{\mu}_f - \hat{\mu}_m = 463.775 - 714.473 = -250.698$. It appears that men eat more fruits and vegetables per day than women

3* ☐ There is no significant difference between men's and women's intake of fruits and vegetables per day, as the relevant p-value is 0.07506

4 ☐ Yes, there is a significant difference between men's and women's intake of fruits and vegetables per day given that the relevant p-value is 0.02614

5 ☐ No, there is no significant difference in the intake of fruit and vegetables per day for men and women since the relevant p-value is 0.0002135

———— FACIT-BEGIN ————

Consider again the random sample comprising 20 employees. In fact, the 20 observations 2 random samples where the 10 observations in the first column is the data for men's daily intake of fruits and vegetables, while the 10 observations in the second column is for women's intake.

Let M_i and F_i be independent random variables, where M_i indicates the ith man's daily intake of fruits and vegetables in this random sample and correspondingly F_i ith woman's intake of fruits and vegetables in this random sample. Assume M_i is normally distributed N(μ_m , σ_m^2) and correspondingly that F_i are normally distributed N(μ_f , σ_f^2).The model parameters are estimated by: $\hat{\mu}_m = 714.473$, $\hat{\sigma}_m^2 = 113464.1 = 336.84^2$ and $\hat{\mu}_f = 463.775$, $\hat{\sigma}_f^2 = 60611.72 = 246.19^2$

We want to examine whether there are differences in men's and women's mean intake of fruits and vegetables, corresponding to the following hypothesis:

$$H_0 : \mu_f = \mu_m$$
$$H_1 : \mu_f \neq \mu_m$$

The hypothesis is tested on the level $\alpha = 0.05$

There must be made a Welch two-sample t-test Cf. Method 3.60 eNote 3 page 64. The test statistics is determined by:

$$t_obs = \frac{(\bar{m}-\bar{f})}{\sqrt{s_m^2/n_m+s_f^2/n_f}} = \frac{714.473-463.775}{\sqrt{336.84^2/10+246.19^2/10}} = 1.9001$$

The degrees of freedom is determined by:

$$\nu = \frac{\left(\frac{s_m^2}{n_m}+\frac{s_f^2}{n_f}\right)^2}{\frac{(s_m^2/n_m)^2}{n_m-1}+\frac{(s_f^2/n_f)^2}{n_f-1}} = \frac{\left(\frac{336.84^2}{10}+\frac{246.19^2}{10}\right)^2}{\frac{(336.84^2/10)^2}{9}+\frac{(246.19^2/10)^2}{9}} = 16.481$$

Since T is t-distributed with $\nu = 16.481$ degrees of freedom, the p-value determined by

$$p = 2 \cdot P(T > |t_{obs}|) = 2 \cdot P(T > 1.9001) = 0.07506$$

```
2*(1-pt(abs(1.9001),16.481))

## [1] 0.07506054
```

As $p > 0.05$ we accept H_0, i.e. there is no significant difference between men and women mean intake of fruits and vegetables. It appears that the answer 3, is the right solution.

———— FACIT-END ————

Question VI.3 (11)

What is the upper quartile (75% percentile) for the 10 intake for women, based on the textbook definition of this?

1* ☐ 561.23

2 ☐ 262.28

3 ☐ 431.375

4 ☐ 709.97

5 ☐ 246.19

———— FACIT-BEGIN ————

The ordered sample sample for the random sample of the 10 women daily intake of fruits and vegetables is determined by:

130.96, 257.80, 262.28, 324.19, 421.93, 440.82, 552.96, 561.23, 730.55, 955.03

According to Definition 1.4 Median eNote 1 page 10, respectively Definition 1.6 Quantiles and Percentiles eNote 1 page 12 the upper quartile is determined based on the ordered sample.

As n = 10, $np = 7.5$, the upper quartile is the 8th observation, that is, 561.23, so the correct answer is 1.

```
f <- c(262.28, 730.55, 324.19, 421.93, 561.23, 552.96, 130.96, 440.82, 955.03, 257.80)
quantile(f, type=2)

##      0%     25%     50%     75%    100%
## 130.960 262.280 431.375 561.230 955.030
```

———— FACIT-END ————

Continues on page 17

Exercise VII

A study investigated dioxin emissions from a Danish incineration plant. Parts of the measured variables are shown in the table below. The 3 variables are: Dioxin measured in "parts per millon", load of the plant measured as relative deviation from a reference, and the content of water in the emitted gas (measured in %). As seen in the table, there are in total 23 measurements. Average and empirical standard deviation ("sample standard deviation") are listed at the bottom of the table.

	Dioxin (ppm)	Load	H_20 (%)
	DIOX	NEFF	H2O
1	984.10	0.2560	13.78
2	662.00	0.3520	14.59
3	270.90	-0.0200	12.55
⋮	⋮	⋮	⋮
21	112.70	0.0490	13.84
22	94.20	0.1350	14.18
23	323.20	0.2820	12.56
$\bar{x}$	329.16	-0.0266	12.589
s	254.95	0.2105	1.980

The primary interest in the study is related to the question: can dioxin emissions be influenced by adjusting the load. For this purpose the following R code is executed (the data input is, however, omitted)

```
fit1 <- lm(DIOX ~ NEFF)
summary(fit1)

##
## Call:
## lm(formula = DIOX ~ NEFF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -348.41 -116.61  -22.98  101.19  496.16
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)    347.8       44.7   7.781 0.000000128 ***
## NEFF           702.2      215.3   3.262     0.00373 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 212.6 on 21 degrees of freedom
## Multiple R-squared:  0.3362,Adjusted R-squared:  0.3046
## F-statistic: 10.64 on 1 and 21 DF,  p-value: 0.00373
```

Continues on page 18

Hence, the following model is examined

$$\texttt{DIOX}_i = \beta_0 + \beta_1 \texttt{NEFF}_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

Question VII.1 (12)

At significance level $\alpha = 0.05$, what is the conclusion about the effect of the load on dioxin emissions (both conclusions and argument must be correct)?

1 ☐ There is an effect since $1.3 \cdot 10^{-7} < 0.05$, and $\beta_1 > 0$ because $347.8 > 0$

2 ☐ There is an effect since $702.2 > 347.2$, and $\beta_1 > 0$ because $3.26 > 0$

3* ☐ There is an effect since $0.0037 < 0.05$, and $\beta_1 > 0$ because $702.2 > 0$

4 ☐ There is no evidence of an effect as $3.26 < 7.78$.

5 ☐ There is no evidence of an effect as $0.0037 > \frac{0.05}{100}$.

———— FACIT-BEGIN ————

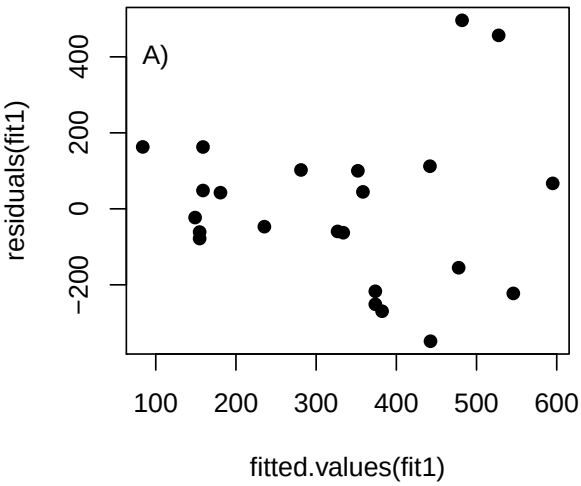
In this case we test the hypothesis

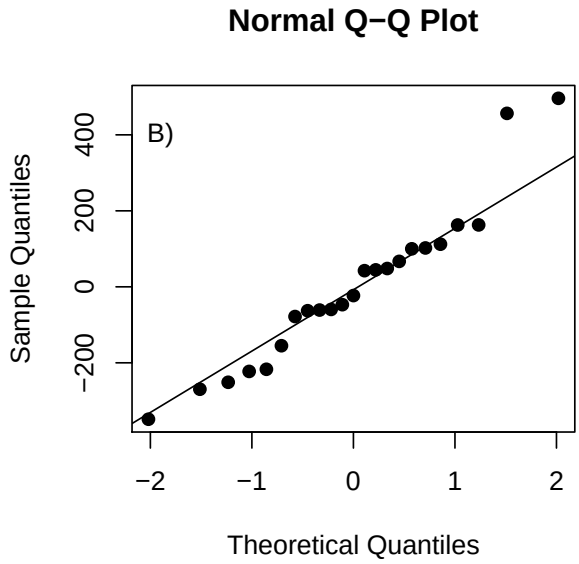
$$H_0: \quad \beta_1 = 0 \tag{1}$$

against the two-sided alternative. The p-value for this hypothesis is given directly in the table as 0.00373, since $0.00373 < 0.05$ there is an effect on the specified level. Since also $\hat{\beta}_1 = 702.2 > 0$ we have that $\beta_1 > 0$ (on the specified level). Hence the correct answer is no. 3.

———— FACIT-END ————

In order to investigate whether the conditions for using the model are satisfied, 2 residual plots are shown in the figure below.





Continues on page 20

Question VII.2 (13)

Which assumptions are primarily examined in each of the 2 plots (both assumptions and figure reference must be correct)?

1* ☐ Variance homogeneity (A) and the normal distribution assumption (B)

2 ☐ $E(\epsilon) = 0$ (A) and $V(\epsilon) = \sigma^2$ (B)

3 ☐ Variance-homogeneity (A) and assumption of linearity (B)

4 ☐ $E(\epsilon) = 0$ (A) and independence (B)

5 ☐ Independence (A) and variance homogeneity (B)

———— FACIT-BEGIN ————

Figure (A) is used to check variance homogeneity, independece or missing structures, while B is used for shecking the normal assumption, hence the correce answer is no. 1.

Lets just have a loook at the other answers for no. 2 the first part $E[\epsilon] = 0$ does not really make sense to test since $\sum e_i$ is always (by construction) equal 0. The second part is actually variance homogeneity (which is not tested in figure B)).

———— FACIT-END ————

Regardless of the outcome of the previous question it is decided to make the analysis on log-tranformed dioxin data. The result of the analysis conducted in R is shown below (some of the numbers are, however, replaced by letters)

```
> fit2 <- lm(log(DIOX) ~ NEFF)
> summary(fit2)

Call:
lm(formula = log(DIOX) ~ NEFF)

Residuals:
     Min       1Q   Median       3Q      Max
-1.29588 -0.44048  0.05093  0.49403  0.94119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.5927        A     B  < 2e-16 ***
NEFF          1.8416        C     D     E
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6266 on 21 degrees of freedom
Multiple R-squared:  0.2862,Adjusted R-squared:  0.2522
F-statistic:  8.42 on 1 and 21 DF,  p-value: 0.00853
```

Continues on page 22

Question VII.3 (14)

What is D?

1 ☐ D= $\frac{0.623^2}{21} = 0.019$

2 ☐ D= $0.623 \cdot \sqrt{\frac{1}{22 \cdot 0.211^2}} = 0.63$

3* ☐ D= $\frac{1.84}{\text{C}}$

4 ☐ D= $\frac{\text{C}}{\text{B}}$

5 ☐ D= $\frac{0.623}{\sqrt{22}} = 0.13$

———— FACIT-BEGIN ————

The model is in this case

$$\log(\texttt{DIOX}_i) = \beta_0 + \beta_1 \texttt{NEFF}_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

D is the test statistic for the hypothesis

$$H_0: \quad \beta_1 = 0$$

against the twosided alternative, the teststatistic is in this case given by

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}} = \frac{1.84}{\text{C}}$$

where 1.84 is the estimate for β_1 and C is the standard error for $\hat{\beta}_1$ ($\hat{\sigma}_{\beta_1}$). For completeness the full R-output is given below.

```
fit2 <- lm(log(DIOX) ~ NEFF)
summary(fit2)

##
## Call:
## lm(formula = log(DIOX) ~ NEFF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29588 -0.44048  0.05093  0.49403  0.94119
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)   5.5927     0.1317  42.450 < 0.0000000000000002 ***
```

```
## NEFF          1.8416     0.6346   2.902              0.00853 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6266 on 21 degrees of freedom
## Multiple R-squared:  0.2862,Adjusted R-squared:  0.2522
## F-statistic:  8.42 on 1 and 21 DF,  p-value: 0.00853
```

———— FACIT-END ————

Question VII.4 (15)

What is the usual 95% confidence interval for the slope in the model for log(DIOX)?

1 ☐ $1.84 \pm 1.72 \cdot \texttt{B}$

2 ☐ $1.84 \pm 2.08 \cdot 0.2862$

3 ☐ $1.84 \pm 1.72 \cdot \texttt{D}$

4 ☐ $1.84 \pm 2.08 \cdot 0.623$

5* ☐ $1.84 \pm 2.08 \cdot \texttt{C}$

———— FACIT-BEGIN ————

The slope is 1.84 (directly from the R-output), and the standard error for the slope is C, to to get a 95% confidence interval we need to multiply the C by the 0.975 quantile in the t-distribution with 21 degrees of freedom,

$$1.85 \pm C \cdot t_{0.975} \tag{2}$$

the quantile in the t-distribution is calculated by

```
qt(0.975,df=21)
```

```
## [1] 2.079614
```

This is answer no. 5.

———— FACIT-END ————

Continues on page 24

It is now decided to investigate whether water vapor should be included in the model. For this purpose, a multiple regression model is formulated

$$\log(\text{DIOX}_i) = \beta_0 + \beta_1 \text{NEFF}_i + \beta_2 \text{H2O}_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

In order to investigate the model, the following R-code has been executed (including the result)

```
fit3 <- lm(log(DIOX) ~ NEFF + H2O)
summary(fit3)

##
## Call:
## lm(formula = log(DIOX) ~ NEFF + H2O)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11709 -0.36741  0.05337  0.36192  0.90410
##
## Coefficients:
##             Estimate Std. Error t value     Pr(>|t|)
## (Intercept)   7.4704     0.8098   9.225 0.0000000121 ***
## NEFF          2.1963     0.5955   3.688      0.00146 **
## H2O          -0.1484     0.0633  -2.345      0.02948 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5687 on 20 degrees of freedom
## Multiple R-squared:  0.4401,Adjusted R-squared:  0.3841
## F-statistic:  7.86 on 2 and 20 DF,  p-value: 0.003028
```

Question VII.5 (16)

What are the parameter estimates for the model?

1* ☐ $\hat{\beta}_0 = 7.47$, $\hat{\beta}_1 = 2.20$, $\hat{\beta}_2 = -0.148$ og $\hat{\sigma} = 0.569$

2 ☐ $\hat{\beta}_0 = 9.22$, $\hat{\beta}_1 = 3.69$, $\hat{\beta}_2 = -2.35$ og $\hat{\sigma} = 0.4401$

3 ☐ $\hat{\beta}_0 = 7.47$, $\hat{\beta}_1 = 2.20$, $\hat{\beta}_2 = -0.148$ og $\hat{\sigma} = 0.384$

4 ☐ $\hat{\beta}_0 = 9.22$, $\hat{\beta}_1 = 3.69$, $\hat{\beta}_2 = -2.35$ og $\hat{\sigma} = 0.569$

5 ☐ $\hat{\beta}_0 = 9.22$, $\hat{\beta}_1 = 3.69$, $\hat{\beta}_2 = -2.35$ og $\hat{\sigma} = 7.86$

———— FACIT-BEGIN ————

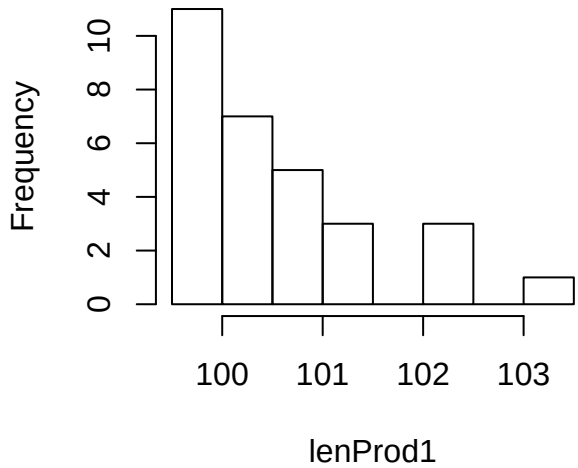
The estimates of β_0-β_1 can be read in the first column of the result (the one denoted "`Estimate`"), this is answer no. 1. The estimate of the residual standard deviation ($\hat{\sigma}$) is 0.569 (the one denoted "`Residual standard deviation`"). This is also in no. 1, hence the correct answer is no. 1.

———— FACIT-END ————

Continues on page 26

Exercise VIII

In a production, it is anticipated that part of the production must be discarded due to a minimum length requirement. It is found that it is economically feasible if not more than 25% of the produced elements are discarded. An experiment is carried out with a particular production method and the length of 50 produced items are observed. The observations are loaded and stored in the vector `lenProd1`. A histogram of the observations is



A confidence interval for the lower quartile (i.e. the 25% quantile) must be calculated without any assumptions of the distribution. The following R code is run:

```
## Simulate 10000 samples
k = 10000
simSamples = replicate(k, sample(lenProd1, replace = TRUE))

simStat = apply(simSamples, 2, quantile, probs=0.25)
quantile(simStat, c(0.005,0.025,0.05,0.95,0.975,0.995))

##     0.5%     2.5%       5%      95%    97.5%    99.5%
##  99.5825  99.7225  99.7600 100.0575 100.1025 100.2300

simStat = apply(simSamples, 2, quantile, probs=0.5)
quantile(simStat, c(0.005,0.025,0.05,0.95,0.975,0.995))

##     0.5%     2.5%       5%      95%    97.5%    99.5%
##  99.9450  99.9850 100.0000 100.5900 100.6550 100.8801

simStat = apply(simSamples, 2, quantile, probs=0.75)
quantile(simStat, c(0.005,0.025,0.05,0.95,0.975,0.995))

##     0.5%     2.5%       5%      95%    97.5%    99.5%
## 100.3000 100.4625 100.5125 101.4300 101.9550 102.1450
```

Continues on page 27

Note that the option `probs` is "passed on" to the `quantile` function, such that for each of the three calls to `apply` a different quantile is calculated by the `quantile` function.

Question VIII.1 (17)

What is the 95% confidence interval for the lower quartile (i.e. the 25% quantile) for the length?

1* ☐ $[99.72, 100.10]$

2 ☐ $[100.00, 100.59]$

3 ☐ $[99.59, 100.23]$

4 ☐ $[100.46, 101.96]$

5 ☐ $[100.49, 101.43]$

———— FACIT-BEGIN ————

To find the correct estimate of the 95% confidence interval for the lower quantile, we need to first find the one of the three calculation of `simStat` which is of the lower quartile (i.e. the 25% quantile). The argument `probs` indicate the quantile to be calculated, hence the first which has `probs=0.25` is the right one. Next we need to find the 2.5% and 97.5% quantile of the simulated statistic, hence the estimated interval is

$$[99.72, 100.10]$$

———— FACIT-END ————

Question VIII.2 (18)

In the following Q denotes a quartile, such that Q_1 is the lower quartile, Q_2 is the median and Q_3 is the upper quartile. In which of the following two-sided tests would the null hypothesis have been rejected on significance level $\alpha = 0.01$ under the assumptions and simulation results presented above?

1 ☐ $H_0 : Q_1 = 100$ vs. $H_1 : Q_1 \neq 100$

2 ☐ $H_0 : Q_2 = 100$ vs. $H_1 : Q_2 \neq 100$

3* ☐ $H_0 : Q_2 = 101$ vs. $H_1 : Q_2 \neq 101$

4 ☐ $H_0 : Q_3 = 101$ vs. $H_1 : Q_3 \neq 101$

5 ☐ $H_0 : Q_3 = 102$ vs. $H_1 : Q_3 \neq 102$

———— FACIT-BEGIN ————

The null hypothesis will be rejected if the value tested for falls out of the confidence interval calculated with the same significance level, as used for the test. Hence, the null hypothesis $H_0 : Q_2 = 101$ is the only one falling outside the respective 0.5% and 99.5% CI.

———— FACIT-END ————

Continues on page 29

Exercise IX

A new wind turbine is to be build on a site and some investigations of the wind conditions on the site have been carried out. The outcome is that the average hourly wind speed on the site can be represented with the probability density function plotted below in Figure 1:

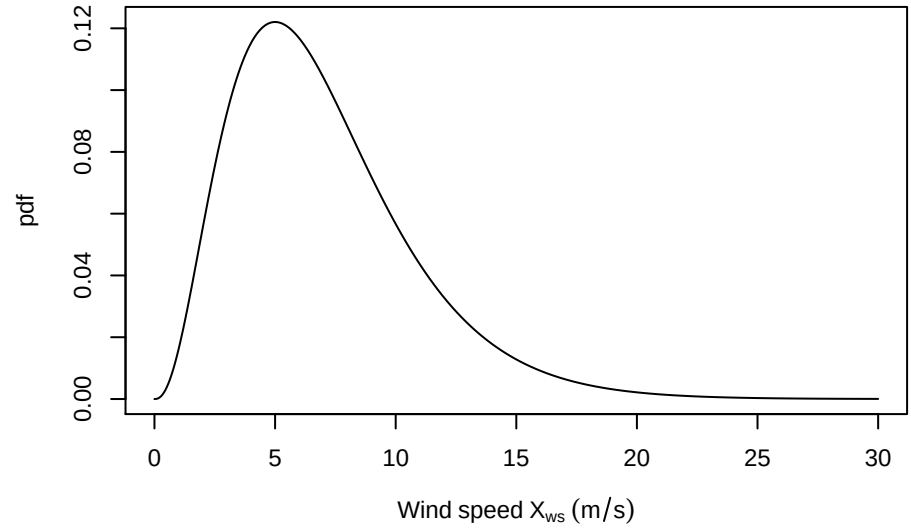


Figure 1: *Probability density function (pdf) for the wind speed* X_{ws}.

In order to investigate the power production of a wind turbine build on the site a function called the 'power curve' for the wind turbine is used (it is the power output as a function of the wind speed, it has nothing to do with the power of a statistical test). The power curve used is plotted below in Figure 2:

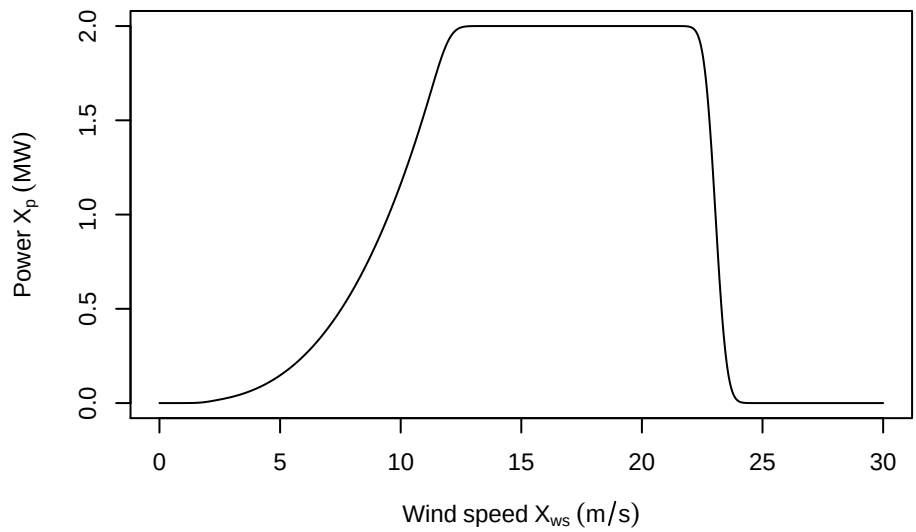


Figure 2: *Power curve, i.e. the function between the wind speed* X_{ws} *and the output power* X_{p}.

It can be seen that if the wind speed is 5 m/s the output power will be around 0.15 MW and at 15 m/s the output power will be 2 MW. This function can be applied directly on average hourly values of the wind speed and gives then hourly average values of output power.

Continues on page 30

Let X_{ws} be the average hourly wind speed in m/s and the power output in MW

$$X_{\text{p}} = f_{\text{powercurve}}(X_{\text{ws}})$$

where $f_{\text{powercurve}}()$ is the power curve function.

Question IX.1 (19)

From the plot of the pdf in Figure 1 conclude which of the following statements is not correct (Note: you must mark the FALSE statement - four of the statements are correct!):

1 ☐ $P(X_{\text{ws}} > 12) \approx 0.10$

2 ☐ $P(X_{\text{ws}} < 5) \approx 0.34$

3 ☐ $P(X_{\text{ws}} > 10) \approx 0.19$

4 ☐ $P(X_{\text{ws}} > 0) \approx 1$

5* ☐ $P(X_{\text{ws}} < 15) \approx 0.04$

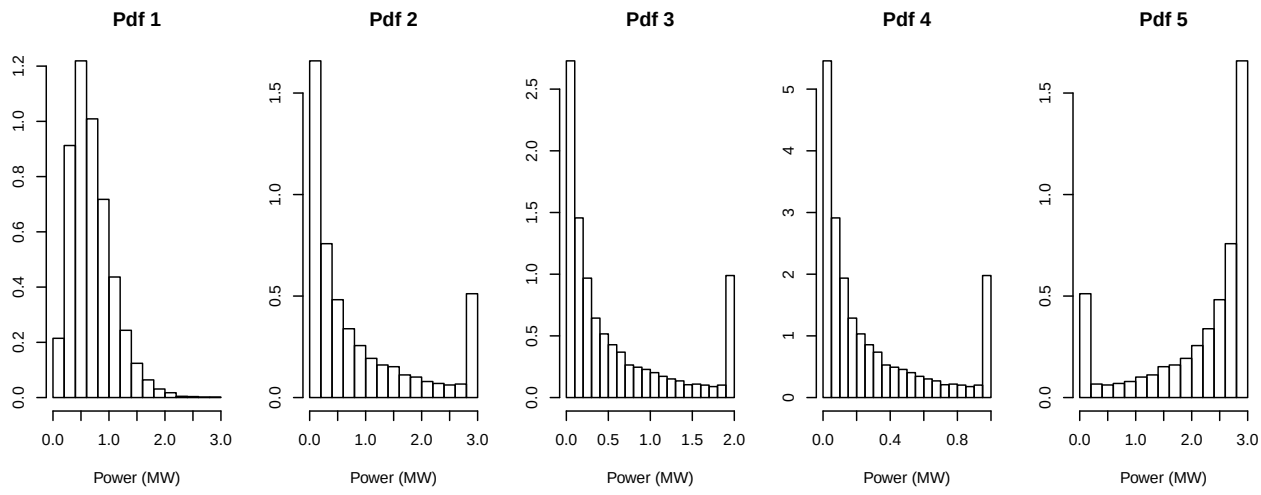
——— FACIT-BEGIN ———

Reading from the plot of the wind speed pdf is clear that the statement $P(X_{\text{ws}} < 15) \approx 0.04$ is not correct, since most of the probability mass (the area below the pdf) is below 15 m/s.

——— FACIT-END ———

Question IX.2 (20)

The probability density function of the hourly output power X_{p} is found by simulation. Which of the following pdfs can be the probability density function of the hourly output power at the site, i.e. the pdf of X_{p}?



Continues on page 31

1 ☐ Pdf 1

2 ☐ Pdf 2

3* ☐ Pdf 3

4 ☐ Pdf 4

5 ☐ Pdf 5

———— FACIT-BEGIN ————

Since the power cannot be higher that 2 MW, which can be seen from the power curve which saturates at 2 MW, Pdf 1, 2 and 5 can be excluded. Further, Pdf 4 has zero probability of being higher than 1 MW, which cannot be the case either. The correct pdf is Pdf 3: it has most probablity mass below 0.5 MW, which because most probability mass of the wind speed pdf is below 7 m/s, which is the point where output power is approximately 0.5 MW. Further, the saturation of the power curve at 2 MW gathers all probability mass in the range approx. 12 to 23 m/s around 2 MW, which creates the peak in the output power pdf around 2 MW.

———— FACIT-END ————

Question IX.3 (21)

In wind power forecasting it is very important to include the forecast uncertainty. It can be described by the variance of the power output forecast σ_{p}^2. In the range from 5 to 10 m/s the power curve is constructed by the following relation

$$f_{\mathrm{powercurve}}(X_{\mathrm{ws}}) = aX_{\mathrm{ws}}^3 \quad \textit{for } 5 < X_{\mathrm{ws}} < 10 \tag{3}$$

Further, it is known that the variance of the wind speed forecast in the same range is σ_{ws}^2.

Which of the following expressions calculates an approximation to the variance of the output power forecast σ_{p}^2 for a wind speed X_{ws} in the range from 5 to 10 m/s?

1 ☐ $\sigma_{\mathrm{p}}^2 = X_{\mathrm{ws}}^3 \sigma_{\mathrm{ws}}^2$

2* ☐ $\sigma_{\mathrm{p}}^2 = 9a^2 X_{\mathrm{ws}}^4 \sigma_{\mathrm{ws}}^2$

3 ☐ $\sigma_{\mathrm{p}}^2 = \int_5^{10} \sigma_{\mathrm{ws}}^2 3ax^2 dx$

4 ☐ $\sigma_{\mathrm{p}}^2 = \int_5^{10} \sigma_{\mathrm{ws}}^2 x^3 dx$

5 ☐ $\sigma_{\mathrm{p}}^2 = a^2 \sigma_{\mathrm{ws}}^2$

———— FACIT-BEGIN ————

We need to consider error propagation through a non-linear function and the power curve aX_{ws}^3 is a non-linear function. Hence we can use the error propagation rule in Method 4.6. We need the derived function with respect to X_{ws}

$$\frac{\partial f_{\text{powercurve}}}{\partial x_{\text{ws}}} = 3aX_{\text{ws}}^2$$

We know the variance of the wind speed forecast σ_{ws}^2, hence the correct expression is found by inserting (and squaring the partial derivative) in the Method 4.6 formula

$$\sigma_{\text{p}}^2 = 9a^2X_{\text{ws}}^4\sigma_{\text{ws}}^2$$

———— FACIT-END ————

Continues on page 34

Exercise X

A supermarket chain would like to track the trend in sales of organic meat. Therefore, they have for four years conducted a survey among their customers, asking whether the customers bought organic meat. The distribution of the answers is seen in the table below.

	2011	2012	2013	2014
Bought organic meat	68	72	81	90
Bought non-organic meat	432	428	419	410

Question X.1 (22)

The supermarket chain wants to test the hypothesis that the proportion buying organic meat is the same each year.

$$H_0 : p_1 = p_2 = p_3 = p_4$$

Here p_1 is the proportion that buys organic meat in 2011, p_2 is the proportion that buys organic meat in 2012 etc.

What is the expected number of organic meat purchases in 2014, under the hypothesis of equal proportions each year?

1 ☐ 144.69

2 ☐ 250.00

3* ☐ 77.75

4 ☐ 43.48

5 ☐ 422.25

——— FACIT-BEGIN ———

We are looking for the number

$$\begin{aligned} e_{1,4} &= \frac{\text{Row 1 total} \cdot \text{Column 4 total}}{\text{Grand total}} \\ &= \frac{(68+72+81+90)\cdot(90+410)}{68+72+81+90+432+428+419+410} \\ &= \frac{311 \cdot 500}{2000} = \frac{155500}{2000} = 77.75 \end{aligned}$$

So the correct answer is

3 □ 77.75

———— FACIT-END ————

Continues on page 36

Question X.2 (23)

A χ^2 distributed test statistic is used in order to test the hypothesis

$$H_0 : p_1 = p_2 = p_3 = p_4$$

What is the contribution $q_{No,2011}$ to the test statistic χ^2_{obs} from the respondents, who answer that they bought non-organic meat in 2011?

1* ☐ $q_{No,2011} = 0.2251$

2 ☐ $q_{No,2011} = 1.2227$

3 ☐ $q_{No,2011} = 9.75$

4 ☐ $q_{No,2011} = 0.0231$

5 ☐ $q_{No,2011} = 0.2201$

———— FACIT-BEGIN ————

From Method 7.21 we know that the χ^2-test statistic χ^2_{obs} is calculated as a sum

$$\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

The contribution from the respondents in 2011 who answered no:

$$q_{No,2011} = \frac{(432 - 422.25)^2}{422.25} = 0.2251$$

So the correct answer is

1 ☐ $q_{No,2011} = 0.2251$

———— FACIT-END ————

Question X.3 (24)

The supermarket chain has now conducted the test of the hypothesis

$$H_0 : p_1 = p_2 = p_3 = p_4$$

to track the trend in sales of organic meat.

In this case the relevant test statistic becomes 4.3977.

Which of the following R commands calculates the p-value for the hypothesis test?

1 ☐ `pchisq(4.3977, 3)`

2 ☐ `2*(1-pnorm(4.3977))`

3* ☐ `1-pchisq(4.3977, 3)`

4 ☐ `2*(1-pchisq(4.3977, 4))`

5 ☐ `1-pchisq(4.3977, 6)`

———— FACIT-BEGIN ————

From Method 7.21 we see that the test statistic should be compared with a χ^2-distribution with (2-1)(4-1)=3 degrees of freedom. The test probability is now:

```
  1-pchisq(4.3977, 3)

## [1] 0.2215987
```

So the correct answer is

3 ☐ `1-pchisq(4.3977, 3)`

Doing the analysis in R

```
 study <- matrix(c(  68 ,72, 81 ,90,432, 428, 419 , 410 ), nrow=2, byrow=TRUE)
  colnames(study) <- c("2011", "2012", "2013", "2014")
  rownames(study) <- c("Organic", "Non-Organic")
chi <- chisq.test(study); chi

##
##  Pearson's Chi-squared test
##
## data:  study
## X-squared = 4.3977, df = 3, p-value = 0.2216
```

———— FACIT-END ————

Continues on page 38

Exercise XI

Studies have shown that teenage girls have a lower life satisfaction than boys. Therefore, a team of first-year students decided to study life satisfaction among their peers. The results of their study were as follows.

	High life satisfaction	Lower life satisfaction
Men	68	208
Women	18	74

Question XI.1 (25)

What is the correct 95% confidence interval for the estimate of the difference between the proportion of high life satisfaction for men and women?

1 ☐ $(0.2464 - 0.1957) \pm 1.64 \cdot \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}} = (-0.029; 0.131)$

2 ☐ $(0.2464 - 0.1957) \pm 1.96 \cdot \sqrt{(\frac{0.2464(1-0.2464)}{276})^2 + (\frac{0.1957(1-0.1957)}{92})^2} = (0.047; 0.054)$

3 ☐ $\frac{0.2464}{0.1957} \pm 1.96 \cdot \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}} = (1.16; 1.35)$

4 ☐ $(0.2464 - 0.1957) \pm 3.84 \cdot \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}} = (-0.137; 0.238)$

5* ☐ $(0.2464 - 0.1957) \pm 1.96 \cdot \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}} = (-0.045; 0.146)$

———— FACIT-BEGIN ————

Let p_1 be the proportion of high life satisfaction in men and p_2 the proportion of high life satisfaction in women. According to Method 7.14 the 95% confidence interval for the estimated difference $\hat{p}_1 - \hat{p}_2$ is given as

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2}\hat{\sigma}_{\hat{p}_1-\hat{p}_2}$$

The estimates $\hat{p}_1$ and $\hat{p}_2$ are the observed proportions with high life satisfaction.

$$\begin{aligned} \hat{p}_1 &= \frac{68}{68+208} = 0.2464 \\ \hat{p}_2 &= \frac{18}{18+74} = 0.1957 \end{aligned}$$

The estimated standard error is

$$\hat{\sigma}_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

If we insert the number of men $n_1 = 276$ and the number of women $n_2 = 92$ then we get

$$\hat{\sigma}_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}}$$

Finally, notice $z_{1-\alpha/2} = 1.96$ is the 97.5% percentile in a standard normal distribution.

So the correct answer is

5 ☐ $(0.2464 - 0.1957) \pm 1.96 \cdot \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}} = (-0.045; 0.146)$

———— FACIT-END ————

Continues on page 40

Question XI.2 (26)

Now we want to test the hypothesis that the proportion of high life satisfaction is the samen for men and women. I.e. we testing the hypothesis (at significance level $\alpha = 0.05$).

$$\begin{aligned} H_0 : p_1 &= p_2 \\ H_A : p_1 &\neq p_2 \end{aligned}$$

Here p_1 is the proportion of high life satisfaction amongst men and p_2 is the proportion of high life satisfaction amongst women.

What is the conclusion to this test? (Both the conclusion and the argumentation must be correct).

1 ☐ H_0 is rejected, since the test statistic $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{0.2337(1-0.2337)(\frac{1}{276+92})}} = 2.298$ leads to a p-value of 0.02

2* ☐ H_0 is accepted, since the test statistic $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{0.2337(1-0.2337)(\frac{1}{276}+\frac{1}{92})}} = 0.995$ leads to a p-value of 0.32

3 ☐ H_0 is accepted, since the test statistic $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{\frac{0.2464(1-0.2464)}{276}+\frac{0.1957(1-0.1957)}{92}}} = 1.038$ leads to a p-value of 0.15

4 ☐ H_0 is accepted, since the test statistic $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{0.2337(1-0.2337)(\frac{1}{276+92})}} = 2.298$ leads to a p-value of 0.02

5 ☐ H_0 is rejected, since the test statistic $z_{obs} = \frac{(0.2464-0.1957)}{\frac{0.2464(1-0.2464)}{276}+\frac{0.1957(1-0.1957)}{92}} = 21.3$ leads to a p-value < 0.0001

———— FACIT-BEGIN ————

According to Method 7.17 the hypothesis is tested using the test statistic

$$z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$$

Here we have

$$\begin{aligned} n_1 &= 276 \\ n_2 &= 92 \\ \hat{p}_1 &= \frac{68}{68+208} = 0.2464 \\ \hat{p}_2 &= \frac{18}{18+74} = 0.1957 \\ \hat{p} &= \frac{68+18}{276+92} = 0.2337 \end{aligned}$$

So

$$z_{obs} = \frac{0.2464 - 0.1957}{\sqrt{0.2337(1-0.2337)(\frac{1}{276}+\frac{1}{92})}} = 0.995$$

We want a two-sided test and calculate $2P(Z > z_{obs})$

```
2*(1-pnorm(0.995))

## [1] 0.3197363
```

So the correct answer is

2 ☐ H_0 is accepted, since the test statistic $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{0.2337(1-0.2337)(\frac{1}{276}+\frac{1}{92})}} = 0.995$ leads to a p-value of 0.32

———— FACIT-END ————

Continues on page 42

Exercise XII

18 test persons evaluated the bass quality of 3 different headphones, so that all 18 persons evaluated all 3 headphones such that the data consist of 54 observations of bass quality on a scale between 0 and 150. The average of the three headphone bass qualities were:

Headphone	Average
1	53.5
2	55.5
3	97.1

Question XII.1 (27)

How is the $SS(Tr)$ calculated in the 2-way analysis of variance, which compares the mean bass quality for the three headphones? (Where "Tr" now refers to the 3 headphones)

1* ☐ $18 \cdot (53.5 - 68.7)^2 + 18 \cdot (55.5 - 68.7)^2 + 18 \cdot (97.1 - 68.7)^2$

2 ☐ $(53.5 - 68.7)^2 + (55.5 - 68.7)^2 + (97.1 - 68.7)^2$

3 ☐ $\frac{(53.5-68.7)^2}{53.5} + \frac{(55.5-68.7)^2}{55.5} + \frac{(97.1-68.7)^2}{55.5}$

4 ☐ $\frac{(53.5-68.7)}{53.5} + \frac{(55.5-68.7)}{55.5} + \frac{(97.1-68.7)}{55.5}$

5 ☐ $3 \cdot (53.5 - 68.7)^2 + 3 \cdot (55.5 - 68.7)^2 + 3 \cdot (97.1 - 68.7)^2$

———— FACIT-BEGIN ————

With 18 observations ($b = 18$) for each treatment in a 2-way ANOVA the defining formula for $SS(Tr)$ gives:

$$18 \cdot (53.5 - 68.7)^2 + 18 \cdot (55.5 - 68.7)^2 + 18 \cdot (97.1 - 68.7)^2$$

, since the mean of the three means become 68.7. So the correct answer is 1).

———— FACIT-END ————

Question XII.2 (28)

If, in line with the above, we let "persons" constitute "blocks", we are given that $SS(Bl) = 6003.5$ and that $SSE = 7160.3$ in the 2-way analysis of variance. What will the F-test statistic for the hypothesis that the 18 persons have the same mean value be?

1 ☐ $F_{obs} = \frac{18 \cdot 6003.5}{210.6/3}$

2 ☐ $F_{obs} = \frac{3 \cdot 6003.5}{7160.3/17}$

3* ☐ $F_{obs} = \frac{6003.5/17}{7160.3/34}$

4 ☐ $F_{obs} = \frac{(6003.5-210.6)^2}{7160.3}$

5 ☐ $F_{obs} = \frac{(6003.5/18-210.6)}{\sqrt{(210.6)}}$

———— FACIT-BEGIN ————

The F-statistic is

$$F_{obs,Bl} = \frac{MS(Bl)}{MSE} = \frac{6003.5/17}{210.6},$$

as the $MSE = 210.6$ (PBB: I will have to change this!!!!)

———— FACIT-END ————

Continues on page 44

Question XII.3 (29)

The hypothesis of no difference in mean bass quality of the three headphones is by the usual test evaluated by which sampling distribution?

1 ☐ z-distribution (= standard normal distribution)

2 ☐ t-distribution with 53 degrees of freedom

3 ☐ χ^2-distribution with 53 degrees of freedom

4* ☐ F-distribution with 2 and 34 degrees of freedom

5 ☐ F-distribution with 3 and 51 degrees of freedom

———— FACIT-BEGIN ————

According to Theorem 8.22 the right sampling distrubution is the F-distribution with $l-1 = 17$ and $(k-1)(l-1) = 2 \cdot 17 = 34$, so the correct answer is 4).

———— FACIT-END ————

Question XII.4 (30)

What will the 95% confidence interval be for the mean difference between headphone 2 and 1? (It can be assumed that this is a "pre-planned" comparison)

1 ☐ $2 \pm 2 \cdot 210.6$

2 ☐ $2 \pm 2.03 \cdot \sqrt{210.6}$

3 ☐ $2 \pm 1.96 \cdot \frac{210.6}{54}$

4 ☐ 2 ± 1.96

5* ☐ $2 \pm 2.03 \cdot \sqrt{2 \cdot 210.6 \frac{1}{18}}$

———— FACIT-BEGIN ————

We use the post hoc method box for oneway anova combined with the 2-way adaption:

1. Use the MSE and/or SSE from the two-way analysis
2. Use $(l-1)(k-1)$ as denominator DF

So:

$$2 \pm 2.03 \cdot \sqrt{2 \cdot 210.6 \frac{1}{18}}$$

So the correct answer is 5).

———— FACIT-END ————

THE EXAM IS FINISHED. ENJOY THE SUMMER!

Technical University of Denmark *Page 1 of 43 pages.*

Written examination: 20. August 2017

Course name and number: **Introduction to Statistics (02323 og 02402)**

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

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

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

The questionnaire contains 43 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 swimming team goes on an weekly training camp with a focus on training the swimming stroke front crawl. A test is carried out where the time, for each swimmer swimming the same distance in front crawl, is measured. The test is carried out before and after the camp.

The measured times are stored (in the same order for the swimmers) in the following vectors in R: `before` holds the times before and `after` holds the times after the training camp.

The following hypothesis must be tested

$$\mu_{\text{after}} - \mu_{\text{before}} = 0$$
$$\mu_{\text{after}} - \mu_{\text{before}} \neq 0$$

where μ_{before} and μ_{after} denotes the mean times for the entire team before and after the camp.

Question I.1 (1)

Which of the following R-calls correctly calculates the p-value for a t-test of the hypothesis?

1 ☐ `t.test(after, before, mu=0)`

2 ☐ `t.test(after, before, mu=-10)`

3 ☐ `t.test(after, before, mu=10)`

4 ☐ `t.test(after, mu=10)`

5* ☐ `t.test(after-before, mu=0)`

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

See Section 3.2.3. Since there is a measurement for each swimmer before and after the camp the correct way to analyze the data is a paired t-test, and they are ordered such that the time for each swimmer is in the same place in `before` as in `after`. The paired analysis is carried out by using a single-sample t-test on the differences, this is done in Answer 5. Answer 1 to 3 assume independent samples (non-paired) and Answer 4 only test the speed after the camp.

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

Question I.2 (2)

The p-value of the test was calculated to 0.00287. Can the null hypothesis be rejected at significance level $\alpha = 5\%$ (both conclusion and argument must be correct)?

1* ☐ Yes, since the p-value is below the significance level the null hypothesis is rejected

2 ☐ No, since the p-value is below the significance level the null hypothesis is accepted

3 ☐ Yes, since the p-value is over the significance level the null hypothesis is rejected

4 ☐ No, since the p-value is over the significance level the null hypothesis is accepted

5 ☐ More information is needed in order to decide against the null hypothesis

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

Since the p-value is less than the significance level ($0.00287 < 0.05$), the null hypothesis is rejected (See Method 3.36)

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

Question I.3 (3)

Each day at the training camp, there is a random drawing about who should do the dishes. There must be 4 each day for doing the dishes and there are in total 35 participants. For each participant there is equally high probability of being drawn each day. Calculate the probability that a participant will not do the dishes at all during training camp, which includes 7 evenings with dish washing.

1 ☐ 1 - $\binom{7}{0} \cdot 0.144^0 \cdot (1 - 0.144)^{7-0} = 0.57$

2 ☐ $\binom{5}{2} \cdot 0.144^2 \cdot (1 - 0.144)^{5-2} = 0.09$

3 ☐ $\binom{7}{7} \cdot 0.798^7 \cdot (1 - 0.798)^{7-7} = 0.21$

4* ☐ $\binom{7}{7} \cdot 0.886^7 \cdot (1 - 0.886)^{7-7} = 0.43$

5 ☐ $\binom{5}{2} \cdot 0.886^2 \cdot (1 - 0.886)^{5-2} = 0.01$

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

See Definition 2.20. The probability that a participant does not have to do the dishes at a specific day is $1 - \frac{4}{35} = 0.886$. Since the probability is the same every day, these must be

independent draws and the probability distribution for the number of times a participant have do the dishes is a binomial with $n = 7$ and $p = 0.886$. Hence the probability can be calculated by

$$\binom{7}{7} \cdot 0.886^7 \cdot (1 - 0.886)^{7-7} = 0.43 \tag{1}$$

or in R by

```
## Sandsynligheden for ikke at bliver trukket alle 7 dage
dbinom(7, 7, 1-4/35)

## [1] 0.4276176
```

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

Continues on page 5

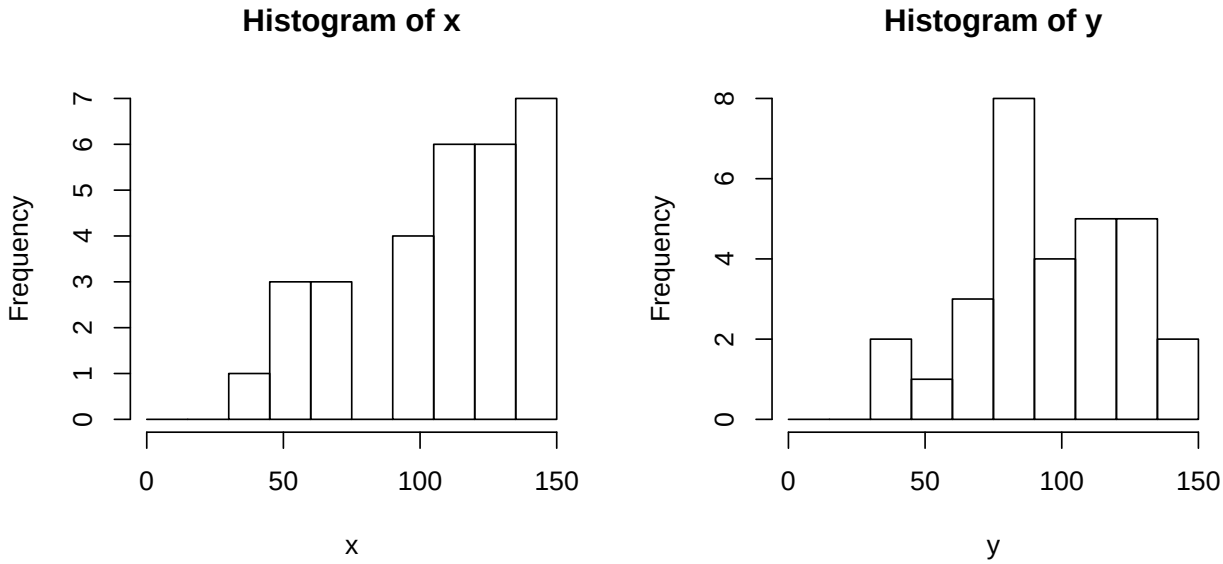
Exercise II

In connection with the exam in Introduction to Statistics, it is desired to examine whether foreign students are doing well. The score of the exam is calculated as a number between -30 and 150, as there are 30 questions and a wrong answer gives -1 point and a correct answer gives 5 points. There can only be given one answer to each question.

Two random samples of the score has been taken: one for foreign students (`x`) and one for Danish students (`y`). Each sample has 30 observations.

Question II.1 (4)

To assess the most appropriate analysis, a histogram is plotted of each sample:



What is the most appropriate statement based on the given information?

1 ☐ Nothing indicates that the samples don't come from symmetrical distributed populations

2 ☐ The samples cannot be assumed to come from symmetrical distributed populations. This is supported by the histograms, in particular the distribution of `x` appears to be right-skewed

3* ☐ The samples cannot be assumed to come from symmetrical distributed population. This is supported by the histograms, in particular the distribution of `x` appear to be left-skewed

4 ☐ The populations from which the samples are taken can both be assumed to be exponentially distributed

5 ☐ The populations from which the samples are taken can both be assumed to be normally distributed

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

- The histogram of `x` show that the empirical distribution is highly skewed, hence Answer 1 is wrong
- The `x` data is left skewed (since the mean is smaller than the median) hence Answer 2 is wrong
- Answer 3 is correct (see the arguments for 1 and 2 being wrong)
- Exponential data is right-skrewed, hence Answer 4 is wrong
- Normally distributed data is symmetrical hence 5 is wrong

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

Question II.2 (5)

It is decided that the best analysis is included in the following `R` code:

```
## Number of simulations
k <- 10000
## Simulate each sample k times
simxsamples <- replicate(k, sample(x, replace=TRUE))
simysamples <- replicate(k, sample(y, replace=TRUE))
## Calculate the sample mean differences
simmeandifs <- apply(simxsamples,2,mean) - apply(simysamples,2,mean)
## Quantiles of the differences gives the CI
quantile(simmeandifs, c(0.005,0.995))
```

```
##  0.5% 99.5%
## -9.23 31.63
```

```
quantile(simmeandifs, c(0.025,0.975))
```

```
##   2.5%  97.5%
## -4.125 26.106
```

```
## CI for the median differences
simmediandifs <- apply(simxsamples,2,median) - apply(simysamples,2,median)
quantile(simmediandifs, c(0.005,0.995))
```

```
##   0.5%  99.5%
## -10.42  43.05
```

```
quantile(simmediandifs, c(0.025,0.975))
```

```
##   2.5%  97.5%
## -3.975 39.525
```

Which of the following statements is correct?

1* ☐ Non-parametric bootstrap confidence intervals have been calculated for differences between two populations

2 ☐ Parametric bootstrap confidence intervals have been calculated for differences between two populations

3 ☐ Confidence intervals for differences between two populations have been calculated under the assumption of normal distributions

4 ☐ Confidence intervals for differences between two populations have been calculated under the assumption of exponential distributions

5 ☐ Confidence intervals for differences between two populations have been calculated under the assumption of Poisson distributions

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

The `R` code calculates non-parametric bootstrap confidence intervals for the differences in scores between the two populations (DK and foreign students). It is done both for the mean (with levels 99% and 95%), and the same confidence intervals for the median. It is non-parametric because no assumption about the distribution is made, which is carried out by sampling directly from the observations with the `sample()` funtion (instead of e.g. using `rnorm()` which would be under assumption of normal distribution).

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

Question II.3 (6)

The following hypothesis should be tested at significance level $\alpha = 5\%$

$$H_0 : q_{0.5,\mathtt{x}} = q_{0.5,\mathtt{y}}$$
$$H_1 : q_{0.5,\mathtt{x}} \neq q_{0.5,\mathtt{y}}$$

where $q_{0.5,\mathtt{x}}$ denotes the 50% quantile for foreign students and $q_{0.5,\mathtt{y}}$ denotes the 50% quantile for Danish students.

Which of the following statements is correct (not all of the statements are necessarily meaningful)?

1 ☐ H_0 is rejected and it can be concluded that Danish students perform significantly better than foreign students at the indicated level of significance

2 ☐ H_0 is rejected and it can be concluded that foreign students perform significantly better than Danish students at the indicated level of significance

3* ☐ H_0 is not rejected and it cannot be concluded that Danish students perform significantly different than foreign students at the indicated level of significance

4 ☐ H_0 is not rejected and it can be concluded that Danish students perform significantly different than foreign students at the indicated level of significance

5 ☐ None of the above statements are correct

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

The `R` code for the previous question gives the 95% confidence interval for the median as $[-3.975; 39.525]$, hence the null hypothesis is not rejected at level $\alpha = 0.05$ since $(0 \in [-3.975; 39.525])$. It's close to the boundary, so maybe with a slightly larger sample we would have concluded a significant difference, but we have not and must accept that.

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

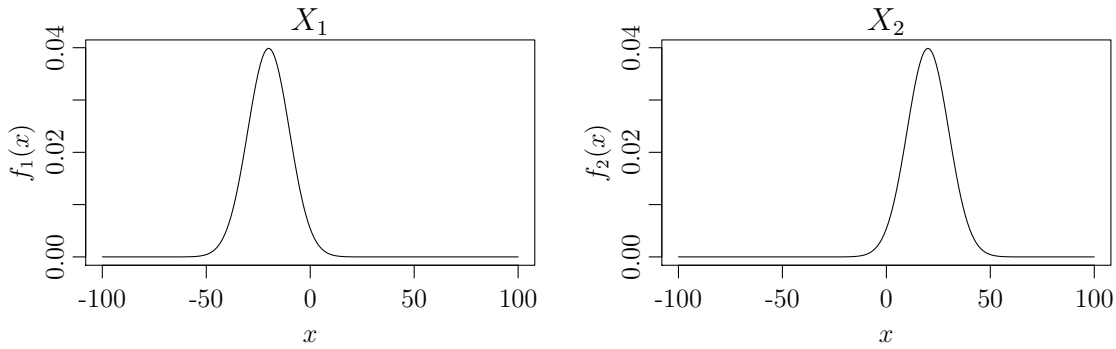
Continues on page 9

Exercise III

Let two independent random variables be given by

$$X_1 \sim N(-20, 10^2) \quad \text{and} \quad X_2 \sim N(20, 10^2).$$

Their probability density functions (pdfs) are then:

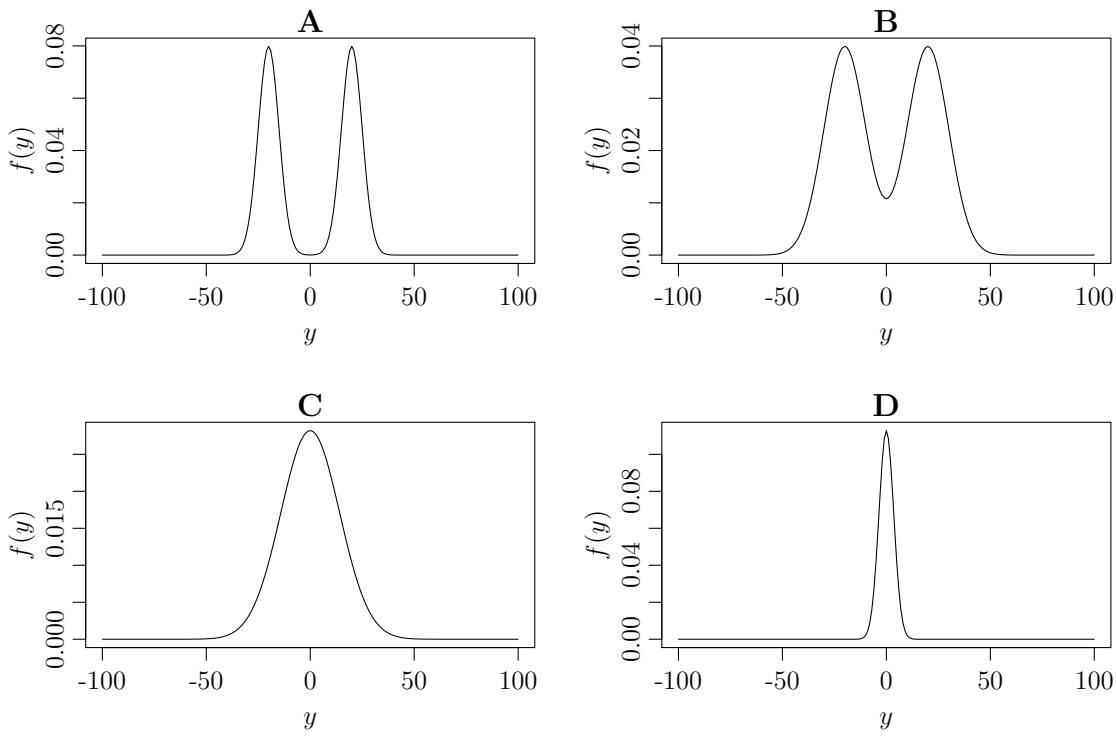


Question III.1 (7)

Now a new random variable is defined by

$$Y = X_1 + X_2.$$

Which of the following plots is then the pdf for Y?



1 ☐ Plot A

2 ☐ Plot B

3* ☐ Plot C

4 ☐ Plot D

5 ☐ None of the shown plots can be close to the pdf of Y

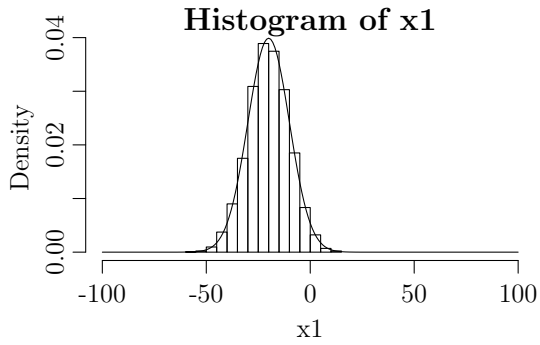
---------- FACIT-BEGIN ----------

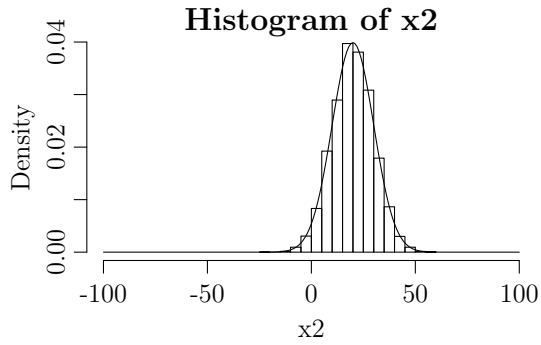
First of all, as stated in Theorem 2.40, sums of independent normal random variables are normal random variables (this exclude plots A and B). Also from Theorem 2.56 we have that

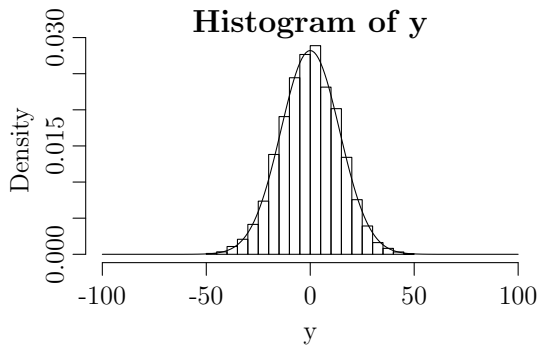
$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[X_1] + \mathrm{E}[X_2] = 0, \\ \mathrm{V}[Y] &= \mathrm{V}[X_1] + \mathrm{V}[X_2] = 200. \end{aligned}$$

The variance of Y is greater than the variance of X_1 and X_2 hence we can exclude Plot D, and the only remaining option is Plot C that have a larger variance the X_1 and X_2, hence plot C is correct. This can be confirmed with simulation:

```
## (x1+x2) = var(x1) + var(x2) = 100 + 100 = 200
x1 <- rnorm(n, mean=-20, sd=10)
x2 <- rnorm(n, mean=20, sd=10)
y <- x1 + x2
par(mfrow=c(2,2), mgp=c(1.6,0.5,0), mar=c(4,3,1,1), tcl=-0.4)
hist(x1, xlim=c(xmin,xmax), prob=TRUE)
lines(xseq, dnorm(xseq,mean=-20,sd=10), type="l")
hist(x2, xlim=c(xmin,xmax), prob=TRUE)
lines(xseq, dnorm(xseq,mean=20,sd=10), type="l")
hist(y, xlim=c(xmin,xmax), prob=TRUE)
lines(xseq, dnorm(xseq,mean=0,sd=sqrt(200)), type="l")
```







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

Question III.2 (8)

Assuming X_1 and X_2 each represent a population and the test for difference in mean value with the commonly used non-paired t-test should be carried out. What is the smallest sample size $n = n_1 = n_2$ that must be taken from each population, at significance level $\alpha = 5\%$, in order to achieve a power of the test of at least 99%?

1* ☐ $n = 4$ observations in each sample

2 ☐ $n = 12$ observations in each sample

3 ☐ $n = 38$ observations in each sample

4 ☐ $n = 69$ observations in each sample

5 ☐ $n = 248$ observations in each sample

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

The difference in mean of the two distributions is 40, and the standard deviation in each os the two groups is 10, hence we can find the the number of observation needed with:

```r
power.t.test(delta=40, sd=10, sig.level=0.05, power=0.99)
```

```
##
##      Two-sample t test power calculation
##
##               n = 3.644287
##           delta = 40
##              sd = 10
##       sig.level = 0.05
##           power = 0.99
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

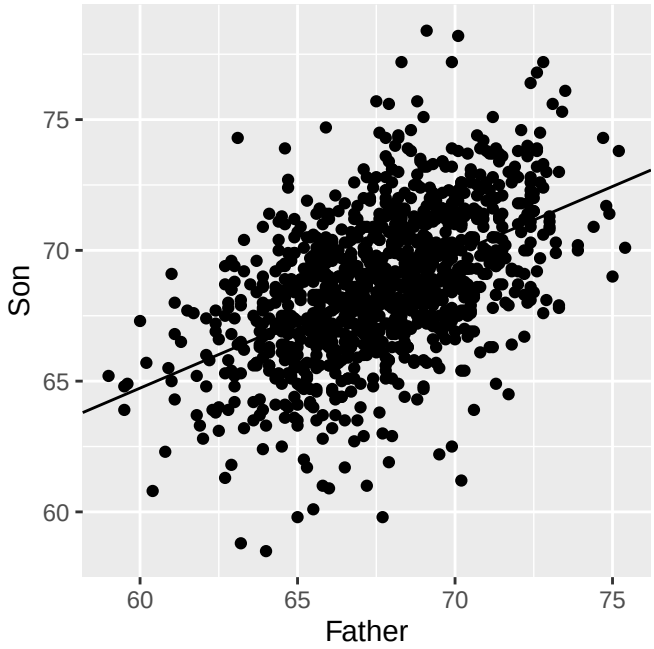
hence the correct answer is $n = 4$, since we must round up to nearest integer. See more in Section 3.3

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

Continues on page 13

Exercise IV

The figure below shows the relation between the height of about 1000 fathers and their sons measured in inches:



The shown regression line describes the fit of the following model

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.},$$

where Y_i is the height of the i'th son and x_i is the height of the i'th father.

Question IV.1 (9)

Which of the following statements is a correct description of the regression line?

1 ☐ The line describes an estimate of the mean height of the sons as a function of their fathers mean height

2 ☐ The line describes an estimate of the linear correlation between the average height of father and son

3 ☐ The line describes an estimate of the fathers mean height as a function of the height of their sons

4* ☐ The line describes an estimate of the sons mean height as a function of the height of their fathers

5 ☐ The line describes the height of a son as a function of the height of the father

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

- The regression line show an estimate of the mean (or expected) height of sons as a function fathers height (not mean height) hence Answer 1 is wrong
- The estimation of correlation cannot be directly derived from the regression line, hence Answer 2 is wrong
- In Answer 3 the relation is reversed (hence it is wrong)
- Answer 4 correctly states that the line describe the mean (or expected) height of sons as a function of fathers heights
- The line only describe an estimate of the sons' mean height (hence the points are scattered around the line), not the actual height. So answer 5 is wrong.

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

Question IV.2 (10)

It is chosen to analyze the data with the following `R` code, where `fs` is a data frame with the columns `Son` and `Father` holding the observed heights:

```
summary(fit <- lm(Son ~ Father, data=fs))
```

Which gives the following result where two numbers are replaced by letters:

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280          A   18.49   <2e-16 ***
Father       0.51401          B   19.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512, Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

How large a proportion of the variation in the height of sons is not explained by the height of the fathers?

1 ☐ Approximately 25%

2 ☐ Approximately 50%

3* ☐ Approximately 75%

4 ☐ Approximately 86.5%

5 ☐ Approximately 66%

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

The proportion of explaned variation is the multiple R^2 value, which can be read as 0.2512 or approximately 25% from the R output, hence the variation not explained will is approximately 75%.

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

Continues on page 16

Question IV.3 (11)

What is the estimate of the standard deviation of the coefficient for `Father`?

1 ☐ $\hat{\sigma}_{\beta_1} = 0.514/2.438 = 0.211$

2 ☐ $\hat{\sigma}_{\beta_1} = 2.438/1076 = 0.00227$

3 ☐ $\hat{\sigma}_{\beta_1} = 0.514 \cdot 19.00 = 9.77$

4 ☐ $\hat{\sigma}_{\beta_1} = 33.89/18.49 = 1.83$

5* ☐ $\hat{\sigma}_{\beta_1} = 0.514/19.0 = 0.027$

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

From Theorem 5.12 we have the formula for the test statistic

$$T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}},$$

and we know that the default t-test printed out by `summary()` is with the null hypothesis that the slope is 0:

$$H_0 : \beta_{0,1} = 0$$

so we can write

$$t_{\beta_1,\text{obs}} = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\beta_1}},$$

where $t_{\beta_1,\text{obs}}$ is the observed T_{β_1} which is seen from the R output to be 19.0.

Rearranged it gives and values found in the result

$$\hat{\sigma}_{\beta_1} = \frac{\hat{\beta}_1 - 0}{t_{\beta_1,\text{obs}}} = \frac{0.514}{19.0} = 0.027.$$

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

Question IV.4 (12)

Given the following calculations in `R`, what is a 95% confidence interval for the mean height of sons of fathers who are 75 inches tall?

```
mean(fs$Father); var(fs$Father)

## [1] 67.68683
## [1] 7.539566

mean(fs$Son); var(fs$Son)

## [1] 68.68423
## [1] 7.930949
```

1* ☐ $33.893 + 0.514 \cdot 75 \pm 1.96 \cdot 2.438 \cdot \sqrt{\frac{1}{1078} + \frac{(75-67.687)^2}{7.540 \cdot (1078-1)}}$

2 ☐ $33.893 + 0.514 \cdot 75 \pm 1.96 \cdot 2.438 \cdot \sqrt{\frac{1}{1078} + \frac{(75-67.687)^2}{7.540}}$

3 ☐ $33.893 + 0.514 \cdot 75 \pm 1.96 \cdot 2.438^2 \cdot \sqrt{\frac{1}{1078} + \frac{(75-67.687)^2}{7.540 \cdot (1078-1)}}$

4 ☐ $33.893 + 0.514 \cdot 75 \pm 1.96 \cdot 2.438 \cdot \sqrt{\frac{1}{1077} + \frac{(75-67.687)^2}{7.540 \cdot (1077-1)}}$

5 ☐ $33.893 + 0.514 \cdot 75 \pm 1.65 \cdot 2.438^2 \cdot \sqrt{\frac{1}{1077} + \frac{(75-67.687)^2}{7.540 \cdot (1077-1)}}$

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

The general formula for the confidence interval for a point on the line is given in Method 5.18

$$\beta_0 + \beta_1 x_{\text{new}} \pm t_{\alpha/2} \cdot \sigma \cdot \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}$$

we can find $\beta_0 = 33.89$, $\beta_1 = 0.514$, $\sigma = 2.438$, and $n = 1076 + 2 = 1078$ in the `summary` output above. In addition we got $\bar{x} = 67.69$ in the calculation above, note also that $S_{xx} = (n-1)s_x^2$ where s_x^2 is the empirical variance of farthers height, hence $S_{xx} = 1077 \cdot 7.540$. The degrees of fredom for the t-distibution is $n - 2 = 1076$ which gives $t_{0.975} = 1.96$ Inserting the numbers we get Answer 1 as the correct answer. This can also be checked by the folllowing R calculations.

```
n <- 1078
sxx <- 7.540 * (n-1)
xnew=75
xbar <- 67.687
sigma <- 2.438
beta0 <- 33.893
beta1 <- 0.514

round(beta0 + beta1*xnew + c(-1, 1) * qt(0.975, df=n-2) * sigma *
  sqrt((1/n) + ((xnew - xbar)^2/sxx)), 2)
```

```
## [1] 72.03 72.86

round(predict(fit, data.frame("Father"=75), interval="conf"), 2) # Check

##     fit   lwr   upr
## 1 72.44 72.03 72.86
```

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

Question IV.5 (13)

Now information about each family's monthly income is obtained and the following model is setup

$$Y_i = \beta_0 + \beta_1 \cdot x_{1,i} + \beta_2 \cdot x_{2,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.},$$

where Y_i is the height of the i'th son, $x_{1,i}$ is height of the i'th father, and $x_{2,i}$ is the income for the i'th family.

Under the following two assumptions:

- Rich families eat better and a better diet has a significant positive effect, which gives the sons of the family a higher growth
- There is independence between the father's height and the family's income

what is the consequence of adding the income into the model (not all answers are necessarily meaningful)?

1* ☐ Inclusion of income in the model will contribute to reducing the residual variance ($\hat{\sigma}^2$) and the uncertainty of the regression coefficient for the father's height (β_1) will be reduced

2 ☐ Inclusion of income in the model will contribute to reducing the residual variance ($\hat{\sigma}^2$), but this will not affect the uncertainty of the regression coefficient for the father's height (β_1)

3 ☐ As the fathers height is independent of the fathers income, the inclusion of income in the model will not affect the estimate of β_1 or the uncertainty of it

4 ☐ Inclusion of income in the model will use one more degree of freedom, such that a confidence interval for β_1 may be expected to be wider than if incomes were not included in the model

5 ☐ One must expect a high degree of multicollinearity between the estimates of β_1 and β_2, so the model must be reduced to a simple linear regression model

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

Lets go through the possibilities one by one:

- Answer 1: Including an effect that has a significant effect will reduce the residual variation, and with independence between fathers height and family income the unceartaity for β_1 will be reduced. Hence 1 is correct

- Answer 2: The residual variantion has a direct effect on the uncertainty of the parameters, hence 2 cannot be correct
- Answer 3: With the agrument in ans 2, this cannot be correct either
- Answer 4: The effect of using a degree of freedom is very small, hence the effect of reducing the variance will dominate, hence 4 is not correct
- Answer 5: Since we assume tat these are independent we will not expect multicollinarity

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

Continues on page 21

Exercise V

In humans there are a variety of different genetic determined blood type systems. The most well-known are probably the AB0- and Rhesus-systems. Another blood type system is the so-called MN blood type system, which is determined by a single gene Glycophorin A (GPA). In the GPA-gene there are two alleles M and N, such that a human may have the genotype (blood type) MM, MN, or NN.

The distribution of blood types in the MN blood type system is now seeked estimated from a sample of volunteer students of two different Philippine universities. One university, University of the Philippines-Diliman, here shortened UPD, is the country's largest university where students come from all over the country. The second university, Isabela State University, here abbreviated ISU, is a small university where the students primarily come from the local area. The following table lists the distribution of genotypes among the students in the samples from the two universities:

Bloodtype	UDP	ISU
MM	19	43
MN	15	7
NN	17	9

Question V.1 (14)

State the χ^2 test statistic and the conclusion of the test in which the MN blood type distribution in the two universities are compared (both test size and conclusion must be correct).

1 ☐ The test statistic is $\chi^2 = 14.15$, its distribution has 2 degrees of freedom and the test shows that there is some evidence for a difference in the MN blood type distribution at the two universities

2* ☐ The test statistic is $\chi^2 = 14.15$, its distribution has 2 degrees of freedom and the test shows that there is very strong evidence for a difference in the MN blood type distribution at the two universities

3 ☐ The test statistic is $\chi^2 = 3.76$, its distribution has 2 degrees of freedom and the test shows that there is not found any evidence for a difference in the MN blood type distribution at the two universities

4 ☐ The test statistic is $\chi^2 = 4.57$, its distribution has 1 degree of freedom and the test shows that there is evidence for a difference in the MN blood type distribution at the two universities

5 ☐ The test statistic is $\chi^2 = 3.76$, its distribution has 1 degree of freedom and the test shows that there is weak evidence for a difference in the MN blood type distribution at the two universities

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

The easiest way to solve this is by using `chisq.test`:

```
## Define the table
Bloodtype <- matrix(c(19,43,15,7,17,9),nrow=3,byrow=T)
colnames(Bloodtype) <- c("UDP","ISU")
rownames(Bloodtype) <- c("MM","MN","NN")
## Answer:
chisq.test(Bloodtype)

##
##  Pearson's Chi-squared test
##
## data:  Bloodtype
## X-squared = 14.154, df = 2, p-value = 0.0008443
```

We can see that the test-statistics is 14.15 and the degrees of freedom is 2, the p-value is 0.00084, and hence very strong evidence against the null hypothesis. The null hypothesis being that there is no difference in the distribution of blood cells between the universities. This is Answer 2.

We could also solve this by "hand"-calculations

```
## By hand:
mat <- Bloodtype
Exp <- rowSums(mat) %o% colSums(mat) / sum(mat)
Chisq.val <- sum((mat - Exp)^2 / Exp)
df <- prod(dim(mat) - 1)
pchisq(Chisq.val, df, lower=FALSE)

## [1] 0.0008443032
```

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

Question V.2 (15)

A biological population is said to be in Hardy-Weinberg (HW) equilibrium if the proportion of genotypes can be written as

$$
\begin{aligned}
p_{\mathrm{MM}} &= p^2, \\
p_{\mathrm{MN}} &= 2pq, \\
p_{\mathrm{NN}} &= q^2.
\end{aligned}
$$

Where p and q are the allele frequencies for M and N, respectively. They are calculated by

$$p = \frac{2 \cdot X_{\text{MM}} + X_{\text{MN}}}{2n},$$
$$q = \frac{2 \cdot X_{\text{NN}} + X_{\text{MN}}}{2n},$$

where $X_{\text{bloodtype}}$ is the observed number of the blood type and n is the sample size. Thus for UDP

$$p_{\text{MN}} = 2 \cdot \frac{2 \cdot X_{\text{MM}} + X_{\text{MN}}}{2n} \cdot \frac{2 \cdot X_{\text{NN}} + X_{\text{MN}}}{2n} = 0.4992,$$

is set as the proportion of MN blood type under HW-equilibrium.

A simple test to decide whether the population on UDP is not in HW-equilibrium can therefore be of the hypothesis

$$H_0 : p_{\text{MN,UDP}} = 0.4992$$
$$H_1 : p_{\text{MN,UDP}} \neq 0.4992$$

i.e. if the observed proportion of MN blood type on UDP $p_{\text{MN,UDP}}$ is equal to the proportion under HW-equilibrium.

We want to test whether it can be rejected that the genotypes on UDP are in HW-equilibrium. What is the usually applied test statistic for this test?

1 ☐ The test statistic is $\chi^2 = 2(1.99 + 4.30 + 2.32) = 17.2$

2* ☐ The test statistic is $z_{\text{obs}} = \frac{15-25.46}{\sqrt{25.46 \cdot (1-\frac{25.46}{51})}} = -2.93$

3 ☐ The test statistic is $z_{\text{obs}} = \frac{15-51}{\sqrt{51 \cdot (1-\frac{15}{51})}} = -6.00$

4 ☐ The test statistic is $\chi^2 = (1.99^2 + 4.30^2 + 2.32^2)/2 = 13.9$

5 ☐ The test statistic is $\chi^2 = 1.99 + 4.30 + 2.32 = 8.6$

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

This is a large sample test for a single proportion so we use the standard normal distribution (as stated in Theorem 7.10)

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

and the values are found for UDP and inserted

$$np_0 = 0.4992 \cdot 51 = 25.46,$$
$$z_{\text{obs}} = \frac{15 - 25.46}{\sqrt{25.46(1 - \frac{25.46}{51})}} = \frac{15 - 25.46}{\sqrt{25.46(1 - \frac{25.46}{51})}} = -2.93.$$

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

Question V.3 (16)

For another type of test for HW-equilibrium the test statistic is found to $\chi^2 = 24.52$ and under the null hypothesis it will follow a χ^2-distribution with 1 degree of freedom. What is the p-value and conclusion of the test using a significance level of 0.001?

1 ☐ p-value is `pchisq(24.52, df=1)` ≈ 1 and the hypothesis of HW-equilibrium cannot be rejected

2 ☐ p-value is `1 - pchisq(24.52, df=1)` < 0.001 and the hypothesis of HW-equilibrium cannot be rejected

3* ☐ p-value is `1 - pchisq(24.52, df=1)`< 0.001 and the hypothesis of HW-equilibrium is rejected

4 ☐ p-value is `1 - pnorm(sqrt(24.52))` < 0.001 and the hypothesis of HW-equilibrium cannot be rejected

5 ☐ p-value is `1 - pnorm(sqrt(24.52))` < 0.001 and the hypothesis of HW-equilibrium is rejected

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

The solution is to find that the correct p-value is calculated in `R` by:

```
## Either
pchisq(24.52, df=1, lower.tail=FALSE)

## [1] 7.354249e-07

## Or
1 - pchisq(24.52, df=1)

## [1] 7.354249e-07
```

so the p-value is below 0.001 and thus the conclusion is that the null hypothesis is rejected.

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

Question V.4 (17)

For theoretical reasons it has been suggested that the frequencies of genotypes MM and NN in the underlying population are the same and it is now of interest to investigate this on the basis of the observations from UDP. Assuming that the proportions for MM and NN are independent a 90% confidence interval for the difference in the proportion of MM and NN ($p_{\text{MM,UDP}} - p_{\text{NN,UDP}}$) is given by:

1* ☐ $2/51 \pm 1.64\sqrt{\frac{19\cdot 32}{51^3} + \frac{17\cdot 34}{51^3}}$

2 ☐ $2/51 \pm 1.96\sqrt{\frac{19\cdot 32}{51^3} + \frac{17\cdot 34}{51^3}}$

3 ☐ $2/51 \pm 1.64\sqrt{\frac{19\cdot 32}{51^2} + \frac{17\cdot 34}{51^2}}$

4 ☐ $2/51 \pm 1.96\sqrt{\frac{19\cdot 32}{51^2} + \frac{17\cdot 34}{51^2}}$

5 ☐ $2/51 \pm 1.68\sqrt{\frac{19\cdot 32}{51^3} + \frac{17\cdot 34}{51^3}}$

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

We use Theorem 7.15 and insert the values:

```
## Answer ("manually"):
sd.pid <- sqrt(19*32/51^3 + 17*34/51^3)
CI <- (19/51 - 17/51) + c(-1, 1) * qnorm(0.95) * sd.pid
round(CI, 3)

## [1] -0.116  0.195
```

Notice that $\sqrt{\frac{19\cdot 32}{51^3} + \frac{17\cdot 34}{51^3}} = \sqrt{\frac{\frac{19}{51}(1-\frac{19}{51})}{51} + \frac{\frac{17}{51}(1-\frac{17}{51})}{51}}$

Alternative we use the prob.test in R as shown in example 7.19

```
## Answer using prob.test
event <- c(19,17)
n <- c(51,51)
prop.test(event, n, conf.level=0.90, correct=FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  event out of n
## X-squared = 0.17172, df = 1, p-value = 0.6786
```

```
## alternative hypothesis: two.sided
## 90 percent confidence interval:
##  -0.1163144  0.1947457
## sample estimates:
##    prop 1    prop 2
## 0.3725490 0.3333333
```

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

Continues on page 27

Question V.5 (18)

What is the usual test statistic for the test that the proportions of MM and NN are equal at UDP, i.e. $H_0 : p_{\text{MM,UDP}} = p_{\text{NN,UDP}}$?

1 ☐ $z_{\text{obs}} = \frac{2}{51\sqrt{\frac{19\cdot 32}{51^2}+\frac{17\cdot 34}{51^2}}}$

2 ☐ $z_{\text{obs}} = \frac{2/51}{\sqrt{\frac{19\cdot 32}{51^3}+\frac{17\cdot 34}{51^3}}}$

3* ☐ $z_{\text{obs}} = \frac{2}{51\sqrt{\frac{6}{17}\frac{11}{17}\frac{2}{51}}}$

4 ☐ $z_{\text{obs}} = \frac{2/51}{\sqrt{\frac{6}{17}\frac{6}{19}\frac{2}{51}}}$

5 ☐ $z_{\text{obs}} = \frac{2/51}{\sqrt{\frac{19\cdot 34}{51^3}+\frac{17\cdot 32}{51^3}}}$

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

Use Method 7.18 and insert the values:

```
x1 <- 17;
x2 <- 19;
n1 <- n2 <- 15+19+17
p1 <- 17/n1
p2 <- 19/n2
(delta.p <- p2 - p1)

## [1] 0.03921569

(phat <- (17+19)/(51+51))

## [1] 0.3529412

(zobs <- delta.p / sqrt(phat*(1-phat)*(1/n1 + 1/n2)))

## [1] 0.4143877

## The formula from the answer give the same answer
2/51/sqrt((6/17)*(11/17)*(2/51))

## [1] 0.4143877
```

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

Continues on page 28

Exercise VI

A sample with the following 10 observations is taken:

```
x <- c(-1.63, -1.37, -1.21, -0.60, -0.36, -0.26, -0.18, 0.02, 0.29, 0.39)
```

Notice that the observations have been sorted in the code above.

The sample mean and sample standard deviation are calculated:

```
mean(x)

## [1] -0.491

sd(x)

## [1] 0.7003
```

Question VI.1 (19)

What is the sample variance?

1 ☐ $s^2 = 0.21$

2* ☐ $s^2 = 0.49$

3 ☐ $s^2 = 1.46$

4 ☐ $s^2 = 1.70$

5 ☐ $s^2 = 2.36$

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

The sample variance is simply the squared standard deviation (Definition 1.11) $s^2 = 0.07^2 = 0.49$, or in R:

```
var(x)

## [1] 0.4904
```

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

Question VI.2 (20)

What is the first quartile of the sample?

1 ☐ $Q_1 = -1.37$

2 ☐ $Q_1 = -1.29$

3* ☐ $Q_1 = -1.21$

4 ☐ $Q_1 = -0.91$

5 ☐ $Q_1 = -0.60$

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

See Definition 1.7. With n=10 observations we get $pn = 0.25 \cdot 10 = 2.5$ and hence the first quartile is $Q_1 = x_{(3)} = -1.21$.

In R we can get this by:

```
quantile(x, type=2,prob=0.25)

##   25%
## -1.21
```

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

Question VI.3 (21)

Which of the following is a correct 95% confidence interval for the mean of the population from which the sample is taken?

1 ☐ $-0.491 \pm t_{0.975}\frac{0.490}{\sqrt{10}} = [-0.84, -0.14]$ where $t_{0.975} = 2.26$ is a quantile in t-distribution with 9 degrees of freedom

2 ☐ $-0.491 \pm t_{0.95}\frac{0.700}{\sqrt{9}} = [-0.92, -0.64]$ where $t_{0.95} = 1.83$ is a quantile in t-distribution with 9 degrees of freedom

3 ☐ $-0.491 \pm t_{0.95}\frac{0.490}{\sqrt{9}} = [-0.79, -0.19]$ where $t_{0.95} = 1.83$ is a quantile in t-distribution with 9 degrees of freedom

4 ☐ $-0.491 \pm t_{0.975}\frac{0.700}{10} = [-0.65, -0.33]$ where $t_{0.975} = 2.26$ is a quantile in t-distribution with 9 degrees of freedom

5* ☐ $-0.491 \pm t_{0.975}\frac{0.700}{\sqrt{10}} = [-0.99, 0.01]$ where $t_{0.975} = 2.26$ is a quantile in t-distribution with 9 degrees of freedom

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

As stated in Method 3.9, the 95% confidence interval is given by

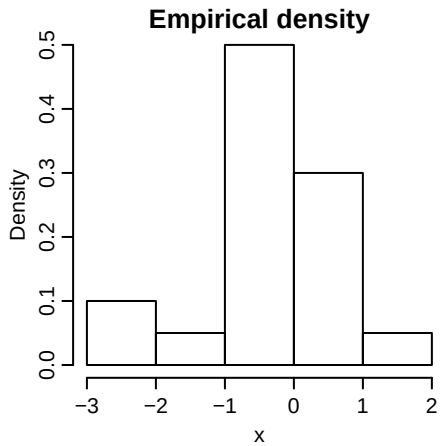
$$\bar{x} \pm t_{1-\alpha/2}\frac{s}{\sqrt{n}}$$

with the numbers given above the correct is Answer 5.

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

Question VI.4 (22)

Another sample is taken and its empirical density is:



What is the size of the sample, i.e. how many observations n are in the sample?

1 ☐ 20

2 ☐ 30

3 ☐ 100

4 ☐ 300

5* ☐ This question cannot be answered with the given information

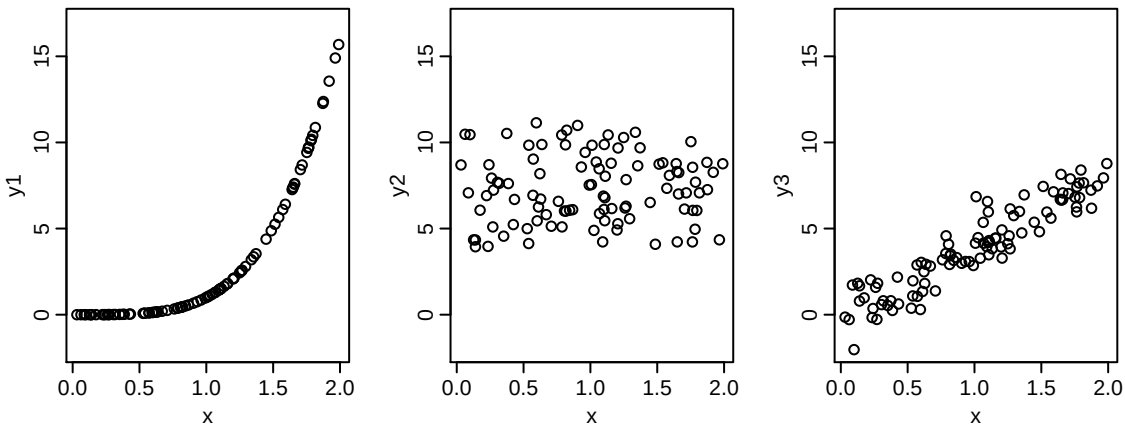
---------- FACIT-BEGIN ----------

Since we only have the relative frequencies (the empirical density), we cannot know how many observations are in the sample.

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

Question VI.5 (23)

The following three plots are of coherent values of x and y for samples from three different populations:



The following statements are about the correlations of the three populations from which the samples were taken. Which of the statements is not very unlikely?

1 ☐ $\rho_{XY_1} = 0$, $\rho_{XY_2} = 0$ and $\rho_{XY_3} = 0.33$

2 ☐ $\rho_{XY_1} = 0$, $\rho_{XY_2} = 0$ and $\rho_{XY_3} = -0.89$

3 ☐ $\rho_{XY_1} = 0$, $\rho_{XY_2} = 0.61$ and $\rho_{XY_3} = 0.91$

4* ☐ $\rho_{XY_1} = 0.87$, $\rho_{XY_2} = 0$ and $\rho_{XY_3} = 0.92$

5 ☐ $\rho_{XY_1} = 0.22$, $\rho_{XY_2} = 0$ and $\rho_{XY_3} = -0.34$

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

From the plots we can see that $\rho_{XY_1} > 0$, $\rho_{XY_2} \approx 0$, $\rho_{XY_2} > 0$, since there is a positive correlation between x and y1 and x and y3, but no visible correlation between x and y2, hence the only plausible is Answer 4.

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

Continues on page 32

Exercise VII

In a finite population of N units with mean $\mathrm{E}[Y] = \mu$ and variance $\mathrm{V}[Y] = \sigma^2$ we are considering a sample with n units Y_i, $i = 1, \ldots, n$. If the sample is taken randomly and without replacement, then the sample mean is $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and the variance is $\mathrm{V}(\bar{Y}) = \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n}$. The interest is now on the sum of the sample $\tau = \sum_{i=1}^{N} Y_i$, which can be estimated by $\hat{\tau} = \frac{N}{n}\sum_{i=1}^{n} Y_i$.

Question VII.1 (24)

What is the variance of the estimator $\hat{\tau}$ i.e. $\mathrm{V}(\hat{\tau})$?

1 ☐ $\mathrm{V}(\hat{\tau}) = \frac{N^2}{n}\sigma^2$

2* ☐ $\mathrm{V}(\hat{\tau}) = N(N-n)\frac{\sigma^2}{n}$

3 ☐ $\mathrm{V}(\hat{\tau}) = \frac{N^2}{n^3}\sigma^2$

4 ☐ $\mathrm{V}(\hat{\tau}) = N^2(1-n)\sigma^2$

5 ☐ $\mathrm{V}(\hat{\tau}) = \frac{N}{n}\sigma^2$

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

Since

$$\hat{\tau} = \frac{N}{n}\sum_{i=1}^{n} y_i = N\bar{y},$$

where $\bar{y}$ symbolizes the sample mean of y we can bring out the N and square it (as described in Theorem 2.54)

$$\begin{aligned}\mathrm{V}[\hat{\tau}] &= \mathrm{V}[N\bar{y}] \\ &= N^2\,\mathrm{V}[\bar{y}] \\ &= N^2\left(\frac{N-n}{N}\right)\frac{\sigma^2}{n} \\ &= N(N-n)\frac{\sigma^2}{n}.\end{aligned}$$

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

Continues on page 33

Exercise VIII

Up until the 1970s in Finland, it was only allowed to sell and serve alcoholic beverages in towns and not in rural areas. When it was wanted to ease the restrictions on alcohol sale in rural areas it raised concerns if this would lead to an increased rate of road accidents. Ahead of easing the restrictions a project was carried out in which: four rural municipalities were granted extraordinary permission to sell alcohol in shops, and four other municipalities were granted permission to, besides selling alcohol in shops, serve alcohol in restaurants and others serving places. Finally, four other rural municipalities without extraordinary permits acted as control. Data on the number of traffic accidents from the 12 selected municipalities over the year the project ran is presented in the following table:

Name	Control	Sale	SaleAndServing
	177	226	226
	225	196	229
	167	198	215
	176	206	188
Sum	745	826	858

and the chosen analyses is an ANOVA. The result is:

```
## Analysis of Variance Table
##
## Response: Accidents
##           Df Sum Sq Mean Sq F value Pr(>F)
## Treatment  A 1696.2  848.08     C     D
## Residuals  B 3670.7  407.86
```

Where `Treatment` is a factor dividing the municipalities into the three groups and `Accidents` is the number of accidents.

Question VIII.1 (25)

To investigate whether the permission to sell alcohol has an effect on the rate of traffic accidents, the average number of traffic accidents in the 3 groups are compared. Assuming that the variance in the number of traffic accidents is constant between the groups, what is then the result of the test for a difference in the mean number of traffic accidents between the 3 groups on significance level $\alpha = 0.05$?

1 ☐ The test statistic $F_{\text{obs}} = 1.232$ which under H_0 follows an F-distribution with 3 and 8 degrees of freedom, gives a p-value of 0.360 and the study therefore gives no reason to believe that an easing of alcohol restrictions will increase number of traffic accidents

2 ☐ The test statistic $F_{\text{obs}} = 2.079$ which under H_0 follows an F-distribution with 2 and 9 degrees of freedom, gives a p-value of 0.181 and the study therefore showes that easing of alcohol restrictions will certainly lead to an increase in the number of traffic accidents

3* ☐ The test statistic $F_{\text{obs}} = 2.079$ which under H_0 follows an F-distribution with 2 and 9 degrees of freedom, gives a p-value of 0.181 and the study therefore gives no reason to believe that an easing of alcohol restrictions will increase number of traffic accidents

4 ☐ The test statistic $F_{\text{obs}} = 4.324$ which under H_0 follows an F-distribution with 2 and 9 degrees of freedom, gives a p-value of 0.0434 and the study therefore shows that easing of alcohol restrictions will lead to a change of the number of traffic accidents

5 ☐ The test statistic $F_{\text{obs}} = 4.324$ which under H_0 follows an F-distribution with 3 and 8 degrees of freedom, gives a p-value of 0.0434 and the study therefore shows that easing of alcohol restrictions will lead to a change of the number of traffic accidents

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

This is a one-way ANOVA. First note that the degrees of freedom are $A = 3 - 1 = 2$, and $B = 4 \cdot 3 - 3 = 9$, these are the degrees of fredom needed to calculate the p-value. Now lets calculate the numbers `C` and `D` as described in Method 8.6

```
(F <- 848.08 / 407.86)

## [1] 2.079341

(p.value <- 1 - pf(F,df1=2,df2=9))

## [1] 0.1809821
```

Hence the only possible correct answers are 2 and 3, but in Answer 2 a wrong conclusion is drawn (with the given p-value=0.181 > 0.05) while Answer 3 correctly states that the there is not evidence against the null-hypothesis (that the number of accident will not increase).

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

Question VIII.2 (26)

What is the estimate of the standard deviation of the errors?

1 ☐ $\hat{\sigma} = 1696.2/(12-1) = 154$

2 ☐ $\hat{\sigma} = \sqrt{1696.2/(3-1)} = 29.1$

3 ☐ $\hat{\sigma} = \sqrt{1696.2/(12-1)} = 11.0$

4* ☐ $\hat{\sigma} = \sqrt{3670.7/(12-3)} = 20.2$

5 ☐ $\hat{\sigma} = 5367.1/(12-3)^2 = 66.3$

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

See the end of Chapter 8.2.2. Since $MSE = s^2$, the number can be calculate directly from the ANOVA table as $\hat{\sigma} = \sqrt{407.86} = 20.2$.

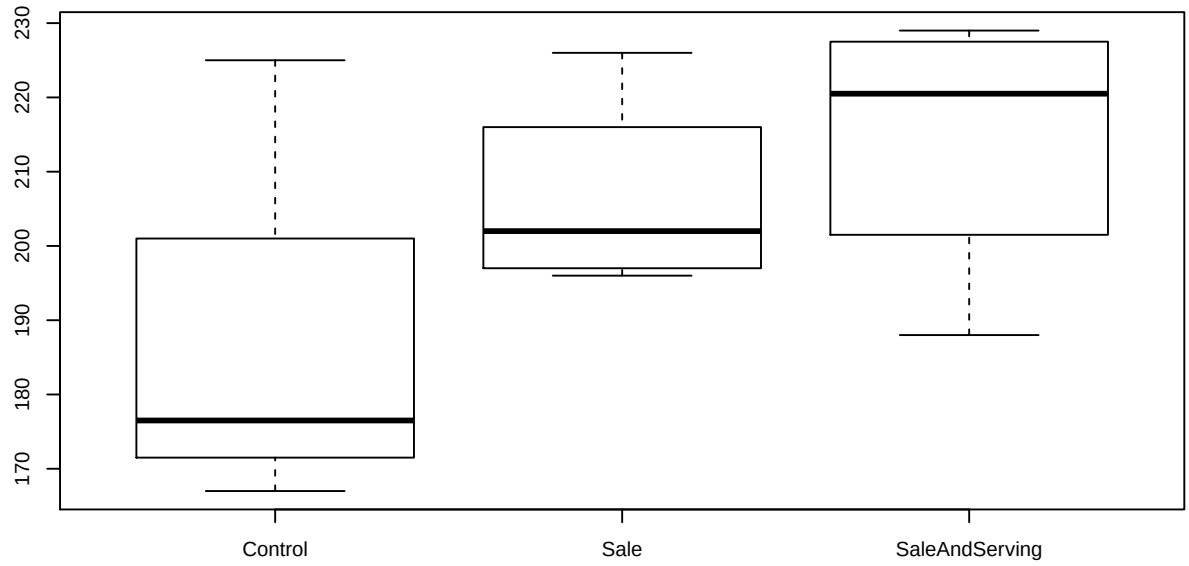
Or more in detail R:

```
(SSE <- (4-1) * (687.58 + 187.666667 + 348.3333)) # Eq 8-6

## [1] 3670.74

(MSE <- SSE/(n-k))

## [1] 407.86

## The answer
sqrt(MSE)

## [1] 20.19554

sqrt(3670.7/(12-3))

## [1] 20.19543
```

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

Question VIII.3 (27)

The assumption of homogeneous variance is validated with the following box plots:



Which of the following statements is the most correct conclusion based on this plot and the informations given (not all the statements are necessarily meaningful)?

1 ☐ Taking the high number of observation into account there is no evidence that the assumption of homogeneous variance is not fulfilled

2 ☐ Taking the high number of observation into account there is evidence that the assumption of homogeneous variance is not fulfilled

3* ☐ Taking the low number of observation into account there is no evidence that the assumption of homogeneous variance is not fulfilled

4 ☐ Taking the low number of observation into account there is evidence that the assumption of homogeneous variance is not fulfilled

5 ☐ Based on the information provided there cannot be drawn any conclusions about the assumption of homogeneous variance

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

The number of observations for each box-plot is 4, which is a small number (hence we can exclude answer 1 and 2). With a low number of observations we will have to accept some differences between the box plots and hence there is no evidence against the hypothesis of homogeneous variance.

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

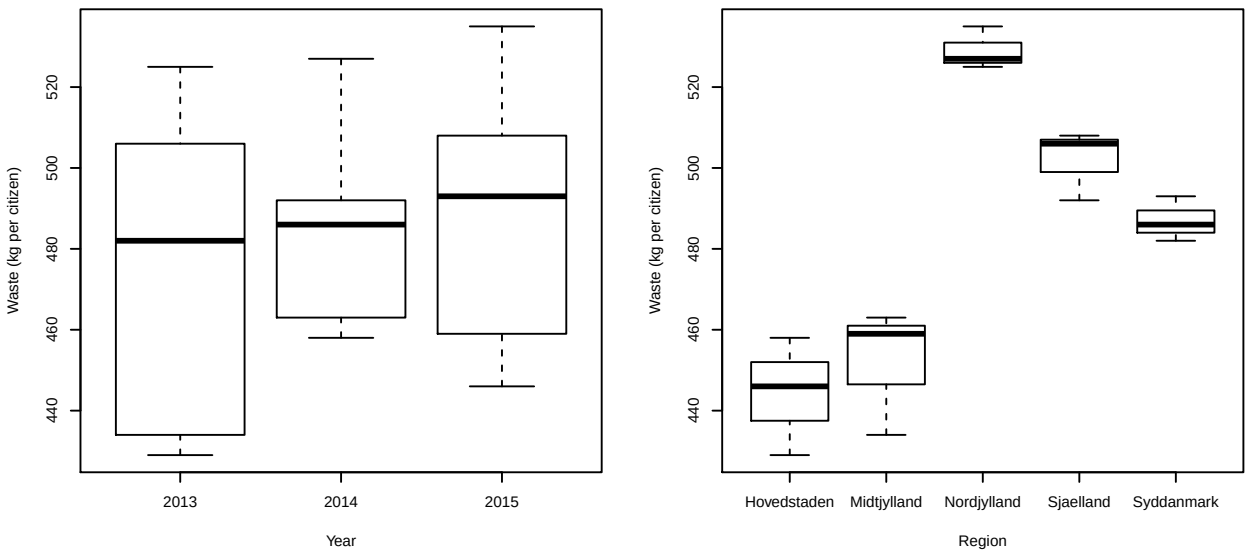
Continues on page 37

Exercise IX

The Environmental and Food Agency collects data on waste in Denmark every year and publishes a report with data and analyses. The report is named "Affaldsstatistik 2015"([1]) and in it one can find the amount of waste (kg) per citizen for the years 2013 to 2015 grouped on regions:

	Hovedstaden	Midtjylland	Nordjylland	Sjaelland	Syddanmark
2013	429	434	525	506	482
2014	458	463	527	492	486
2015	446	459	535	508	493

The following box plot shows waste per citizen grouped on year and on region:



A 2-way ANOVA is carried out and the result is:

```
## Analysis of Variance Table
##
## Response: Waste
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Year       2   463.3   231.7  2.5551    0.1386
## Region     4 14847.1  3711.8 40.9386 2.266e-05 ***
## Residuals  8   725.3    90.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[1] http://www2.mst.dk/Udgiv/publikationer/2017/05/978-87-93614-01-7.pdf

Question IX.1 (28)

Which of the following statements is correct when using a significance level of $\alpha = 5\%$?

1 ☐ From the box plot it can be seen that there is no significant difference in waste between the years, which is also the conclusion from the ANOVA test

2 ☐ Anwers taken out of the exam.

3* ☐ From the box plot it is not possible to conclude if there is a significant difference in waste between the years, but from the ANOVA test no significant difference in waste between the years can be concluded

4 ☐ From the box plot it is not possible no conclude if there is a significant difference in waste between the years, but from the ANOVA test a significant difference in waste between the years can be concluded

5 ☐ None of the statements above are correct

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

Since we have two effects, we cannot make conclusions about the effects of years based on the box plot (since the effect of region is not being accounted for in the boxplot). Hence answer 1 and 2 are both wrong. The ANOVA test shows no significant difference between years, since the p-value$= 0.1386 > 0.05$.

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

Question IX.2 (29)

Further, in the report it is listed how large a proportion of the waste is sorted in the five regions and the proportion of waste that is sorted is calculated for each year and each region. A 2-way ANOVA has been carried out on this data with the following result:

```
## Analysis of Variance Table
##
## Response: Proportion
##           Df    Sum Sq   Mean Sq F value   Pr(>F)
## Year       2 0.0109878 0.0054939  13.054 0.003026 **
## Region     4 0.0173773 0.0043443  10.323 0.003019 **
## Residuals  8 0.0033668 0.0004208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Which one of the following conclusions is correct using a significance level of 5% (both argument and conclusion must be correct)?

1 ☐ Since the p-value > 0.05 for the relevant test, a significant change in the sorted proportion over the years is not detected

2* ☐ Since $P(F > 13.054) < 0.05$ where F follows the relevant F-distribution, a significant change in the sorted proportion over the years is detected

3 ☐ Since $P(T > 0.003) > 0.05$ where T follows the relevant t-distribution, a significant change in the sorted proportion over the years is not detected

4 ☐ Since $P(T < 10.323) < 0.05$ where T follows the relevant t-distribution, a significant change in the sorted proportion over the years is detected

5 ☐ Since $1 - P(T > 10.323) > 0.05$ where T follows the relevant t-distribution, a significant change in the sorted proportion over the years is not detected

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

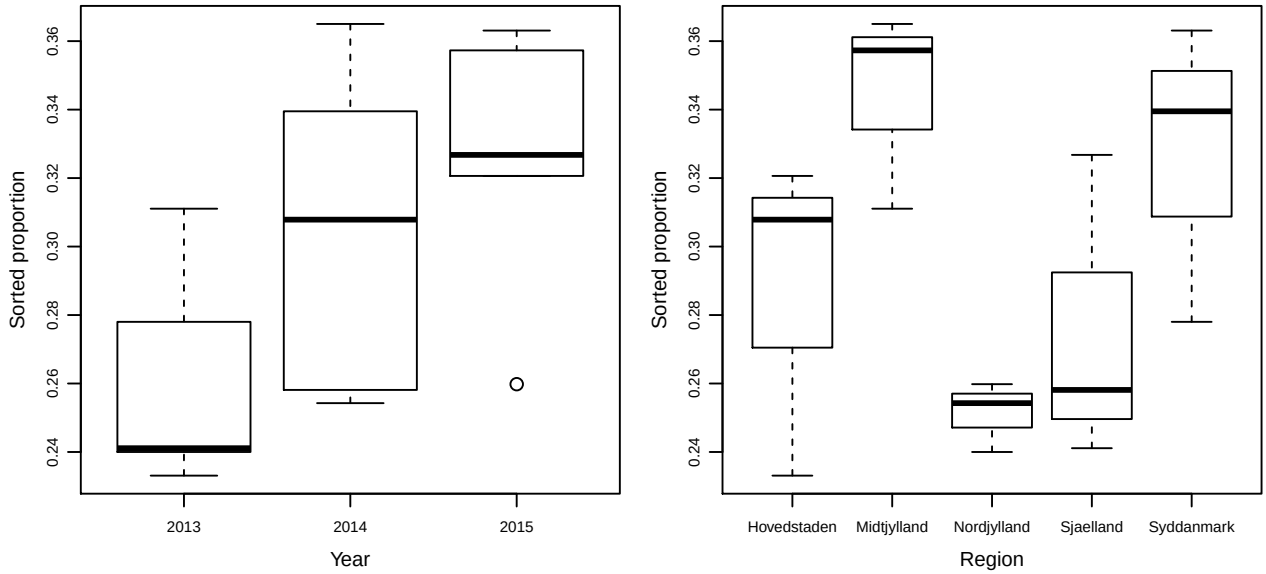
Both p-values are less than 0.05, hence Answer 1 is not correct. Since $P(F > 13.054) = 0.00303 < 0.05$ there is a significant effect of years, hence Answer 2 is correct.

Answer 3, 4 and 5 all uses a t-distribution instead of an f-distribution and hence they are all wrong.

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

Question IX.3 (30)

The box plots showing the proportion of sorted waste by year and by region are:



It is seen that in 2015 there is an observation which is low compared to the others, and it is identified as an outlier according to the modified box plot.

Which of the following statements is not correct (Tip: Remember that there is only one observation for each year in each region)?

1 ☐ The lowest observation in 2013 is from Hovedstaden

2 ☐ The lowest observation in 2015 (i.e. the outlier) is from Nordjylland

3* ☐ Each year Sjaelland has had a higher observation than Hovedstaden

4 ☐ Nordjylland has the lowest median

5 ☐ The 75% quantile for 2014 is higher than the 25% quantile for 2015

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

Lets go through the answers:

1 In the right box plot it is seen that Hovedstaden has the lowest value of all the regions, and 2013 have the lowest value of all years, hence Hovedstaden must have the value in 2013. Thus TRUE

2 All regions have exactly one observation in 2015. Nordjylland has none higher than the outlier, hence it must belong to Nordjylland. Thus TRUE

3 The median for the regions mark exactly the middle observation, since there are only 3 values for each region. Since Hovedstaden has had 2 values above the 2'nd highest value

for Sjaelland, then Hovedstaden must have had a higher value one of the years. Thus NOT TRUE

4 The black bar in the box marks the median, which thus seen in the right-hand box plot to be lowest for Nordjylland. Thus TRUE

5 The 75% quantile is marked by the upper side of the box, thus on the left-hand box plot it is seen to be higher for 2014 than the 25% quantile for 2015 (marked by the lower side of the box). Thus TRUE

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

Continues on page 43

THE EXAM IS FINISHED. Enjoy the late summer!

Technical University of Denmark *Page 1 of 45 pages.*

Written examination: 17. December 2017

Course name and number: **Introduction to Statistics (02323)**

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 18 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	III.1	III.2	IV.1	IV.2	V.1	V.2	V.3
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer	5	2	4	5	3	3	4	5	3	3

Exercise	VI.1	VI.2	VII.1	VIII.1	VIII.2	IX.1	IX.2	IX.3	X.1	XI.1
Question	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Answer	3	2	5	2	1	2	2	2	5	4

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

The questionnaire contains 45 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

We consider pairwise measurements of 2 stochastic variables, X og Y. Both variables can be assumed normally distributed.

x	38	35	47	38	42	41	48	35
y	25	21	26	23	28	27	29	18

Data can be loaded into R using the following command:

```
x <- c(38, 35, 47, 38, 42, 41, 48, 35)
y <- c(25, 21, 26, 23, 28, 27, 29, 18)
```

Question I.1 (1)

Provide an estimate for the correlation coefficient, ρ, between X and Y:

1 ☐ $\hat{\rho} = 0.12$

2 ☐ $\hat{\rho} = 0.22$

3 ☐ $\hat{\rho} = 0.64$

4 ☐ $\hat{\rho} = 0.73$

5* ☐ $\hat{\rho} = 0.82$

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

See Definition 1.19. The easiest way to do this is to use R. We copy into R to read in the data two vectors

```
x <- c(38, 35, 47, 38, 42, 41, 48, 35)
y <- c(25, 21, 26, 23, 28, 27, 29, 18)
```

and the we calculate the estimate of the correlation as the sample correlation

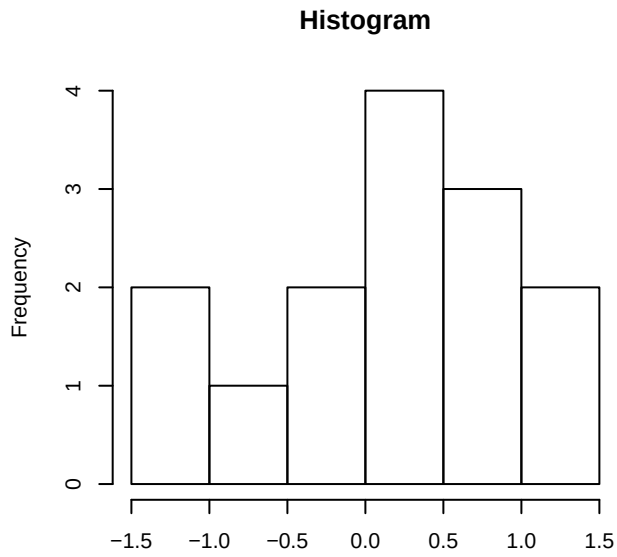
```r
cor(x,y)

## [1] 0.8237548
```

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

Continues on page 4

Exercise II

A sample has been taken from a population. The following histogram has been generated:



Question II.1 (2)

What is the sample size n?

1 ☐ $n = 10$

2* ☐ $n = 14$

3 ☐ $n = 20$

4 ☐ $n = 28$

5 ☐ $n = 40$

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

In the histogram the height of each bar is the number of observations in the sample which is in the interval spanned by the bar on the x-axis. So we can simply add heights of the bars to get the sample size, so

$$n = 2 + 1 + 2 + 4 + 3 + 2 = 14.$$

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

Question II.2 (3)

The sample variance has been calculated to $s^2 = 0.79$. What is the sample standard deviation?

1 ☐ 0.31

2 ☐ 0.40

3 ☐ 0.62

4* ☐ 0.89

5 ☐ 1.58

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

The sample standard deviation, s is found by taking the square root of the sample variance (Definition 1.11)

$$s = \sqrt{s^2} = \sqrt{0.79} = 0.89.$$

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

Continues on page 6

Exercise III

In a study 605 test persons, all with a record of previous heart disease, were randomized to one of two possible diets (A or B), in order to study the effect of diet on health. After an observation period of 4 years the test persons were classified according to health status: (I) dead, (II) cancer, (III) other disease, (IV) well.

Health status

	I	II	III	IV	Total
Diet A	15	24	25	239	303
Diet B	7	14	8	273	302
Total	22	38	33	512	605

The null hypothesis in the study was that there is no association between diet and health.

Question III.1 (4)

State the distribution of the usual test statistics, when assuming that the null hypothesis is true:

1 ☐ The usual test statistics follows a χ^2-distribution with 8 degrees of freedom

2 ☐ The usual test statistics follows a F-distribution with (1, 603) degrees of freedom

3 ☐ The usual test statistics follows a t-distribution with 4 degrees of freedom

4 ☐ The usual test statistics follows a t-distribution with 302 degrees of freedom

5* ☐ The usual test statistics follows a χ^2-distribution with 3 degrees of freedom

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

The setup of the data is a multi-sample proportion setup (chapter 7.4). We must test the hypothesis, that the proportions in each group is equal

$$H_0 : P_1 = p_2 = p_3 = p_4.$$

and under this hypothesis the test statistic follows a χ^2-distribution with $c - 1$ degrees of freedom, and there are 4 groups, so 3 degrees of freedom (Method 7.20).

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

Question III.2 (5)

We now only consider the proportion of test persons who are healthy at the end of the 4 year period. We want to estimate at 95% confidence interval for the difference in proportions of test

persons who are healthy for each of the 2 diets. Which of the suggestions below is the correct code in R to achieve this?

1 ☐ `prop.test(x=c(512), n=c(605), correct=FALSE)`

2 ☐ `prop.test(x=c(303,302), n=c(512,512), correct=FALSE)`

3* ☐ `prop.test(x=c(239,273), n=c(303,302), correct=FALSE)`

4 ☐ `prop.test(x=c(239,273), n=c(605,605), correct=FALSE)`

5 ☐ `prop.test(x=c(239,273), n=c(512,512), correct=FALSE)`

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

Here we are working with proportions in two populations as described in Chapter 7.3. We need the observed proportion which are well for each diet. So on Diet A 239 out of 303 are well and for Diet B 273 out of 302 are well, and these numbers are passed to `prop.test`, which then prints out the estimated confidence interval (same as Example 7.19).

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

Continues on page 8

Exercise IV

In the production of a consumer product 3 subprocesses are involved, denoted A, B and C. The time (in hours) it takes to complete each subprocess is represented with a random variable, which we denote X_A, X_B and X_C, respectively. It can be assumed, that X_A, X_B and X_C are all independent and normally distributed given by $X_A \sim N(12, 2^2)$, $X_B \sim N(25, 3^2)$ and $X_C \sim N(42, 4^2)$.

The total production time, Y, is now defined by

$$Y = X_A + X_B + X_C.$$

Question IV.1 (6)

State the probability that the total production time, Y, exceeds 85 hours:

1 ☐ 0.0081

2 ☐ 0.1080

3* ☐ 0.1326

4 ☐ 0.4180

5 ☐ 0.6301

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

We need to find the mean and variance of Y, which we know is normal distributed, since a linear function of normal distributed random variables is also normal distributed (Theorem 2.56).

We use the identities in Theorem 2.56 to get

$$\mu_Y = \mathrm{E}(Y) = \mathrm{E}(X_A + X_B + X_C) = \mathrm{E}(X_A) + \mathrm{E}(X_B) + \mathrm{E}(X_C) = 12 + 25 + 42 = 79,$$

and

$$\sigma_Y^2 = \mathrm{V}(Y) = \mathrm{V}(X_A + X_B + X_C) = \mathrm{V}(X_A) + \mathrm{V}(X_B) + \mathrm{V}(X_C) = 4 + 9 + 16 = 29.$$

Alternatively we could also have simulated the variance in R.

```
k <- 1000000
X_a <- rnorm(k, 12, 2)
X_b <- rnorm(k, 25, 3)
X_c <- rnorm(k, 42, 4)
Y <-X_a + X_b + X_c
var(Y)
```

```
## [1] 28.99541
```

This we use to look up the probability $P(Y > 85) = 1 - P(Y \leq 85)$ in R by:

```
1 - pnorm(q=85, mean=79, sd=sqrt(29))

## [1] 0.1326027
```

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

Question IV.2 (7)

An engineer is now able to perform some optimization of the process, so that the improved process time Y^*, becomes

$$Y^* = 0.9 \cdot X_A + 0.8 \cdot X_B + X_C,$$

where X_A, X_B and X_C are defined as in the previous question.

State the variance of Y^*:

1 ☐ $\mathrm{V}(Y^*) = (0.9 + 0.8 + 1) \cdot (2^2 + 3^2 + 4^2)$

2 ☐ $\mathrm{V}(Y^*) = (0.9^2 + 0.8^2 + 1^2) \cdot (2^2 + 3^2 + 4^2)$

3 ☐ $\mathrm{V}(Y^*) = 0.9 \cdot 2^2 + 0.8 \cdot 3^2 + 1 \cdot 4^2$

4* ☐ $\mathrm{V}(Y^*) = 0.9^2 \cdot 2^2 + 0.8^2 \cdot 3^2 + 1^2 \cdot 4^2$

5 ☐ $\mathrm{V}(Y^*) = 2^2 + 3^2 + 4^2$

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

Again the identities in Theorem 2.56 to get

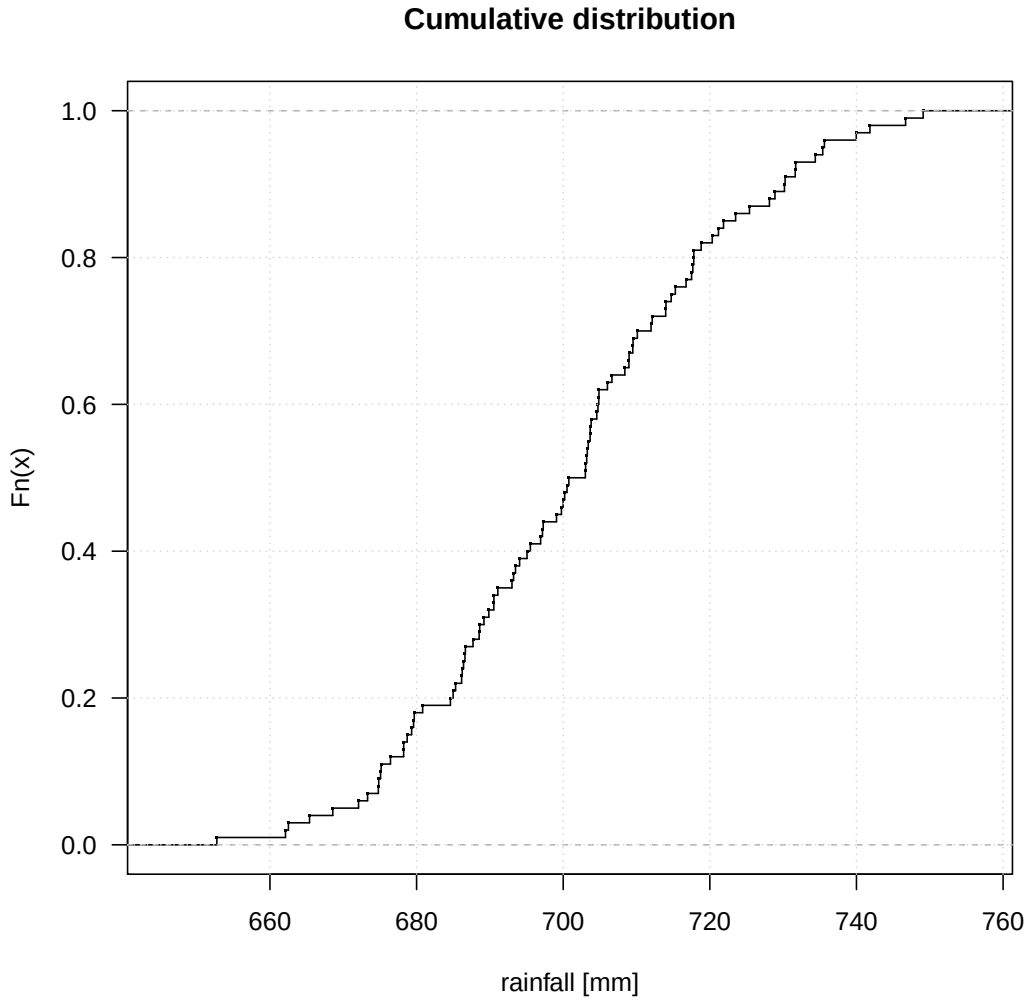
$$\begin{aligned}\sigma^2_{Y^*} = \mathrm{V}(Y^*) &= \mathrm{V}(0.9 \cdot X_A + 0.8 \cdot X_B + X_C,) \\ &= 0.9^2\,\mathrm{V}(X_A) + 0.8^2\,\mathrm{V}(X_B) + \mathrm{V}(X_C) \\ &= 0.9^2 \cdot 2^2 + 0.8^2 \cdot 3^2 + 1^2 \cdot 4^2\end{aligned}$$

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

Continues on page 10

Exercise V

The yearly rainfall has been registered within a region for the last 100 years. It can be assumed that the rainfall is independent from year to year. The cumulative distribution for the yearly rainfall is shown in the figure below:



The following summary of the data has been conducted by the use of R, where the yearly rainfall measurements are stored in the variable `rainfall`:

```
> var(rainfall)
[1] 412.7042
> summary(rainfall)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  652.8   686.6   701.9   701.3   714.9   749.1
```

Continues on page 11

Question V.1 (8)

Which of the following statements is not correct?

1 ☐ The estimate of the standard deviation of the mean $\hat{\sigma}_{\bar{X}}$, becomes $\frac{\sqrt{412.7042}}{10}$ mm

2 ☐ The 50% quantile for the 100 observations is 701.9 mm

3 ☐ The standard deviation of the sample, s, for the 100 measurements is $\sqrt{412.7042}$ mm

4 ☐ 50% of the 100 observations are between 686.6 and 714.9 mm

5* ☐ The estimated coefficient of variation for the 100 observations becomes $\frac{412.7042}{701.9}$

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

Lets go through them one by one:

1. TRUE statement. The formula for the estimate is $\frac{s}{\sqrt{n}}$ (also called the standard error of the mean). See Definition 3.7
2. TRUE statement. Seen from the `summary()` call
3. TRUE statement. Standard deviation is the square root of the variance
4. TRUE statement. 686.6 is the first quartile (25% quantile) and 714.9 is the third quartile (75% quantile), and certainly 50% of the observations lies between the 25% and 75% quantile
5. FALSE statement. The estimated coefficient of variation is $\hat{V} = \frac{s}{\bar{x}} = \frac{\sqrt{412.7042}}{701.3}$. See Definition 1.12.

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

Question V.2 (9)

Provide a 95% confidence interval for the variance of the rainfall based on the 100 observations, still assumed to be normally distributed:

1 ☐ $[\frac{20.31512^2 \cdot 134.6416}{99}; \frac{20.31512^2 \cdot 69.22989}{99}]$

2 ☐ $[\frac{20.31512^2 \cdot 99}{134.6416}; \frac{20.31512^2 \cdot 99}{69.22989}]$

3* ☐ $[\frac{412.7042 \cdot 99}{128.422}; \frac{412.7042 \cdot 99}{73.36108}]$

4 ☐ $[\frac{412.7042 \cdot 99}{123.2252}; \frac{412.7042 \cdot 99}{77.04633}]$

5 ☐ $[\frac{20.31512 \cdot 99}{123.2252}; \frac{20.31512 \cdot 99}{77.04633}]$

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

We find the formula for a $1 - \alpha$ confidence interval for the variance of a normal distributed population in Method 3.19 and insert the values

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

The chi-square quantiles are found in R as

```
qchisq(c(0.025, 0.975), 99)

## [1]  73.36108 128.42199
```

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

Continues on page 13

Question V.3 (10)

We continue with the exercise from the previous page. The following code in R has now been run:

```
k = 10^5
Q5 = function(x){ quantile(x, 0.95) }
samples = replicate(k, sample(rainfall, replace = TRUE))
simvalues = apply(samples, 2, Q5)
interval = quantile(simvalues, c(0.025,0.975))
```

which gives the result:

```
> interval
    2.5%    97.5%
728.9515 742.0814
```

What has been calculated in the vector `interval`?

1 ☐ A 95% confidence interval for the mean of the yearly rainfall (parametric bootstrap)

2 ☐ A 95% confidence interval for the 5% quantile of the yearly rainfall (parametric bootstrap)

3* ☐ A 95% confidence interval for the 95% quantile of the yearly rainfall (non-parametric bootstrap)

4 ☐ A 95% confidence interval for the 2.5% and 97.5% quantile of the yearly rainfall (non-parametric bootstrap)

5 ☐ A 95% confidence interval for the 2.5% and 97.5% quantile of the yearly rainfall (parametric bootstrap)

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

We look at the R code and see that it is a bootstrapping is carried out by simulating the sample 100000 times, and not assuming any distribution (since the `sample` function is used), therefore it is non-parametric.

The statistic calculated for each simulated sample is the 95% quantile and since the quantiles taken for these values are the 2.5% and the 97.5%, then the results is a 95% confidence interval for the 95% quantile.

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

Continues on page 14

Exercise VI

We consider an experiment that can result in one of two possible outcomes, here denoted A or B. The probability of outcome A is denoted $P(A)$. By defintion we get the probability of outcome B as $P(B) = 1 - P(A)$.

Question VI.1 (11)

Assume that we observe a random variable, X, which counts the number of times that we observe the outcome A out of $n = 300$ independent trials of the experiment. If we assume that $P(A) = 0.40$ in a single trial, what is then the expected number $\mathrm{E}(X)$ and variance $\mathrm{V}(X)$?

1 ☐ $\mathrm{E}(X) = 300 \cdot 0.4 \cdot (1 - 0.4)$ and $\mathrm{V}(X) = 300^2 \cdot 0.4$

2 ☐ $\mathrm{E}(X) = 300 \cdot 0.4$ and $\mathrm{V}(X) = 300^2 \cdot 0.4 \cdot 0.6$

3* ☐ $\mathrm{E}(X) = 300 \cdot 0.4$ and $\mathrm{V}(X) = 300 \cdot 0.4 \cdot 0.6$

4 ☐ $\mathrm{E}(X) = 300 \cdot 0.4 \cdot 0.6$ and $\mathrm{V}(X) = 300^2 \cdot 0.4^2 \cdot 0.6^2$

5 ☐ $\mathrm{E}(X) = 300 \cdot 0.4 \cdot 0.6$ and $\mathrm{V}(X) = 300 \cdot 0.4^2 \cdot 0.6^2$

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

X follows a Binomial distribution with $p = 0.4$ and we have a formula for the mean and variance defined in Theorem 2.21, which we use to get

$$\mu = \mathrm{E}(X) = np = 300 \cdot 0.4,$$
$$\sigma^2 = \mathrm{V}(X) = np(1-p) = 300 \cdot 0.4 \cdot 0.6.$$

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

Question VI.2 (12)

Regardless of your answer to the previous question we now want to estimate the probability $P(A)$ based on the $n = 300$ trials. From the $n = 300$ trials we count that in 120 of these the outcome was A and in the remaining 180 trials the outcome was B. Provide a 95% confidence interval for the probability $P(A)$:

1 ☐ [0.33, 0.48]

2* ☐ [0.35, 0.46]

3 ☐ [0.35, 0.42]

4 ☐ [0.31, 0.53]

5 ☐ [0.29, 0.54]

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

Using the inbuilt function in R the results is

```
prop.test(120, 300, correct=FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  120 out of 300, null probability 0.5
## X-squared = 12, df = 1, p-value = 0.000532
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3461652 0.4563634
## sample estimates:
##   p
## 0.4
```

whereas using the formula in method 7.3 gives a slightly different result is obtained

```
n <- 300
x <- 120
phat <- x/n
phat + c(-1,1) * qnorm(p=0.975) * sqrt(phat*(1-phat)/n)

## [1] 0.3445638 0.4554362
```

This is due to a numerical rounding by R and can occur sometimes. The answer is in any case closest to the answer marked correct [0.35, 0.46].

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

Continues on page 16

Exercise VII

An engineer is examining the quality in a batch of raw materials. The quality demand is that the purity of the raw material is at least 90%. The engineer takes a sample of 10 independent measurements from the batch and saves the measured values (in %) of the purity in a vector `x`.

He then runs the following code in R

```
> x <- c(90.6, 90.3, 88.9, 87.5, 87.6, 88.1, 87.5, 88, 88, 89.6)
> n <- length(x)
> tobs <- (mean(x) - 90) / (sd(x) / sqrt(n))
> pt(tobs, df=n-1)
```

Which yields the following output

```
[1] 0.002279236
```

Question VII.1 (13)

Based on the calculations listed above, and assuming that the measurements of the purity are normally distributed and applying a significance level of $\alpha = 0.05$, what can the engineer conclude?

1 ☐ The engineer can conclude that the purity of the raw material is at least 88.6%

2 ☐ The engineer can conclude that the mean purity of the raw material is at most 88.6%

3 ☐ The engineer has with probability 99.7% shown that the mean purity of the raw material is 90%

4 ☐ The engineer can assume that the mean purity of the raw material is 90%

5* ☐ The engineer can reject that the mean purity of the raw material is 90%

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

We can see from the way that `tobs` is calculated that the null hypothesis is that the $\mu = 90$ (See Method 3.23) Since the p-value is 0.0046 and thus much lower than $\alpha = 0.05$. This leads to the conclusion that the null hypothesis, that the mean purity is 90%, must be rejected.

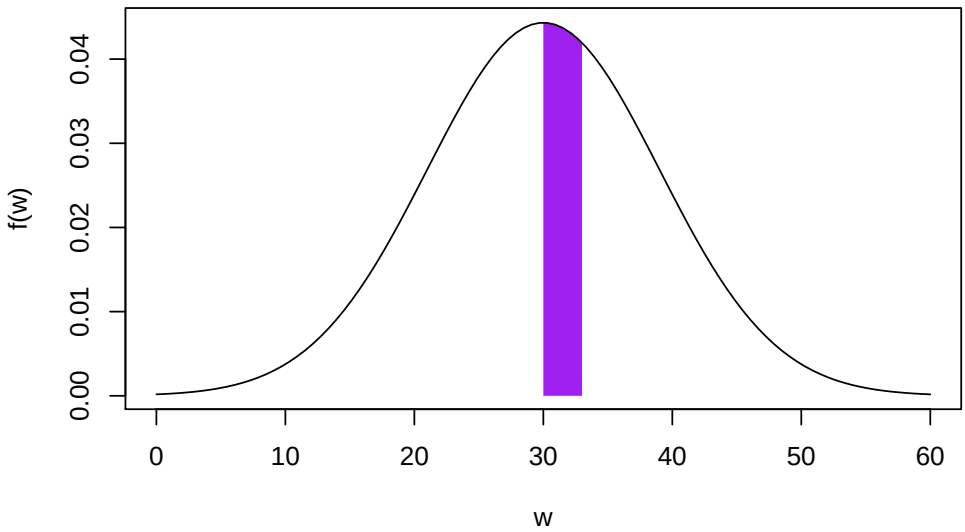
---------- FACIT-END ----------

Continues on page 17

Exercise VIII

We consider a random variable W with density function $f(w) = \frac{1}{9\sqrt{2\pi}} e^{-\frac{(w-30)^2}{162}}$.

The density function is shown in the figure below, where the probability $P(30 < W < 33)$ is shown as the shaded area.



Question VIII.1 (14)

Calculate the probability $P(30 < W < 33)$:

1 ☐ 0.09

2* ☐ 0.13

3 ☐ 0.24

4 ☐ 0.34

5 ☐ 0.84

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

The answer is obtained from recognizing the the formula for the probability density function (pdf) for the normal distribution in definition 2.37

$$f(w) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(w-\mu)^2}{2\cdot\sigma^2}}$$

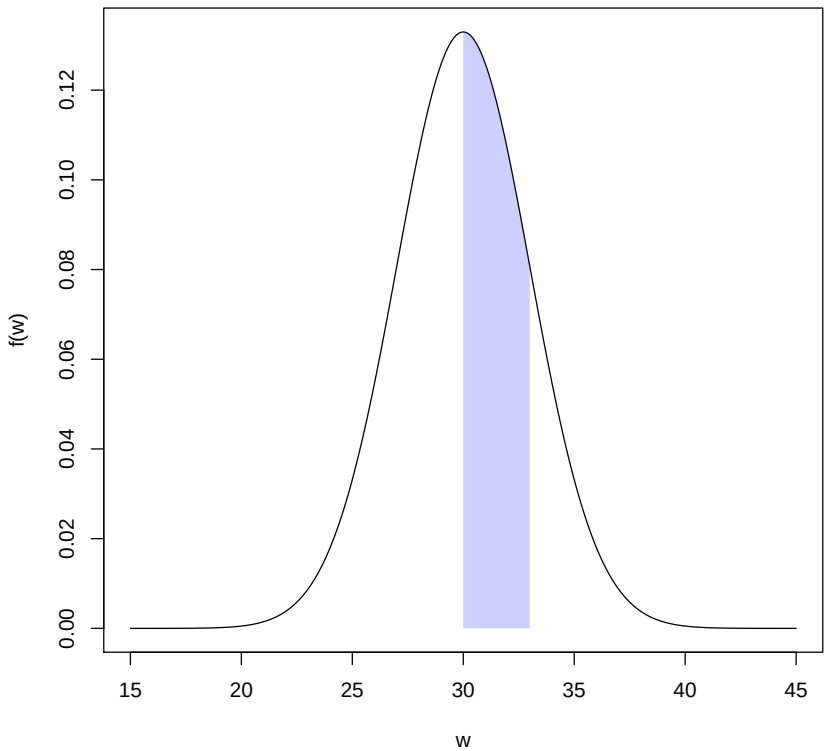
and thus to find the mean $\mu = 30$ and variance $\sigma = 9$. These are then used to obtain

$$P(30 < W < 33) = P(X < 33) - P(X < 30)$$

in R

```
pnorm(33, mean=30, sd=9) - pnorm(30, mean=30, sd=9)

## [1] 0.1305587
```

SINCE in the original exam the plot was which indeed was wrong, it was of the normal distri-



bution with mean $\mu = 30$ and variance $\sigma = 3$

```
pnorm(33, mean=30, sd=3) - pnorm(30, mean=30, sd=3)

## [1] 0.3413447
```

then the Answer 4 is also counted as correct!

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

Continues on page 19

Question VIII.2 (15)

We consider a situation where we take 3 different samples denoted A, B, and C. All three samples are from the population characterized by the density $f(w) = \frac{1}{9\sqrt{2\pi}} e^{-\frac{(w-30)^2}{162}}$ as in the previous question.

Sample A is of size $n_A = 10$ and the estimated mean is denoted $\hat{\mu}_A$. Sample B is of size $n_B = 30$ and the estimated mean is denoted $\hat{\mu}_B$. Sample C is of size $n_C = 100$ and the estimated mean is denoted $\hat{\mu}_C$.

The question is now whether the sample mean will exceed the value 33, even when the population mean is equal to 30.

Which statement is correct?

1* ☐ $P(\hat{\mu}_A \geq 33) > P(\hat{\mu}_B \geq 33)$

2 ☐ $P(\hat{\mu}_C \geq 33) > P(\hat{\mu}_A \geq 33)$

3 ☐ $P(\hat{\mu}_C \geq 33) = P(\hat{\mu}_B \geq 33)$

4 ☐ $P(\hat{\mu}_A \geq 33) = P(\hat{\mu}_B \geq 33) \cdot P(\hat{\mu}_C \geq 33)$

5 ☐ $P(\hat{\mu}_A \geq 33) = \frac{1}{2} P(\hat{\mu}_B \geq 33)$

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

Lets go through the statements:

1. TRUE statement. The $\hat{\mu}$ is the sample mean, which we know follow the distribution $\hat{\mu} \sim N(\mu, \sigma^2/n)$ (Theorem 3.3), so we get the following

$$\hat{\mu}_A \sim N(30, 81/10)$$
$$\hat{\mu}_B \sim N(30, 81/30)$$
$$\hat{\mu}_C \sim N(30, 81/100)$$

and we can actually then realize, that the probability of getting a an outcome above the same value, must be higher for X_A than the two others, since its pdf has higher variance than the others. In R we can check it by:

```
## P(X_A >= 33)
(1-pnorm(q=33, mean=30, sd=sqrt(81/10)))

## [1] 0.1459203

## P(X_B >= 33)
(1-pnorm(q=33, mean=30, sd=sqrt(81/30)))

## [1] 0.03394458
```

2. FALSE statement. Following same argument as above
3. FALSE statement. Since the variance is different, then they are not equal
4. FALSE statement. Be sure by checking the product in R:

```
(1-pnorm(q=33, mean=30, sd=sqrt(81/30))) *
  (1-pnorm(q=33, mean=30, sd=sqrt(81/100)))

## [1] 1.456427e-05
```

5. FALSE statement. Be sure by checking the product in R:

```
0.5 * (1-pnorm(q=33, mean=30, sd=sqrt(81/30)))

## [1] 0.01697229
```

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

Continues on page 22

Exercise IX

The yield from a chemical process, Y_i, is assumed to depend linearly on the temperature, t_i, measured in degrees. In order to achieve insight about this relation, an experiment has been conducted where $n = 50$ pairwise measurements of Y_i and t_i has been taken. It is assumed that the following model can give a reasonable description of the relation

$$Y_i = \beta_0 + \beta_1 \cdot t_i + \varepsilon_i.$$

The residuals in this model are assumed independent and normally distributed with constant variance, i.e. $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Relevant output from the analysis in R is given below:

```
Call:
lm(formula = y ~ t)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0816 -1.4994 -0.2493  1.5175  4.8506

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  65.4919     2.7757  23.595   <2e-16 ***
t             0.1637     0.1103   1.485    0.144
---


Residual standard error: 2.296 on 48 degrees of freedom
Multiple R-squared:  0.04392,Adjusted R-squared:  0.024
F-statistic: 2.205 on 1 and 48 DF,  p-value: 0.1441
```

Question IX.1 (16)

Which of the following statements is correct when the significance level $\alpha = 0.05$ is applied?

1 ☐ The yield increases by 16.37% when the temperature increase one degree

2* ☐ There is no significant linear relation between temperature and yield

3 ☐ The test statistics for no effect of temperature on yield (i.e. the null hypothesis $H_0 : \beta_1 = 0$) is 23.595

4 ☐ A 95% confidence interval for the effect of temperature, β_1, is [-0.132027, 0.4594821]

5 ☐ The correlation between temperature and yield is 0.04392

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

Lets go through the answers one by one:

1. FALSE statement. The yield is estimated to increase 0.1637 units (we are not informed about the units) per degree, which is not the same as 16.37% (increasing some proportion per degree, would also lead to an exponential relation, not linear)

2. TRUE statement. The test of the null hypothesis

   $$H_0 : \beta_1 = 0$$

   leads to a p-value of 0.144, which is not below the significance level $\alpha = 0.05$ and since this is equivalent to testing for correlation equal to zero

   $$H_0 : \rho = 0$$

   there is not found a significant linear relation between the yield and the temperature

3. FALSE statement. Since, the test statistic for no effect is 1.485

4. FALSE statement. The lower limit of the CI is $0.1637 - 1.96 * 0.1103 = -0.052$ and the upper is $0.1637 + 1.96 * 0.1103 = 0.380$

5. FALSE statement. The correlation is $\sqrt{r^2} = \sqrt{0.04392} = 0.21$

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

Continues on page 24

Question IX.2 (17)

We continue with the exercise from the previous page. It turns out that the pH of the process may influence the yield, and since pH has been measured, it is decided to include it into the model, which in its extended form becomes:

$$Y_i = \beta_0 + \beta_1 \cdot t_i + \beta_2 \cdot pH_i + \varepsilon_i.$$

Estimation of the model parameters gives the following output in R:

```
Call:
lm(formula = y ~ t + pH)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7253 -1.2818 -0.2978  1.0724  4.4488

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.46756    4.09799  12.071 5.25e-16 ***
t            0.24113    0.09315   2.589   0.0128 *
pH           2.37090    0.50097   4.733 2.06e-05 ***
---
Signif. codes:  0  '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.91 on 47 degrees of freedom
Multiple R-squared:  0.3525,Adjusted R-squared:  0.3249
F-statistic: 12.79 on 2 and 47 DF,  p-value: 3.667e-05
```

Give estimates for the model parameters, i.e. β_0, β_1, β_2 and σ_ε^2

1 ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (4.09799,\ 0.24113,\ 2.37090,\ 0.3525)$

2* ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (49.46756,\ 0.24113,\ 2.37090,\ 1.91^2)$

3 ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (49.46756,\ 0.24113,\ 2.37090,\ 1.91 \cdot 47)$

4 ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (4.09799,\ 0.09315,\ 0.50097,\ 1.91 \cdot 47)$

5 ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (2.37090,\ 0.50097,\ 4.733\ ,\ 1.91)$

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

The estimates are read directly from the printed output. See Example 6.3

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

Continues on page 25

Question IX.3 (18)

We continue with the exercise from the previous page and the model

$$Y_i = \beta_0 + \beta_1 \cdot t_i + \beta_2 \cdot pH_i + \varepsilon_i.$$

Provide a 95% confidence interval for the effect on yield when pH increases one unit:

1 ☐ $0.24113 \pm 2.01174 \cdot 0.09315$

2* ☐ $2.37090 \pm 2.01174 \cdot 0.50097$

3 ☐ $(49.46756 + 0.24113 + 2.37090) \pm 2.01174 \cdot (4.09799 + 0.09315 + 0.50097)$

4 ☐ $2.37090 \pm 0.509920 \cdot 0.50097$

5 ☐ $(49.46756 + 0.24113 + 2.37090) \pm 0.509920 \cdot 0.50097$

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

See Method 6.5. The confidence interval for the effect of pH is found inserting the printed values into

$$\hat{\beta}_2 \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_2}$$

using the t-distribution with $n - (p + 1) = 47$ degrees of freedom to find the quantile $t_{1-\alpha/2}$:

```
qt(p=0.975, df=47)

## [1] 2.011741
```

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

Exercise X

Assume there exists a dice with 10 sides and where the probability for each of the 10 outcomes, $1, 2, \ldots, 10$, is the same. Consider the discrete random variable X with density $f(x) = 0.1$ for $x \in (1, 2, \ldots, 10)$.

Question X.1 (19)

Give the mean value of X:

1 ☐ $\frac{1}{(10-1)} \sum_{i=1}^{10} x_i = 6.11$

2 ☐ $\frac{1}{(10-6.11)} \sum_{i=1}^{10} |x_i - 6.11| = 6.48$

3 ☐ $\frac{1}{(10)} \sum_{i=1}^{10} (x_i - 6.11)^2 = 8.62$

4 ☐ $\sum_{i=1}^{10} \frac{10-1}{10} x_i \cdot 0.1 = 4.95$

5* ☐ $\sum_{i=1}^{10} x_i \cdot 0.1 = 5.50$

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

See Definition 2.13. We use the formula for calculating the mean value of a discrete random variable

$$\sum_{i=1}^{n} x_i f(x_i)$$

and insert the values. In R:

```
sum(1:10*0.1)

## [1] 5.5
```

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

Continues on page 28

Exercise XI

The yield of a process is $\mu = 60$ mg/l. Certain changes to the process are being planed and it is desirable to be able to prove an effect on the mean yield if the change is at least 5 mg/l (i.e. a two-sided test).

An engineer is now going to plan an experiment to evaluate the effect of the process changes. He wants to decide how large a sample is needed. The sample size has to be large enough to detect the relevant effect (5 mg/l) with a power of 0.8 when applying a significance level of $\alpha = 0.05$. It can be assumed that the standard deviation is $\sigma = 10$ mg/l.

Question XI.1 (20)

Based on the information above, and by applying the function `power.t.test` in R, one concludes that, if an equal number of measurements are taken, then the minimum number of measurements n needed becomes:

1 ☐ $n \simeq 256$ measurements

2 ☐ $n \simeq 128$ measurements

3 ☐ $n \simeq 64$ measurements

4* ☐ $n \simeq 34$ measurements

5 ☐ $n \simeq 27$ measurements

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

Based on the given information the planned test is a one-sample test, since it is not stated that a sample should be taken before the change, only that the yield before is $\mu = 60$ mg/l. See Example 3.67.

```
power.t.test(delta=5, sd=10, sig.level=0.05, power=0.8, type="one.sample")
```

```
##
##      One-sample t test power calculation
##
##               n = 33.3672
##           delta = 5
##              sd = 10
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Rounding up to $n \simeq 34$ measurements.

Since, it is not completely clear, that the it should not be a two-sample setup – one could argue that a nothing in the information given prevents it from being a two-sample test – then Answer 3 is also taken as correct, since:

```
power.t.test(delta=5, sd=10, sig.level=0.05, power=0.8, type="two.sample")

##
##      Two-sample t test power calculation
##
##               n = 63.76576
##           delta = 5
##              sd = 10
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Further, since it is also not specified that n is the number of measurements is in each group (and not the total), then Answer 2 is also taken as correct.

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

Continues on page 30

Exercise XII

In a study the aim is to investigate the possible cholesterol lowering effect of a product. 9 test persons had their cholesterol level measured (denoted `x1`). After 3 months, while using the product, the same 9 test persons had their cholesterol level measured again (denoted `x2`). Data is shown in the table below:

Person	1	2	3	4	5	6	7	8	9
x1	63.5	66.7	59.2	57.4	63.9	63.2	60.7	62.6	63.3
x2	51.3	51.9	57.8	50.2	54.6	43.3	51.2	40.4	52.2

The following code is now run in R, in order to test whether the change over time can be assumed to be zero ($H_0 : \delta = 0$):

```
x1 <- c(63.5, 66.7, 59.2, 57.4, 63.9, 63.2, 60.7, 62.6, 63.3)
x2 <- c(51.3, 51.9, 57.8, 50.2, 54.6, 43.3, 51.2, 40.4, 52.2)
```

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 = -5.6354, df = A, p-value = B
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.847799  C
sample estimates:
mean of the differences
-11.95556
```

Question XII.1 (21)

What conclusion can be made when applying a significance level of $\alpha = 0.05$?

1 ☐ We can show an effect since $\mu_D = -11.95556$

2 ☐ We can not show an effect since the upper limit of the confidence interval is 7.063312

3 ☐ We can not show an effect since the lower limit of the confidence interval is -7.063312

4* ☐ We can show an effect since the p-value is $4.897 \cdot 10^{-4}$

5 ☐ We can show an effect since the p-value is $2.394 \cdot 10^{-4}$

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

The standard statistical test for this setup is a paired two-sample t-test. The R output is from `t.test()`, and the easiest way to solve this is by copying and running

```
x1 <- c(63.5, 66.7, 59.2, 57.4, 63.9, 63.2, 60.7, 62.6, 63.3)
x2 <- c(51.3, 51.9, 57.8, 50.2, 54.6, 43.3, 51.2, 40.4, 52.2)
## The call is then either "t.test(x2, x1, paired=TRUE)" or
t.test(x2-x1)

##
##  One Sample t-test
##
## data:  x2 - x1
## t = -5.6354, df = 8, p-value = 0.0004897
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -16.847799  -7.063312
## sample estimates:
## mean of x
## -11.95556
```

and from the p-value we can find the correct answer. See section 3.1.7 for more examples.

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

Continues on page 32

Exercise XIII

A biologist is interested in examining the effect of 4 different growth inhibitors, denoted V_1, V_2, V_3 og V_4. The 4 growth inhibitors are added to samples from the same cell line and growth after one week is measured Y_{ij} (number of cells per cm^2). 8 replicates are made for each growth inhibitor, i.e. we have a total of 32 measurements. As the measurements can be assumed normally distributed, it is chosen to apply the following analysis of variance model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}.$$

In this model α_i denotes the effect of growth inhibitor i ($i = 1, 2, 3, 4$), μ is the overall average ε_{ij} are the errors, assumed independent and normally distributed with mean zero and standard deviation σ_ε.

An analysis of variance is performed for the above model and the output is given below. Please note that the output is incomplete as some numbers are replaced by the symbols A, B and C.

```
Analysis of Variance Table

Response: growth
          Df Sum Sq Mean Sq F value    Pr(>F)
treatment  A 281.07       B       C 0.0001409 ***
Residuals 28 268.46   9.588
```

Question XIII.1 (22)

Provide the usual test statistics (denoted by `C`) in order to test for equal mean effect of the 4 growth inhibitors

1* ☐ 9.77

2 ☐ 7.23

3 ☐ 2.95

4 ☐ 4.57

5 ☐ 16.11

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

As stated in Theorem 8.6, we can calculate the observed test statistic by

$$F_{\text{obs}} = \frac{SS(Tr)/(k-1)}{SSE/(n-k)} = \frac{281.07/(4-1)}{268.46/(32-4)} = 9.77,$$

where

- $SS(Tr)$ is the variance explained by the effect of the treatment
- SSE is the variance remaining after the model (sum of squared error)
- n is the total number of observations
- k is the number of groups

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

Continues on page 34

Question XIII.2 (23)

We now want to calculate a post hoc 95% confidence interval for a difference in mean between growth inhibitor V_1 and V_2, here denoted $I_{0.95}(V_1 - V_2)$. From the experiment it is known that the estimated mean difference between V_1 and V_2 is 4.5. State the interval $I_{0.95}(V_1 - V_2)$:

1 ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 2.048 \cdot \frac{9.588}{12} \cdot \sqrt{28}$

2* ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 2.048 \cdot \sqrt{9.588} \cdot \sqrt{2/8}$

3 ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 2.306 \cdot \frac{\sqrt{9.588}}{\sqrt{12}}$

4 ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 2.306 \cdot 9.588^2 \cdot \sqrt{1/8}$

5 ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 1.960 \cdot \frac{9.588}{\sqrt{8}}$

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

See method 8.9. The post hoc confidence interval for the difference is

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2}\sqrt{\frac{SSE}{n-k}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

So we use the t-distribution with $n - k = 32 - 4 = 28$ degrees of freedom

```
qt(p=0.975, df=28)

## [1] 2.048407
```

and insert the values

$$4.5 \pm 2.048 \cdot \sqrt{\frac{268.46}{28}\left(\frac{1}{8} + \frac{1}{8}\right)},$$

which we cannot directly find among the answers, so we shorten it

$$4.5 \pm 2.048 \cdot \sqrt{9.588\left(\frac{2}{8}\right)},$$

and finally find the answer

$$4.5 \pm 2.048 \cdot \sqrt{9.588} \cdot \sqrt{2/8}.$$

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

Exercise XIV

We consider a continuous random variable random, where the well-known cumulative distribution function $F(x)$ is given by $P(X \leq x) = 1 - e^{-x/2}$, where $x > 0$.

Question XIV.1 (24)

Provide the mean of X:

1 ☐ $\frac{1}{2}$

2 ☐ 1

3* ☐ 2

4 ☐ $\frac{3}{2}$

5 ☐ 4

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

It is recognized as the cdf of the exponential distribution (Definition 2.48), which is verified by

$$\int_0^x \lambda e^{\lambda y} dy = \left[-e^{-\lambda y} + c\right]_0^x = -e^{-\lambda x} + e^0 = 1 - e^{-\lambda x}$$

and it can be seen that $\lambda = \frac{1}{2}$. Using the formula for the mean of an exponential distribution (Theorem 2.49)

$$\mu = \frac{1}{\lambda} = 2.$$

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

Continues on page 36

Exercise XV

A biologist is examining the bio-diversity within an area and has measured the number of different type of plants per 10 m^2 in different places in the area. She has obtained a total of 30 independent measurements, y_i, and these are in in the vector `Yobs` in R.

Question XV.1 (25)

The biologist would like to estimate a 95% confidence interval for the coefficient of variation for the bio-diversity (number of different type of plants per 10 m^2) by applying the non-parametric bootstrap. Which of the following suggestions in R is most suitable to achieve this?

1 ☐
```
samples = replicate(10000,rnorm(30,mean(Yobs),sd(Yobs))
results = apply(samples,2,sd)/apply(samples,2,mean)
quantile(results, c(0.025,0.975))
```

2 ☐
```
samples = replicate(10000,sample(Yobs,replace=TRUE))
results = apply(samples,2,var)/apply(samples,2,sd)
quantile(results, c(0.025,0.975))
```

3 ☐
```
samples = replicate(10000,rnorm(30,mean(Yobs),sd(Yobs))
results = apply(samples,2,var)/apply(samples,2,median)
quantile(results, c(0.025,0.975))
```

4 ☐
```
samples = replicate(10000,sample(Yobs,replace=FALSE))
results = apply(samples,2,sd)/apply(samples,2,mean)
quantile(results, c(0.025,0.975))
```

5* ☐
```
samples = replicate(10000,sample(Yobs,replace=TRUE))
results = apply(samples,2,sd)/apply(samples,2,mean)
quantile(results, c(0.025,0.975))
```

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

In the code in Answer 1 and 3 the samples are simulated using `rnorm()`, hence a normal distribution is assumed and it is not non-parametric bootstrapping (but parametric).

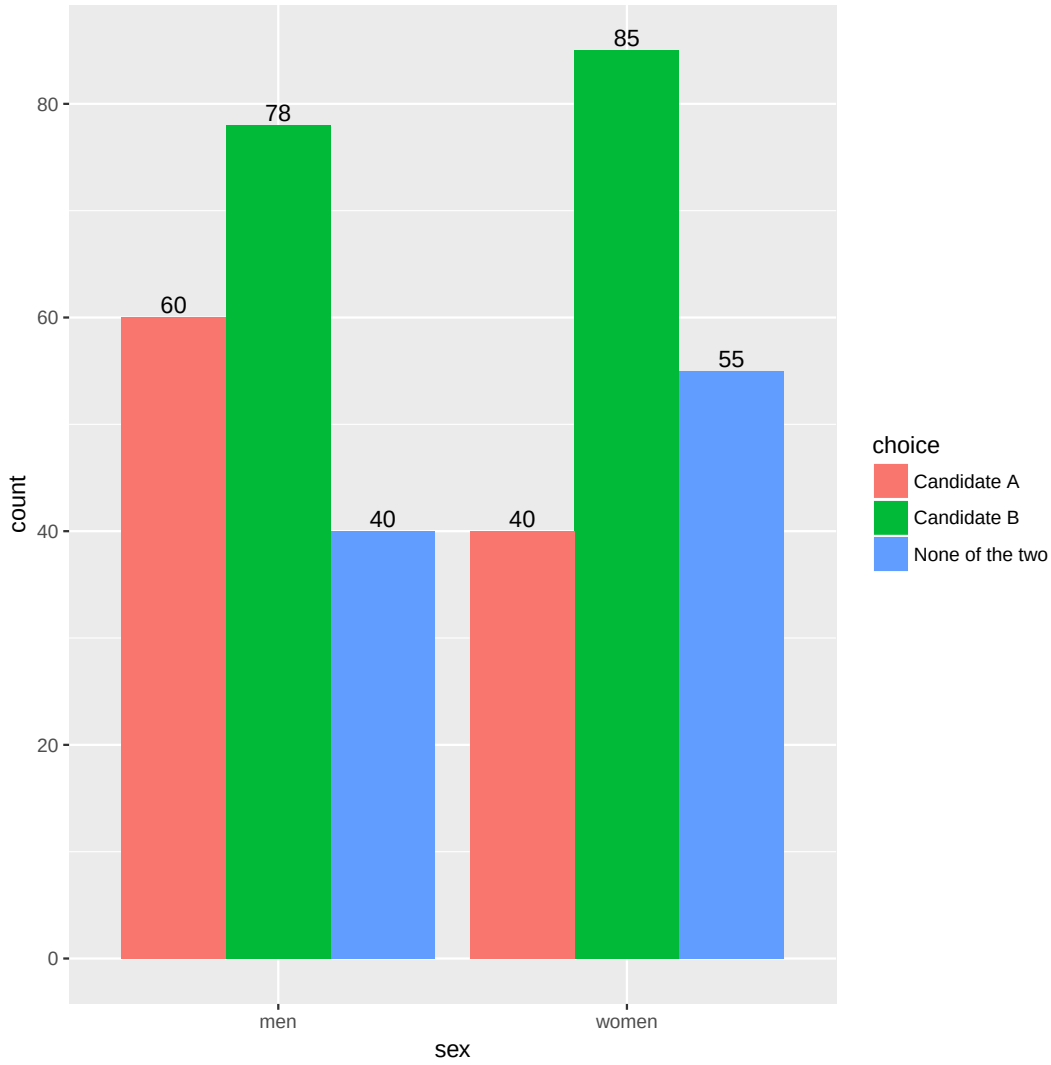
In Answer 2 it is not the coefficient of variation which is calculated by
`apply(samples,2,var)/apply(samples,2,sd)`,
which it is in Answer 4 and 5 by
`apply(samples,2,sd)/apply(samples,2,mean)`.
The difference between 4 and 5 is that in Answer 4 the samples are drawn without replacement
`sample(Yobs,replace=FALSE)`,
which is wrong, where in Answer 5 the samples are drawn correctly with replacement
`sample(Yobs,replace=TRUE)`. See Chapter 4.3 for more on non-parametric bootstrap.

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

Continues on page 37

Exercise XVI

In a study 178 men and 180 women were asked to answer whom of 2 political candidates, A or B, they preferred. Alternatively, they could answer "none of the two". The distribution of the answers is shown in the figure below.



Continues on page 38

Question XVI.1 (26)

It is seen from the figure that we observe that 85 out of the 180 women prefer Candidate B. If we can assume the same distribution of answers by gender, how many women out of the 180 would we expect to prefer Candidate B?

1 ☐ $\frac{163}{358} \cdot \frac{95}{358} \cdot 358$

2 ☐ $\frac{100}{358} \cdot \frac{223}{358} \cdot 358$

3 ☐ $\frac{95}{358} \cdot \frac{190}{358} \cdot 358$

4* ☐ $\frac{163}{358} \cdot \frac{180}{358} \cdot 358$

5 ☐ $\frac{95}{358} \cdot \frac{180}{358} \cdot 358$

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

See chapter 7.2. The total number of respondents are $n = 180 + 178 = 358$ and if we assume the same distribution of answers by gender, i.e. the under the hypothesis that the proportion of men and women prefering B is equal

$$H_0 : p_{\text{men},B} = p_{\text{women},B} = p,$$

then

$$p = \frac{\text{"Total number for B"}}{\text{"Total number"}} = \frac{78 + 85}{358} = \frac{163}{358}.$$

It is then simply this fraction we expect out of the total number of women

$$\frac{163}{358} \cdot 180,$$

which is then expressed a little longer by

$$\frac{163}{358} \cdot \frac{180}{358} \cdot 358.$$

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

Question XVI.2 (27)

Provide the usual test statistics when you want to conduct the test of whether the distribution of answers is the same for men and women:

1 ☐ $\chi^2_{\text{obs}} = 5.9915$

2* ☐ $\chi^2_{\text{obs}} = 6.6581$

3 ☐ $\chi^2_{\text{obs}} = 16.212$

4 ☐ $\chi^2_{\text{obs}} = 8.3836$

5 ☐ $\chi^2_{\text{obs}} = 4.5067$

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

Maybe the easiest is to copy example 7.21 from the book of testing multiple proportions

```
prop <- matrix(c(60, 78, 40, 40, 85, 55), ncol = 3, byrow = TRUE)
rownames(prop) <- c("Men", "Women")
colnames(prop) <- c("A", "B", "None")
prop

##         A  B None
## Men    60 78   40
## Women 40 85   55

chisq.test(prop, correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  prop
## X-squared = 6.6581, df = 2, p-value = 0.03583
```

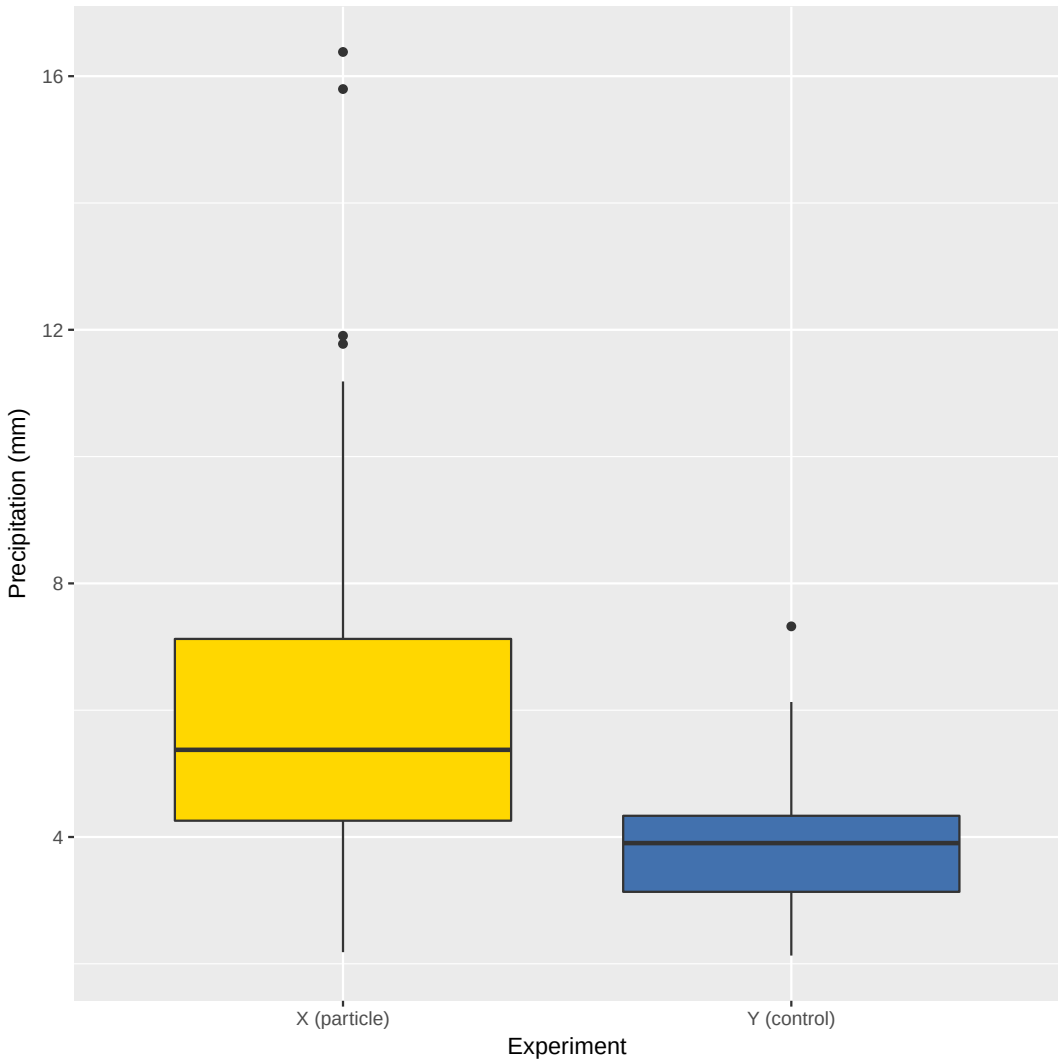
---------- FACIT-END ----------

Continues on page 40

Exercise XVII

Cloud seeding is a form of weather modification that can be used to increase the amount of precipitation that falls from the clouds, by dispersing substances (small particles) e.g. aluminiumoxid into the clouds to modify their development.

In an experiment the aim was to study the effect of cloud seeding by using a new type of particles. The amount of precipitation (mm precipitation per day) for 35 days with cloud seeding using the new particles is denoted X_i, $(i = 1, 2, \ldots, 35)$. This was compared to the amount of precipitation on 30 days without cloud seeding, denoted Y_j, $(j = 1, 2, \ldots, 30)$. Measurements were only taken on days where there was sufficient humidity in the air to make the experiment relevant. Data from the experiment is shown in the figure below.



Continues on page 41

We now want to analyze the data described on the previous page using R. Data x_i is stored in the vector `x` and data y_j is stored in the vector `y`, and the following code has been run:

```
k <- 10^4
resultX <- replicate(k, sample(x, replace = TRUE))
resultY <- replicate(k, sample(y, replace = TRUE))
result <- apply(resultX, 2, median) - apply(resultY, 2, median)
quantile(result, c(0.5, 0.025,0.975))
```

Which gives the result

```
      50%      2.5%     97.5%
1.6283069 0.2843492 2.4233546
```

Question XVII.1 (28)

If we apply a significance level of $\alpha = 0.05$ what can then be concluded?

1* ☐ The median for X is significantly higher than the median for Y

2 ☐ The median for X is 62.8% higher than the median for Y

3 ☐ Precipitation for X is between 28.4% and 142.3% higher than precipitation for Y

4 ☐ The mean precipitation can be assumed equal for the two methods

5 ☐ The medianen for Y is $[0.28; 2.42]$ higher than the median for X

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

In the R code a 95% non-parametric bootstrap confidence interval for the difference in median is calculated, and since 0 is not contained in the interval, then the hypothesis

$$H_0 : q_{0.5,X} = q_{0.5,Y}$$

must be rejected on significance level $\alpha = 0.05$, thus concluded that

$$H_1 : q_{0.5,X} \neq q_{0.5,Y}$$

and further, since $X - Y$ was calculated and the interval is on the positive side, then it can be concluded that $q_{0.5,X} > q_{0.5,Y}$.

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

Continues on page 42

Question XVII.2 (29)

In a different experiment using cloud seeding a different kind of particles were examined. Also in this experiment the amount of precipitation was compared when the particles were used to a situation with no use of particles. In this study, however, it was decided to log transform (the natural logarithm) the data before comparing the groups. By transforming the data it can be assumed that data in the two groups follows a normal distribution. The data is summarized in the table below (unit is log mm precipitation).

	Particles, X (log mm precipitation)	Control, Y (log mm precipitation)
Estimated mean	$\hat{\mu}_X = 1.573$	$\hat{\mu}_Y = 1.314$
Estimated variance	$\hat{\sigma}^2_X = 0.333$	$\hat{\sigma}^2_Y = 0.171$
Number of observations	$n_X = 35$	$n_Y = 30$

We now want to test whether the means of the 2 groups can be assumed equal, i.e.

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y$$

It is given that the usual test statistics assuming the null hypothesis becomes 2.0958 with 61.19 degrees of freedom. State the p-value and conclusion when a significance level of $\alpha = 0.05$ is applied:

1 ☐ p-value $\simeq 0.82$ i.e. H_0 is accepted

2 ☐ p-value $\simeq 0.41$ i.e. H_0 is rejected

3 ☐ p-value $\simeq 0.21$ i.e. H_0 is accepted

4 ☐ p-value $\simeq 0.10$ i.e. H_0 is rejected

5* ☐ p-value < 0.05 i.e. H_0 is rejected

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

This is a two-sample t-test and we get the information we need from $t_{\text{obs}} = 2.0958$ and degrees of freedom is 61.19, so the p-value is calculated by

```
2 * (1-pt(abs(2.0958), df=61.19))

## [1] 0.04024393
```

which is lower than 0.05, so we reject the null hypothesis.

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

Continues on page 43

Exercise XVIII

At a Christmas marked there is a lottery. 24 balls are placed in bowl. On each of 4 balls there is a picture of a star. On each of the remaining 20 balls there is a picture of an elf. The lottery is now played so that 2 balls are drawn without replacement from the bowl. If both balls show a picture of a star then you have won a prize!

Question XVIII.1 (30)

You participate in the game once. Provide the probability of winning a prize:

1 ☐ $\frac{80}{276}$

2 ☐ $\frac{56}{276}$

3 ☐ $\frac{40}{276}$

4 ☐ $\frac{16}{276}$

5* ☐ $\frac{6}{276}$

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

This is drawing without replacement, hence we must use the hypergeometric distribution (Chapter 2.3.2). However, to get most easily to the answer in the presented form, we can use the basic definition of probability

$$P(\text{success}) = \frac{x}{n},$$

where x is the number of successes in a population of size n. We need possible successful combinations, where a ball with a star is drawn. In the first draw one out of the four must be drawn and in the second draw one out of the three remaining must be drawn, thus

$$x = 4 \cdot 3 = 12.$$

The number of elements in the population (of possible draws) is

$$n = 24 \cdot 23 = 552,$$

since in the first draw there are 24 balls and in the second there are one less. Put together this gives

$$\frac{12}{552} = \frac{6}{276}.$$

Alternatively, the x number of successful combinations could be calculated by

```r
dhyper(x=2, m=4, n=20, k=2)
```

```
## [1] 0.02173913
```

which multiplied with the population size gives x

```r
dhyper(x=2, m=4, n=20, k=2) * (24*23)
```

```
## [1] 12
```

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

The exam is finished. Have a great Christmas vacation!

Technical University of Denmark *Page 1 of 43 sider.*

Written examination: 28. May 2017

Course name and number: **Introduction to Statistics (02323 and 02402)**

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 11 exercises. To answer the questions you need to fill in the prepared 30-question multiple choice form (on 6 seperate pages) in CampusNet. **There is one and only one correct answer to each question**.

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"). Also, if more answers are 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 answer of the exercises should be given by filling in and submitting via the exam module in CampusNet. The table sheet here is ONLY to be used as an "emergency" alternative.

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

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

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

In case of "emergency": Remember to provide your **study number**. The questionnaire contains 43 pages. Please check that your questionnaire contains them all.

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

The island of Moen is the first official Dark Sky area in the Nordic area. A man went there to watch stars and shooting stars. It was a clear night and he had read that he should expect to see 2 shoothing stars per minute.

It can be assumed that the intensity of shooting stars is constant and that they arrive independently of each other.

Question I.1 (1)

Which distribution gives the best description of the waiting time in minutes between two shooting stars?

1* ☐ An exponential distribution with $\lambda = 2$

2 ☐ A Poisson distribution with $\lambda = 2$

3 ☐ A Normal distribution with $\mu = 1$ and $\sigma = \sqrt{2}$

4 ☐ A uniform distribution with $\alpha = 0$ and $\beta = 2$

5 ☐ A Poisson distribution with $\lambda = 1/2$

——— FACIT-BEGIN ———

As described in Section 2.5.4, waiting times between independent arrivals are best described by an exponential distribution. As seen from Definition 2.27, λ is the average intensity (the number of events per time period), so $\lambda = 2$.

——— FACIT-END ———

Question I.2 (2)

What is the probability that there is no shooting stars in a given 4 minutes interval?

1 ☐ 0.8824969

2 ☐ 0.1353353

3 ☐ 0.9996645

4* ☐ 3.3546263×10^{-4}

5 ☐ There must have been at least one shooting star in a given 4 minutes interval.

———— FACIT-BEGIN ————

The number of independent events in a given time interval is Poisson distributed. The parameter in the Poisson distribution is the expected number of events which is 2 per minute times 4 minutes = 8 (See equation 2-33 about scaling). So the probability of observing zero can be found as `ppois(0,8)` or `pexp(4,2, lower.tail=FALSE)`.

———— FACIT-END ————

Continues on page 4

Exercise II

When inspecting cars the combustion is tested by measuring the exhaust gasses. Assume that a given car has a true exhaust of particles of 0.12 g/km where the EURO2 norm limit is 0.08 g/km. The measurement has a standard deviation of 0.02 g/km. It is assumed that the measurements are normally distributed around the true exhaust. A car is tested by a single measurement.

Question II.1 (3)

What is the probability that the car is tested as having too high exhaust of particles?

1 ☐ 0.6113513

2 ☐ 0.0227501

3 ☐ 0.3886487

4* ☐ 0.9772499

5 ☐ 0.6113513

———— FACIT-BEGIN ————

The distribution of the measurements for the car in question is: $X \sim N(0.12, 0.02^2)$ So the answer is:

$$P(X > 0.08)$$

Or in R:

```
1-pnorm(0.08, m=0.12, sd= 0.02)
```

```
## [1] 0.9772499
```

———— FACIT-END ————

Continues on page 5

Exercise III

In a purification experiment, the so-called yield was measured after dosing a certain amount of enzyme. The response variable was the yield percentage in relation to the theoretical highest obtainable level (X). Data from 10 different test samples from the experiment were:

x_i
74.7
74.2
74.1
69.6
75.4
76.3
76.7
75.6
72.0
74.3
$\bar{x} = 74.29$
$s = 2.115$

Question III.1 (4)

What is the 80% percentile for these data using the definition from the book?

1 ☐ 74.10

2 ☐ 74.50

3* ☐ 75.95

4 ☐ 74.29

5 ☐ 75.60

———— FACIT-BEGIN ————

See Definition 1.7. Either we can use R (remember type=2, to ensure that the book's definition is used)

```
quantile(x, 0.80, type=2)
```

```
##   80%
## 75.95
```

Or we do it manually and take the mean of the 8th and 9th observation

```
## Or:
sort(x)

##  [1] 69.6 72.0 74.1 74.2 74.3 74.7 75.4 75.6 76.3 76.7

mean(c(75.6,76.3 ))

## [1] 75.95
```

———— FACIT-END ————

Continues on page 7

Question III.2 (5)

Assuming that $X \sim N(\mu, \sigma^2)$ and applying the usual estimated parameters ($\mu = \bar{x}$ and $\sigma = s$), what is the only statement that can be correct:

1 ☐ More than 99% of the population is within $[72.18, 76.40]$ (In R: `mean(x) + c(-1, 1) * sd(x)`)

2 ☐ More than 99% of the population is within $[70.06, 78.52]$ (In R: `mean(x) + c(-1, 1) * 2 *sd(x)`)

3 ☐ Around 95% of the population is within $[72.18, 76.40]$ (In R: `mean(x) + c(-1, 1) * sd(x)`)

4* ☐ More than 99% of the population is within $[67.95, 80.63]$ (In R: `mean(x) + c(-1, 1) * 3 *sd(x)`)

5 ☐ Less than 95% of the population is within $[67.95, 80.63]$ (In R: `mean(x) + c(-1, 1) * 3 *sd(x)`)

———— FACIT-BEGIN ————

In a normal distribution approximately 95% of the population are within the mean plus/minus 2 standard deviations. More than 99% are within plus/minus 3 standard deviations. This means that only answer number 4 can be correct.

```
mean(x)
## [1] 74.29
sd(x)
## [1] 2.114737
qnorm(.995)
## [1] 2.575829
1-2*(1-pnorm(3))
## [1] 0.9973002
mean(x) + c(-1,1)*3*sd(x)
## [1] 67.94579 80.63421
mean(x) + c(-1,1)*2*sd(x)
## [1] 70.06053 78.51947
mean(x) + c(-1,1)*1*sd(x)
## [1] 72.17526 76.40474
```

———— FACIT-END ————

Question III.3 (6)

What is the 95% confidence interval for the mean?

1 ☐ $2.262 \pm 74.29\frac{2.115}{\sqrt{10}} = [-47.42, 51.95]$

2 ☐ $74.29 \pm 1.812\frac{2.115}{\sqrt{9}} = [73.01, 75.57]$

3 ☐ $74.29 \pm 1.96\frac{2.115}{\sqrt{9}} = [72.91, 75.67]$

4* ☐ $74.29 \pm 2.262\frac{2.115}{\sqrt{10}} = [72.78, 75.80]$

5 ☐ $74.29 \pm 1.96\frac{2.115^2}{\sqrt{10}} = [71.52, 77.06]$

———— FACIT-BEGIN ————

As seen from Method 3.9, a one-sample confidence interval has the form

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

using $n - 1$ degrees of freedom for the t-distribution. And as

```
qt(0.975, 9)
```

```
## [1] 2.262157
```

it becomes

$$74.29 \pm 2.262\frac{2.115}{\sqrt{10}}$$

which becomes:

```
74.29 + c(-1, 1)* 2.262*2.115/sqrt(10)
```

```
## [1] 72.77713 75.80287
```

Or completely by R:

```
x <- c(74.7,  74.2, 74.1, 69.6, 75.4, 76.3, 76.7, 75.6, 72.0, 74.3)
t.test(x)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 111.09, df = 9, p-value = 1.97e-15
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  72.77721 75.80279
## sample estimates:
## mean of x
##     74.29
```

———— FACIT-END ————

Continues on page 10

Question III.4 (7)

If you make a 95% confidence interval for the standard deviation, which quantiles should then be used?

1 ☐ `qnorm(0.025)` and `qnorm(0.975)`

2* ☐ `qchisq(0.025, 9)` and `qchisq(0.975, 9)`

3 ☐ `qf(0.025, 9, 9)` and `qf(0.975, 9, 9)`

4 ☐ `qt(0.025, 9)` and `qt(0.975, 9)`

5 ☐ `qunif(0.025)` and `qunif(0.975)`

———— FACIT-BEGIN ————

As seen from Method 3.19, the CI for a variance is based on the χ^2-distribution, so 2 is the only possible answer.

———— FACIT-END ————

Question III.5 (8)

The p-value for the hypothesis test of $H_0 : \mu = 70$ is:

1* ☐ `2*(1-pt(4.29/(2.115/sqrt(10)), 9))`

2 ☐ `2*(1-pnorm(70/(2.115/sqrt(10))))`

3 ☐ `(1-pt(-4.29/(2.115/sqrt(9)), 10))`

4 ☐ `1-pnorm(-4.29/(2.115/sqrt(10)))`

5 ☐ `1-qt(2.115/4.29, 9)`

———— FACIT-BEGIN ————

As described in Method 3.23, the observed t-test statistic is:

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{74.29 - 70}{2.115/\sqrt{10}} = \frac{4.29}{2.115/\sqrt{10}}$$

And the p-value then is:

$$2 * P(t > |t_{obs}|) = 2 * P(t > \frac{4.29}{2.115/\sqrt{10}})$$

as the t_{obs} is positive in this case. So answer 1 has the proper R-call.

———— FACIT-END ————

Question III.6 (9)

In a new experiment which is in the planning phase, a 95% confidence interval for the mean with an expected width of around 1 is wanted. Assume that the standard deviation is 2.115. How large a sample does it approximately require to achieve this desired precision?

1 ☐ 5

2 ☐ 1230

3 ☐ 4

4 ☐ 100

5* ☐ 69

———— FACIT-BEGIN ————

Use Method 3.63, with $\alpha = 0.05$ and $ME = 0.5$ (half of the confidence interval):

$$n = \left(\frac{z_{1-\alpha/2} \cdot \sigma}{ME}\right)^2 = \left(\frac{1.96 \cdot 2.115}{0.5}\right)^2 = 68.74 \approx 69$$

———— FACIT-END ————

Continues on page 12

Exercise IV

In a purification experiment, two different doses of an enzyme have been investigated called d_1 and d_2, with the purpose to investigate a possible effect on the yield. The response variable was the yield percentage in relation to the theoretical highest obtainable level. Data from 19 different test samples from the experiment were:

Dose d_1	Dose d_2
74.7	79.6
74.2	77.5
74.1	82.5
69.6	76.7
75.4	78.2
76.3	76.7
76.7	76.6
75.6	78.1
72.0	79.2
74.3	
$\bar{x}_1 = 74.29$	$\bar{x}_2 = 78.34$
$s_1 = 2.115$	$s_2 = 1.898$

The following was run in R:

```
x1 <- c(74.7, 74.2, 74.1, 69.6, 75.4, 76.3, 76.7, 75.6, 72.0, 74.3)
x2 <- c(79.6, 77.5, 82.5, 76.7, 78.2, 76.7, 76.6, 78.1, 79.2)
t.test(x2, x1)

##
##  Welch Two Sample t-test
##
## data:  x2 and x1
## t = 4.4041, df = 17, p-value = 0.0003878
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.112129 5.996760
## sample estimates:
## mean of x mean of y
##  78.34444  74.29000
```

Continues on page 13

Question IV.1 (10)

What is the 99% confidence interval for the difference in means between dose d_2 and dose d_1?

1 ☐ $[2.45, 5.66]$

2* ☐ $[1.39, 6.72]$

3 ☐ $[1.97, 6.14]$

4 ☐ $[2.11, 6.00]$

5 ☐ $[74.29, 78.34]$

———— FACIT-BEGIN ————

From Method 3.47, we see that the 99% two-sample confidence interval has the form

$$\bar{x}_2 - \bar{x}_1 \pm t_{0.995}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

From the R output we can read the degrees of freedom to be 17, rather than calculating it. We use this to find the t-quantile

```
qt(0.995, 17)

## [1] 2.898231
```

the interval becomes

$$4.054 \pm 2.898\sqrt{\frac{2.115^2}{10} + \frac{1.898^2}{9}}$$

which becomes:

```
4.054 + c(-1, 1)* 2.898*sqrt(2.115^2/10 + 1.898^2/9)

## [1] 1.385967 6.722033
```

Or completely by R:

```
t.test(x2, x1, conf.level = 0.99)

##
##  Welch Two Sample t-test
##
```

```
## data:  x2 and x1
## t = 4.4041, df = 17, p-value = 0.0003878
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  1.386305 6.722584
## sample estimates:
## mean of x mean of y
##  78.34444  74.29000
```

———— FACIT-END ————

Question IV.2 (11)

The conclusion of the usual t-test (based on $\alpha = 0.05$) for this situation is (both conclusion and argument must be correct):

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

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

3 ☐ The two means are approximately equal as the p-value is large

4 ☐ The two means are approximately equal as the p-value is small

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

———— FACIT-BEGIN ————

From the R output we see that it is a Welsch two-sample test, hence we are comparing whether the two mean values are different or identical. Since the p-value is 0.0004 it is much smaller than $\alpha = 0.05$. This means that we reject the null hypothesis ($H_0 : \mu_a = \mu_b$). Hence, they are significantly different.

———— FACIT-END ————

Continues on page 15

Question IV.3 (12)

A new study with 2 doses is planned. It is assumed that the standard deviation within each group is 2, and that a t-test on level $\alpha = 0.05$ should be carried out. The following things are run in R:

```
power.t.test(n=30, delta=2, sd=2, sig.level=0.05)

##
##      Two-sample t test power calculation
##
##               n = 30
##           delta = 2
##              sd = 2
##       sig.level = 0.05
##           power = 0.9677083
##     alternative = two.sided
##
## NOTE: n is number in *each* group

power.t.test(power=0.80, delta=1, sd=2, sig.level=0.05)

##
##      Two-sample t test power calculation
##
##               n = 63.76576
##           delta = 1
##              sd = 2
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Only one of the following statements is true. Which?

1 ☐ The risk that a study with $n = 30$ in each group does not find a significant difference between the means, if the real difference is 1, is around 97%

2 ☐ The chance that a study with $n = 30$ in each group finds a significant difference between the means, if the real difference is 1, is around 97%

3* ☐ The chance that a study with $n = 64$ in each group finds a significant difference between the means, if the real difference is 1, is around 80%

4 ☐ The risk that a study with $n = 64$ in each group does not find a significant difference between the means, if the real difference is 1, is around 80%

5 ☐ The risk that a study with $n = 30$ in each group does not find a significant difference between the means, if the real difference is 2, is around 97%

———— FACIT-BEGIN ————

Let us go through the answers:

1 A sample size of 30 is expressed in the first of the two tests, however in this test the difference delta that is specified is 2 (and not 1 like the question). Furthermore, power is expressing the probability of rejecting the null hypothesis (NOT the risk of no rejection).

2 This time the explanation of power is correct, however the wrong delta is still used for the first test.

3 This is the correct answer, since the delta, n and explanation of power is correct.

4 Like answer 1, this interpretation of power is incorrect

5 This refers to test one, and with the correct delta this time. However, the interpretation of what power is, is still incorrect.

See Section 3.3.2 for more on power.

———— FACIT-END ————

Continues on page 17

Exercise V

In Danish power plants materials are being burned to generate electricity and in this combustion CO2 is emitted. It's a gas which enhances the Greenhouse effect and therefore contributes to warming up the atmosphere. This has many negative consequences, and it is of interest to reduce these emissions. This is done by introducing more wind and solar energy production into the system.

Each day, CO2 emissions are calculated (in grams of CO2 equivalent gas) per kWh electricity produced in Denmark based on data from ENTSO-E about the production.

Column 1 of the table below shows the date. In Column 2, average values of CO2 emissions are given for the 15 days with the highest wind energy production in the period from December 1, 2016 to April 1, 2017. Column 3 in the table shows electricity generation with coal, this column is not used in the first two questions.

t	co2intensity (gCO2eq/kWh)	coal (MW)
2016-12-02	230	1016
2016-12-25	205	817
2016-12-26	203	746
2017-01-01	212	948
2017-01-05	292	1448
2017-01-12	260	1398
2017-02-08	317	1409
2017-02-12	321	1578
2017-02-21	235	1102
2017-02-22	268	1325
2017-02-23	233	1187
2017-03-01	253	1195
2017-03-02	260	1093
2017-03-16	212	976
2017-03-22	250	1095

The data is read into R in a data.table `X` and the following is run:

```
## Put observations in x
x <- X$co2intensity
## Number of simulated samples
k <- 100000
n <- length(x)

## Simulation
simsamples <- replicate(k, sample(x, replace = TRUE))
## Calculate the mean of each simulated sample
simmeans <- apply(simsamples, 2, mean)
```

Continues on page 18

```
## Quantiles of the differences gives the CI
quantile(simmeans, c(0.005, 0.995))

##     0.5%    99.5%
## 227.2667 275.2000

quantile(simmeans, c(0.01, 0.99))

##       1%      99%
## 229.3333 272.8000

quantile(simmeans, c(0.025, 0.975))

##     2.5%    97.5%
## 232.3333 269.2667

quantile(simmeans, c(0.05, 0.95))

##       5%      95%
## 235.0000 265.9333

quantile(simmeans, c(0.1, 0.9))

##      10%      90%
## 238.0667 262.4000
```

Question V.1 (13)

A 99% bootstrapped confidence interval for the mean of the CO2-intensity is wanted, without assumptions about the distribution. What is the correct interval?

1* ☐ $[227, 275]$

2 ☐ $[229, 273]$

3 ☐ $[232, 269]$

4 ☐ $[235, 266]$

5 ☐ $[238, 262]$

———— FACIT-BEGIN ————

We need to identify the interval based on the correct quantiles. It is the 99% confidence interval, so $\alpha = 0.01$ and therefore we need to take the $q_{\alpha/2} = q_{0.005}$ to $q_{1-\alpha/2} = q_{0.995}$ interval, which can be read from the output to be $[227, 275]$.

———— FACIT-END ————

Continues on page 20

Question V.2 (14)

Which one of the following statements is not correct? Each statement is about a null hypothesis for the mean level of CO2 intensity μ_{CO2} at high wind energy production, and the conclusion is drawn based on the results of the R code? (Note again: there are 4 true and 1 false statements - you must find the false statement!)

1 ☐ The null hypothesis $H_0 : \mu_{CO2} = 200$ would have been rejected on a 10% significance level

2 ☐ The null hypothesis $H_0 : \mu_{CO2} = 220$ would have been rejected on a 5% significance level

3 ☐ The null hypothesis $H_0 : \mu_{CO2} = 230$ would have been rejected on a 5% significance level

4 ☐ The null hypothesis $H_0 : \mu_{CO2} = 270$ would have been rejected on a 5% significance level

5* ☐ The null hypothesis $H_0 : \mu_{CO2} = 270$ would have been rejected on a 1% significance level

———— FACIT-BEGIN ————

The null hypothesis for a value of the mean is rejected if the value is outside the confidence interval. When looking at the R output above, the first four answers all fall outside the corresponding confidence interval and thus their conclusions are all correct. In the last statement the value 270 falls inside the 90% confidence interval and thus the null hypothesis would not have been rejected.

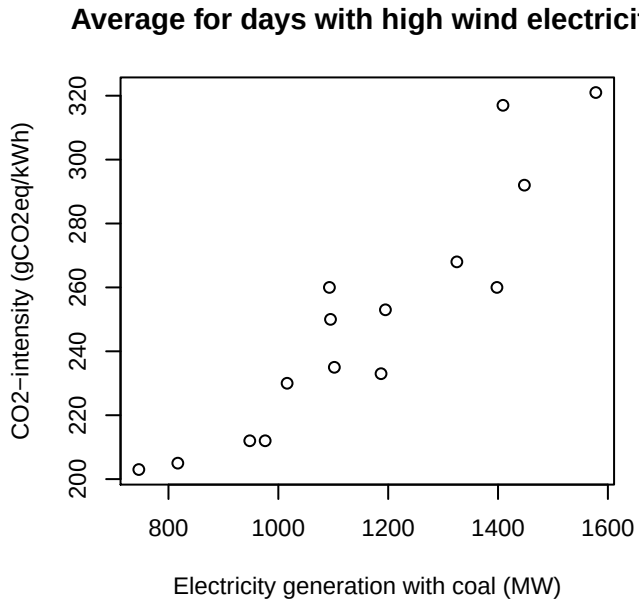
———— FACIT-END ————

Continues on page 21

Question V.3 (15)

In this question, the relationship between the electricity generation with coal (Column 3) and the CO2-intensity at high wind electricity generation (Column 2) is investigated.

To visualize the relation the following scatter plot with the observations is created:



Based on a consideration of the plot, which of the following statements is the most correct conclusion?

1 ☐ The correlation is approximately -1.2

2 ☐ The correlation is approximately 0.1

3 ☐ The correlation is approximately 0

4* ☐ The correlation is approximately 0.9

5 ☐ The correlation is approximately 1.2

———— FACIT-BEGIN ————

The correlation can only be between -1 to 1, and it is certainly not 0, and it is higher that 0.1, so clearly the most correct conclusion is that it is 0.9. In fact it is:

```
cor(X$coal, X$co2intensity)

## [1] 0.9206539
```

———— FACIT-END ————

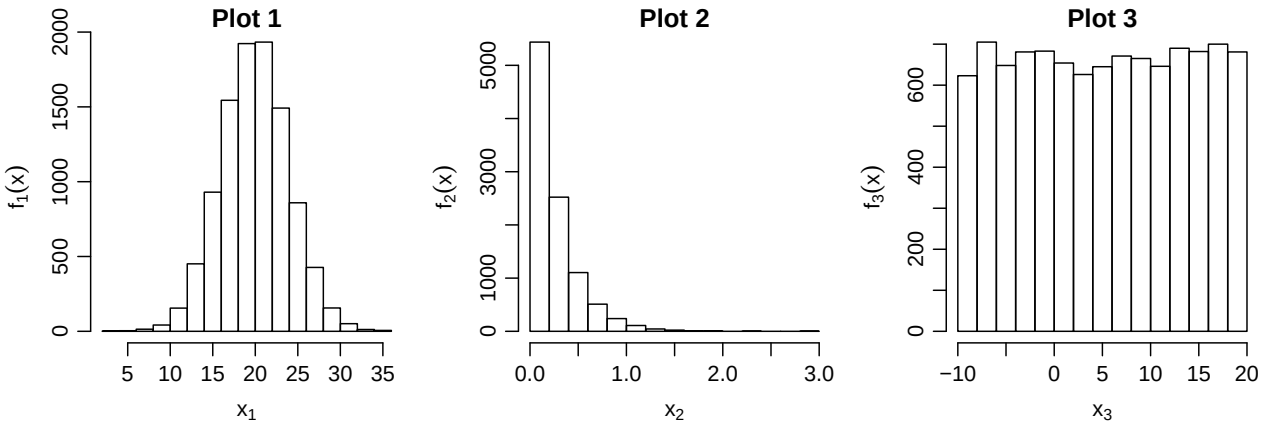
Continues on page 23

Exercise VI

This exercise is about random variables and simulation.

Question VI.1 (16)

The following three histograms are of values from simulations of $n = 10000$ observations from three distributions:



Which of the following distributions are simulated (both the ordering and parameter values must be correct)?

1 ☐ Plot 1: $X_1 \sim N(10, 4^2)$, Plot 2: $X_2 \sim \text{Exp}(4)$ and Plot 3: $X_3 \sim U(5, 20)$

2 ☐ Plot 1: $X_1 \sim U(5, 20)$, Plot 2: $X_2 \sim \text{Exp}(1)$ and Plot 3: $X_3 \sim N(20, 4^2)$

3 ☐ Plot 1: $X_1 \sim U(-10, 20)$, Plot 2: $X_2 \sim \text{Exp}(1)$ and Plot 3: $X_3 \sim N(20, 4^2)$

4* ☐ Plot 1: $X_1 \sim N(20, 4^2)$, Plot 2: $X_2 \sim \text{Exp}(4)$ and Plot 3: $X_3 \sim U(-10, 20)$

5 ☐ Plot 1: $X_2 \sim \text{Exp}(4)$, Plot 2: $X_1 \sim N(10, 4^2)$ and Plot 3: $X_3 \sim U(5, 20)$

——— FACIT-BEGIN ———

From the shape of the empirical distribution (the histograms) we identify: the normal distribution in Plot 1 (Example 2.39), the exponential distribution in Plot 2 (Example 2.50), and the uniform distribution in Plot 3 (Figure 2.3). Now we have to identify the correct parameters, which is easiest done in the uniform, which can be seen to go from $\alpha = -10$ to $\beta = 20$.

——— FACIT-END ———

Continues on page 24

Question VI.2 (17)

Let the random variables $X_i \sim N(2, 4^2)$ for $i = 1, \ldots, 20$ be i.i.d. and define the following random variables as function of these

$$\bar{X} = \frac{1}{20}\sum_{i=1}^{20} X_i,$$

$$S = \sqrt{\frac{1}{19}\sum_{i=1}^{20}(X_i - \bar{X})^2},$$

$$Y = \frac{\bar{X} - 2}{S}.$$

Which distribution does the random variable Y follow?

1 ☐ Y follows the normal distribution $N(0, 1^2)$

2 ☐ Y follows the normal distribution $N(0, 4^2)$

3 ☐ Y follows the χ^2-distribution with 20 degrees of freedom

4 ☐ Y follows t-distribution with 20 degrees of freedom

5* ☐ Y follows t-distribution with 19 degrees of freedom

——— FACIT-BEGIN ———

Comparing to Theorem 3.3 and Definition 1.11, we identify $\bar{X}$ as the sample mean and S as the sample standard deviation, all as a random variables (capitalized letters). Therefore Y is the S-standardized sample mean, which we know, as stated in Theorem 3.5, follows a t-distribution with $n-1$ degrees of freedom. From the sample mean formula we can see that there are $n = 20$ observations in the sample, so the t-distribution has 19 degrees of freedom.

——— FACIT-END ———

Continues on page 25

Exercise VII

A recreational runner wants to measure the effect of his training. For this purpose, he has measured values of average pulse (beats per minute), weeks in the training program and speed (km/h), for a particular stretch he runs frequently.

The recreational runner has decided to measure the effect by examining whether the average speed increases over time (weeks).

Data reading in R is:

```
week <- c(1, 1, 1, 3, 3, 4, 5, 5, 5, 5, 6, 6, 7, 7, 8, 9, 9, 10, 12, 12, 13, 13,
          15, 15, 15, 16, 16)

pulse <- c(137.6, 140.1, 143.0, 148.6, 135.6, 139.0, 155.8, 135.0, 149.0, 133.0,
           135.3, 139.8, 137.2, 137.9, 136.8, 134.6, 152.3, 131.9, 137.2, 160.3,
           130.9, 130.9, 131.8, 131.4, 135.6, 138.6, 136.3)

speed <- c(10.01, 10.02, 10.39, 11.86,  9.65, 10.40, 12.60,  9.80, 11.52,  9.59,
           10.26, 10.42, 10.05, 10.48, 10.03, 10.29, 12.22, 10.27, 10.80, 13.79,
           10.40,  9.49, 10.09, 10.34, 11.18, 11.33, 11.34)
```

Initially, parameters in the following model are estimated

$$\text{speed}_i = \beta_0 + \beta_1 \cdot \text{week}_i + \epsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

Further it is assumed that the ε_i's are independent.

The result of the estimation in R is:

```
summary(lm(speed~week))

##
## Call:
## lm(formula = speed ~ week)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3909 -0.6058 -0.3407  0.2741  2.9492
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.36053    0.38514  26.901   <2e-16 ***
## week         0.04003    0.04046   0.989    0.332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.008 on 25 degrees of freedom
## Multiple R-squared:  0.03766,Adjusted R-squared:  -0.0008306
## F-statistic: 0.9784 on 1 and 25 DF,  p-value: 0.3321
```

Continues on page 26

Question VII.1 (18)

At the significance level $\alpha = 0.05$ what is the conclusion in relation to increased speed (both argument and conclusion must be correct)?

1 ☐ There is no significant effect of the training, since $0.0377 < 0.05$

2 ☐ There is a significant effect of the training, since $26.9 > t_{0.975}$ (where the degrees of freedom for the t-distribution is 25)

3* ☐ There is no significant effect of the training, since $0.332 > 0.05$

4 ☐ There is a significant effect of the training, since $0.04 < 0.05$

5 ☐ There is a significant effect of the training, since $0.989 > 0.95$

———— FACIT-BEGIN ————

The conclusion is that there is no significant effect. This can be based on the p-value for the hypothesis $\beta_1 = 0$ against the two-sided alternative (the p-value is $0.332 > 0.05$). Hence the possibilities are 1 or 3. In 1 the R^2 value is compared with the significance level (which does not make any sense). In 3 the p-value is correctly compared with the significance level. Hence the correct answer is 3. See Method 5.14 for details.

———— FACIT-END ————

Question VII.2 (19)

With the model above what is the 95% confidence interval for β_0?

1 ☐ $10.36 \pm 1.008 \cdot 1.96$

2 ☐ $10.36 \pm \frac{0.385}{\sqrt{25}} \cdot 2.06$

3 ☐ $0.040 \pm 0.0405 \cdot 1.96$

4 ☐ $0.040 \pm \frac{0.989}{25} \cdot 1.96$

5* ☐ $10.36 \pm 0.385 \cdot 2.06$

———— FACIT-BEGIN ————

Method 5.15 describes how to find confidence intervals for a parameter. The estimate of β_0 is 10.36 and the standard error is given directly as 0.385, further we need the 0.975 quantile of the t-distribution with 25 degrees of freedom, which is can be calculated in R:

```r
qt(0.975, 25)
```

```
## [1] 2.059539
```

Hence the correct answer is answer no. 5.

———— FACIT-END ————

Continues on page 28

The recreational runner now decides to investigate a model for the relationship between pulse and velocity (ignoring weeks).

The result of the estimation in R is:

```
summary(lm(speed ~ pulse))

##
## Call:
## lm(formula = speed ~ pulse)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78371 -0.40486 -0.00015  0.35978  0.96661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.06134    1.82173  -2.778   0.0102 *
## pulse        0.11324    0.01308   8.659 5.39e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5139 on 25 degrees of freedom
## Multiple R-squared:  0.7499,Adjusted R-squared:  0.7399
## F-statistic: 74.98 on 1 and 25 DF,  p-value: 5.388e-09
```

Based on the above model, the recreational runner wants an uncertainty interval for a new run. He uses as the assumption that he can hold an average pulse of 160 beats per minute.

As an aid to the task he has calculated the following number:

```
length(week)

## [1] 27

c(mean(week), var(week))

## [1]  8.222222 23.871795

c(mean(pulse), var(pulse))

## [1] 139.09259  59.38225

c(mean(speed), var(speed))

## [1] 10.689630  1.015396
```

Continues on page 29

Question VII.3 (20)

What is the 95% prediction interval for the speed?

1 ☐ $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51 \cdot \sqrt{\frac{1}{27} + \frac{(139.1-160)^2}{27 \cdot 59.38}}$

2 ☐ $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51 \cdot \sqrt{1 + \frac{1}{27} + \frac{(139.1-160)^2}{26 \cdot 1.02}}$

3* ☐ $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51 \cdot \sqrt{1 + \frac{1}{27} + \frac{(139.1-160)^2}{26 \cdot 59.38}}$

4 ☐ $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51^2 \cdot \sqrt{\frac{1}{27} + \frac{(139.1-160)^2}{59.38}}$

5 ☐ $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51 \cdot \sqrt{\frac{1}{27} + \frac{(139.1-160)^2}{59.38^2}}$

———— FACIT-BEGIN ————

The general formula for the prediction interval can be seen from Method 5.18 to be

$$\hat{\beta}_0 + \hat{\beta}_1 x_{new} + t_{0.975} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})}{S_{xx}}}$$

where the quantile of the t-distribution is based on $n-1$ degrees of freedom. In our concrete case we have $\hat{\beta}_0 = -5.06$, $\hat{\beta}_1 = 0.113$, $\hat{\sigma} = 0.51$. These number are taken from the summary of `lm`. With $n = 27$ we have $t_{0.975} = 2.06$. x_{new} is equal 160 in our case, $\bar{x}$ is the average of the measured pulses, i.e. 160. $S_{xx} = \sum_{i=1}^{27}(x_i - \bar{x})^2$ (See Theorem 5.4), which can be calculated by $S_{xx} = 26 s_x^2 = 26 \cdot 59.38$, hence the correct answer is no. 3.

———— FACIT-END ————

The recreational runner now decides to estimate a multiple regression model that contains both weeks and pulse:

$$\text{speed}_i = \beta_0 + \beta_1 \cdot \text{week}_i + \beta_2 \cdot \text{pulse}_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

The result from R is:

```
summary(lm(speed ~ week + pulse))

##
## Call:
## lm(formula = speed ~ week + pulse)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59154 -0.13508 -0.00055  0.15562  0.43438
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.033531   0.945936  -8.493 1.08e-08 ***
## week         0.094014   0.010414   9.028 3.48e-09 ***
## pulse        0.129052   0.006603  19.545 3.02e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2501 on 24 degrees of freedom
## Multiple R-squared:  0.9431,Adjusted R-squared:  0.9384
## F-statistic:   199 on 2 and 24 DF,  p-value: 1.148e-15
```

Continues on page 31

Question VII.4 (21)

At level $\alpha = 0.05$ which of the following statements is correct (both conclusion and argument must be correct)?

1 ☐ Neither the effect of weeks nor the effect of pulse is significant as $0.094 > 0.05$ and $0.129 > 0.05$

2 ☐ Both the effect of weeks and the effect of pulse is significant as $0.01 < 0.05$ and $0.0066 < 0.05$

3 ☐ Since $0.094 < 0.129$ the effect of weeks is significant, while the effect of pulse is not significant

4* ☐ Both the effect of weeks and the effect of pulse is significant since $3.5 \cdot 10^{-9} < 0.05$ and $3.0 \cdot 10^{-16} < 0.05$

5 ☐ Neither the effect of weeks nor the effect of pulse is significant since $3.5 \cdot 10^{-9} < 0.05$ and $3.0 \cdot 10^{-16} < 0.05$

———— FACIT-BEGIN ————

Let's go through the answers:

1 The number 0.094 and 0.129 are parameter estimates, and it does not make sense to compare those with the significance level hence answer no 1 is wrong.

2 In answer 2 standard errors are compared with significance levels, this is also incorrect.

3 In answer 3 the magnitude of the parameters are compared and this does not say anything about significance

4 In answer 4 the p-values of the parameters are correctly compared with the significance level, also the conclusion is correct (p-values < significance level), hence this is the correct answer

5 In answer 5 the p-values are correctly compared with the significance level, however the conclusion is wrong.

———— FACIT-END ————

Question VII.5 (22)

Which statement about the interpretation of the model is correct?

1 ☐ When weeks increase by 0.094 the pulse increase by 0.129

2* ☐ For a given pulse the expected speed increase with 0.094km/h per week.

3 ☐ The model is meaningless since $-8.03 < 0$ and the speed must be positive

4 ☐ Since the degree of explanation is about 0.25, the model have explained about 3/4 of the variation

5 ☐ Since all parameters are significant, it can be seen that all model assumptions are fulfilled

———— FACIT-BEGIN ————

Again we need to go through all the statements.

1 When weeks increase with 1 the expected speed increase by 0.094 (for a given pulse). The two explanatory variables do not say anything about how they affect each other, hence this answer is wrong.

2 It can be seen from the output that the parameter for week is 0.094, hence this means that the expected speed increases with 0.094 every week, hence this answer is correct.

3 The speed cannot be negative but the prediction of -8.03 is only obtained when week=0 and pulse = 0, and a prediction at zero pulse is of course meaningless, but not the model, hence answer no 3 is wrong.

4 About 1/4 of the variation is explained when the degree of explanation is 0.25, hence answer no 4 is wrong.

5 Significance of parameters does not tell anything about the model assumptions (we need diagnostics plots for this), hence answer no 5 is wrong

———— FACIT-END ————

Continues on page 33

Exercise VIII

A person has twice evaluated the sharpness (`Sharpness`) for each of 12 different setups (`Treat`) of images on computer screens, ie. 24 observations of sharpness in total split on 12 setups. The scale is a continuous scale from 0 to 15, in practice, done by marking the value on a line.

The result of the usual analysis of variance of these data gave the following R output, however, some of the values are replaced by the letters A, ..., F:

```
Analysis of Variance Table

Response: Sharpness
          Df Sum Sq Mean Sq F value Pr(>F)
Treat      A   93.7     C       E      F
Residuals  B   51.9     D
```

Question VIII.1 (23)

What are the values of A and B?

1* ☐ A=11 and B=12

2 ☐ A=1 and B=22

3 ☐ A=93.7/51.9 and B= 51.9/22

4 ☐ A=12 and B=24

5 ☐ A=11 and B=23

———— FACIT-BEGIN ————

We need to find the degrees of freedom of a one-way ANOVA as described in Theorem 8.6. As the number of treatments is $k = 12$ and the number of total observations is $n = 24$. Hence $A = k - 1 = 11$ and $B = n - k = 24 - 12 = 12$.

———— FACIT-END ————

Continues on page 34

Question VIII.2 (24)

Similarly, another person evaluated the sharpness (`Sharpness`) a number of times for each of different setups (Treat) of images on computer screens. The result of the usual analysis of variance of these data gave the following R output, however, some of the values are replaced by the letters G,. . . ,J:

```
Analysis of Variance Table

Response: Sharpness
          Df Sum Sq Mean Sq F value Pr(>F)
Treat      7  111.5     G       I      J
Residuals 32   88.4     H
```

What is the value of J?

1 ☐ 1.261

2 ☐ 5.766

3* ☐ 0.0002

4 ☐ 0.2188

5 ☐ 0.3002

———— FACIT-BEGIN ————

Since J is the p-value of a one-way ANOVA, we first need to find F_{obs}. As stated in 8.6 this can be found as:

$$F_{obs} = \frac{MS(Trt)}{MSE} = \frac{SS(Trt)/DF(Trt)}{MSE/DFE} = \frac{111.5/7}{88.4/32} = 5.766$$

And we can then find the p-value as

$$P(F > 5.766) = 1 - P(F \leq 5.766)$$

were F follows an F-distribution with with degrees of freedom $(7, 32)$:

```r
(Fobs <- (111.5/7)/(88.4/32))

## [1] 5.765999

1-pf(Fobs, 7, 32)

## [1] 0.00022284
```

———— FACIT-END ————

Continues on page 35

Exercise IX

Eight experts have each assessed the sharpness (`Sharpness`) for each of 12 different setups (`Treat`) of images on computer screens, ie. 96 observations of sharpness in total split on 12 setups. The scale is a continuous scale from 0 to 15, in practice, done by marking the value on a line:

	Setup 1	Setup 2	Setup 3	Setup 4	Setup 5	Setup 6	Setup 7	Setup 8	Setup 9	Setup 10	Setup 11	Setup 12
Person 1	9.30	4.70	6.60	8.80	5.90	7.20	7.60	5.50	8.10	8.20	6.40	7.40
Person 2	10.20	7.00	8.80	10.70	9.80	7.00	9.20	9.60	8.00	11.80	8.90	10.20
Person 3	11.50	9.50	8.00	12.90	10.00	8.20	11.50	6.40	8.60	11.20	7.70	11.00
Person 4	11.90	6.60	8.20	12.70	5.40	9.00	4.90	8.10	10.10	12.90	8.20	8.70
Person 5	10.70	4.20	5.40	11.40	8.30	7.10	6.80	3.80	9.60	8.60	3.80	10.80
Person 6	10.90	9.10	7.10	11.40	8.60	5.90	8.50	10.50	6.40	11.70	9.50	7.10
Person 7	8.50	5.00	6.30	10.80	6.80	4.60	4.70	8.80	6.70	10.00	5.50	7.50
Person 8	12.60	8.90	10.70	13.50	11.40	8.90	9.50	8.60	7.40	13.50	8.70	9.80

The result of a usual twoway analysis of variance of these data gave the following R-output:

```
## Analysis of Variance Table
##
## Response: Sharpness
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Person     7 122.42 17.4881  8.4596 1.212e-07 ***
## Setup     11 224.28 20.3894  9.8630 6.864e-11 ***
## Residuals 77 159.18  2.0673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question IX.1 (25)

What are the conclusions of the usual hypothesis tests for such an analysis? (Both conclusions and arguments must be correct)

1 ☐ There is a difference between the mean sharpness for persons, but not for setups

2 ☐ There is a difference between the mean sharpness for setups, but not for persons

3 ☐ There is a difference between the variances for persons, but not for setups

4* ☐ There is a difference between the mean sharpness for both setups and persons

5 ☐ There is a difference between the variances for setups, but not for persons

———— FACIT-BEGIN ————

We look at the p-values. Since both the p-values (`1.212e-07` and `6.864e-11` respectively) are below $\alpha = 0.05$, we can reject the null-hypotheses of no difference and conclude that both setup and person make a difference.

———— FACIT-END ————

Continues on page 38

Question IX.2 (26)

Which probability distribution has been used to find the p-value provided for `Setup` in the output?

1 ☐ The z-distribution (= standard normal distribution)

2 ☐ The t-distribution with 159 degrees of freedom

3 ☐ The χ^2-distribution with 159 degrees of freedom

4 ☐ The F-distribution with degrees of freedom 7 and 11

5* ☐ The F-distribution with degrees of freedom 11 and 77

——— FACIT-BEGIN ———

See Theorem 8.22. Since we are trying to find the degrees of freedom associated to the `setup` they are $df1 = l - 1 = 12 - 1 = 11$ and $df2 = (8 - 1)(12 - 1) = 77$.

——— FACIT-END ———

Continues on page 39

Exercise X

In a questionnaire survey under an Introduction to Statistics lecture the participants were asked about different topics. This assignment will cover the analysis of the answers to one of the questions. There were 32 respondents in total.

The question asked was: "Are you worried that we don't do enough to stop climate change?". To this the students answered the following:

Answer	Count
Yes	27
No	5

Question X.1 (27)

Using the "Plus 2" correction when calculating the usual 95% confidence interval for the proportion of students who are worried about the climate (answering yes), one gets:

1 ☐ $0.844 \pm 2.04 \cdot \frac{0.844}{36} = [0.796, 0.892]$

2 ☐ $0.806 \pm 1.69 \cdot \frac{0.806}{36} = [0.768, 0.844]$

3* ☐ $0.806 \pm 1.96\sqrt{\frac{0.806 \cdot 0.194}{36}} = [0.677, 0.935]$

4 ☐ $0.844 \pm 1.69\sqrt{\frac{0.844 \cdot 0.156}{32}} = [0.736, 0.952]$

5 ☐ $0.844 \pm 1.96\sqrt{\frac{0.844 \cdot 0.156}{36}} = [0.725, 0.963]$

——— FACIT-BEGIN ———

The usual formula for confidence interval shown in Method 7.3 requires a large sample size (i.e. high n), however this is not a large sample, so we can use the "plus 2" correction, see Remark 7.7. This is carried out by adding 2 to the count (x) and adding 4 to the number of observations (n), and thereafter using the formulas. So

$$\hat{p}_{\text{plus2}} = \frac{x+2}{n+2} = \frac{29}{36}$$

and then $z_{1-\alpha/2}$ is found in R by `qnorm(0.975)`, and finally

$$\begin{aligned} \hat{p}_{\text{plus2}} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_{\text{plus2}}(1-\hat{p}_{\text{plus2}})}{n+4}} &= 0.806 \pm 1.96\sqrt{\frac{0.806 \cdot 0.194}{36}} \\ &= [0.677, 0.935]. \end{aligned}$$

——— FACIT-END ———

Continues on page 40

Exercise XI

In a questionnaire survey 114 respondents were asked about their traffic preferences. Two questions were asked, which had the following three identical answer options: "Car", "Bike" and "Train or bus".

The two questions were: "If you have 10 km to DTU from your home, what kind of transportation would you prefer during the summer (they take about equal time)?", and: "if you have 10 km to DTU from your home, which kind of transportation would you prefer during the winter (they take about equal time)?".

The following distribution of answers were observed:

		Winter		
		Car	Bike	Train or bus
Summer	Car	27	2	4
	Bike	20	22	11
	Train or bus	13	3	12

The following is run in R:

```
## The data table
tbl <- matrix(c(27, 20, 13, 2, 22, 3, 4, 11, 12), nrow = 3)
rownames(tbl) <- c("Car", "Bike", "Trainorbus")
colnames(tbl) <- c("Car", "Bike", "Trainorbus")
tbl

##            Car Bike Trainorbus
## Car         27    2          4
## Bike        20   22         11
## Trainorbus  13    3         12

## Row sums (distribution for summer)
margin.table(tbl, 1)

##        Car       Bike Trainorbus
##         33         53         28

## Column sums (distribution for winter)
margin.table(tbl, 2)

##        Car       Bike Trainorbus
##         60         27         27

## Chi2-test
chisq.test(tbl, correct = FALSE)

##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 27.608, df = 4, p-value = 1.498e-05
```

Continues on page 41

Question XI.1 (28)

What is the expected count of preferences for: bike in the summer and car in the winter, under the null hypothesis: there is independence between traffic preferences in summer and winter in the surveyed population?

1 ☐ $e_{11} = 114 \cdot \frac{33}{114} \cdot \frac{60}{114} = 17.37$

2 ☐ $e_{23} = 114 \cdot \frac{53}{114} \cdot \frac{27}{114} = 12.55$

3 ☐ $e_{12} = 114 \cdot \frac{33}{114} \cdot \frac{27}{114} = 7.816$

4 ☐ $e_{33} = 114 \cdot \frac{28}{114} \cdot \frac{27}{114} = 6.877$

5* ☐ $e_{21} = 114 \cdot \frac{53}{114} \cdot \frac{60}{114} = 27.89$

———— FACIT-BEGIN ————

Under the null hypothesis there is independence between the preferences, which essentially means that the distribution of the counts in each row is the same as the distribution of counts in the row sums, and the same for the columns. So it means that the distribution in one variable is not changed as a function of the other variable. This is the null hypothesis in Theorem 7.24 and it is tested using the χ^2-test.

We take the two total proportions (in the row and in the column), which is our best estimates if H_0 is true, and multiply with the total number of observations

$$e_{21} = 114 \cdot \frac{53}{114} \cdot \frac{60}{114} = 27.89.$$

———— FACIT-END ————

Question XI.2 (29)

What is the conclusion at significance level 1% of the test for independence of traffic preferences in summer and winter (both conclusion and argument must be correct)?

1 ☐ No significant dependence between traffic preferences is found since the p-value > 0.01

2* ☐ A significant dependence between traffic preferences is found since the p-value < 0.01

3 ☐ No significant dependence between traffic preferences is found since the p-value < 0.01

4 ☐ A significant dependence between traffic preferences is found since the p-value > 0.01

5 ☐ The question cannot be answered with the given information

———— FACIT-BEGIN ————

The p-value of the χ^2-test for independence is calculated in the last line in the R code and we simply read off the p-value to $1.5 \cdot 10^{-5}$, hence much smaller than $\alpha = 0.01$. Thus a significant dependence between traffic preferences is found since p-value < 0.01

———— FACIT-END ————

Continues on page 43

Question XI.3 (30)

There are 60 out of 114 who prefer to drive car in the winter. Would the following null hypothesis

$$H_0 : p_{\text{car,winter}} = 50\%,$$

be rejected at the 5% significance level with the usual test (both conclusion and the p-value must be correct)?

1 ☐ Yes, since the p-value is $0.024 < 0.05$

2* ☐ No, since the p-value is $0.57 > 0.05$

3 ☐ No, since the p-value is $0.40 > 0.05$

4 ☐ Yes, since the p-value is $0.089 > 0.05$

5 ☐ No, since the p-value is $0.21 > 0.05$

———— FACIT-BEGIN ————

We use Method 7.11, since we are dealing with a single proportion. This is easiest calculated using R by:

```
prop.test(60, 114, p=0.5, correct=FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  60 out of 114, null probability 0.5
## X-squared = 0.31579, df = 1, p-value = 0.5741
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4353037 0.6156122
## sample estimates:
##         p
## 0.5263158
```

———— FACIT-END ————

THE EXAM IS FINISHED. ENJOY THE SUMMER!

Technical University of Denmark *Page 1 of 31 pages.*

Written examination: 16. December 2018

Course name and number: **Introduction to Statistics (02323)**

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

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

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

The exam paper contains 31 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

In the analysis of a single sample, 10 measurements are assumed to be independent and sampled from a normal distribution with mean μ and variance σ^2. The sample mean is $\bar{x} = 0.57$, while the sample standard deviation is $s = 0.32$.

Question I.1 (1)

Which of the following is a standard 99% confidence interval for the theoretical standard deviation σ?

1* ☐ $[0.20, 0.73]$

2 ☐ $0.57 \pm 1.96 \cdot 0.32$

3 ☐ $[0.22, 0.58]$

4 ☐ $[0.05, 0.34]$

5 ☐ $[0.03, 0.53]$

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

See Method 3.19. Here, $n = 10$ and $\alpha = 0.01$, so the left and right endpoints are, respectively,

```
sqrt( (10-1)*0.32^2/qchisq(0.995, df = 9) )
```

```
## [1] 0.1976575
```

```
sqrt( (10-1)*0.32^2/qchisq(0.005, df = 9) )
```

```
## [1] 0.7288361
```

which result in the interval $[0.20, 0.73]$ when rounded to two decimals.

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

Exercise II

We would like to determine the median of X_1/X_2, when X_1 and X_2 are independent stochastic variables, which are both normal distributed with mean 1 and variance 1. The distribution of the ratio is not trivial; therefore we resort to simulation to determine an estimate and a confidence interval for the median of the distribution of X_1/X_2.

Question II.1 (2)

First, 10000 medians are simulated, each being the median of 10000 ratios. We store these in R in the vector `medians`:

```
ratio <- replicate(10000, rnorm(10000, mean = 1)/rnorm(10000, mean = 1))
medians <- apply(ratio, 2, median)
```

Subsequently, the sample mean and a series of percentiles are calculated for these 10000 medians:

```
mean(medians)

## [1] 0.6193

quantile(medians, c(0.005, 0.025, 0.05, 0.5, 0.95, 0.975, 0.995), type = 2)

##   0.5%   2.5%     5%    50%    95%  97.5%  99.5%
## 0.5873 0.5949 0.5989 0.6193 0.6402 0.6443 0.6515
```

Which of the following choices yields an estimate for the median of X_1/X_2 and a 95% confidence interval for this median?

1 ☐ Estimate: 1.
95% confidence interval: $[1 - 1.96 \cdot 0.6193, 1 + 1.96 \cdot 0.6193]$.

2* ☐ Estimate: 0.6193.
95% confidence interval: $[0.5949, 0.6443]$.

3 ☐ Estimate: 1.
95% confidence interval: $[1 - 0.5949, 1 + 0.6443]$.

4 ☐ Estimate: 0.6193.
95% confidence interval: $[0.5873, 0.6515]$.

5 ☐ Estimate: 0.6193.
95% confidence interval: $[0.6193 - 0.5949, 0.6193 + 0.5949]$.

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

The estimate is the average of the simulated medians, i.e. the estimated median is 0.6193. The left and right endpoints of the 95% confidence interval are, respectively, the 2.5% and 97.5% quantiles of the simulated medians, so the confidence interval becomes $[0.5949, 0.6443]$.

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

Exercise III

A normal distributed population has mean $\mu = 100$ and standard deviation $\sigma = 15$.

Question III.1 (3)

In a random draw, what is the probability of obtaining an observation below 90?

1* ☐ 0.252

2 ☐ 0.482

3 ☐ 0.518

4 ☐ 0.631

5 ☐ 0.748

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

Let $X \sim N(100, 15^2)$. Then, we may find $P(X < 90) = P(X \leq 90)$ as:

```
pnorm(90, mean = 100, sd = 15)

## [1] 0.2524925
```

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

Question III.2 (4)

If a random sample of $n = 10$ independent observations is drawn from the population, what is the probability that the sample mean is below 90?

1 ☐ 0.000783

2* ☐ 0.0175

3 ☐ 0.146

4 ☐ 0.252

5 ☐ 0.482

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

Let $\bar{X}$ denote the sample mean, i.e. the average of 10 independent random variables $X_1, \ldots, X_{10}$, each with the same distribution as X. Use the mean and variance identities for linear combinations of independent random variables (Theorem 2.54) to compute the mean

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{10}\sum_{i=1}^{10} X_i\right) = \frac{1}{10}\sum_{i=1}^{10}\mathrm{E}(X_i) = \frac{1}{10}\cdot 10\cdot\mu = \mu = 100$$

and the variance

$$\mathrm{V}(\bar{X}) = \mathrm{V}\left(\frac{1}{10}\sum_{i=1}^{10} X_i\right) = \frac{1}{10^2}\sum_{i=1}^{10}\mathrm{V}(X_i) = \frac{1}{10^2}\cdot 10\cdot\sigma^2 = \frac{1}{10}15^2 = 22.5\,.$$

Now, $P(\bar{X} < 90) = P(\bar{X} \leq 90)$ may be found as

```
pnorm(90, mean = 100, sd = sqrt(22.5))
```

```
## [1] 0.01750749
```

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

Question III.3 (5)

Suppose that a random sample of n independent observations is repeatedly drawn from the population, and that the sample variance S^2 is calculated in each repetition. What holds true for S^2?

1 ☐ n^2S^2 is F-distributed with $n-1$ and $n-2$ degrees of freedom.

2 ☐ S^2 is χ^2-distributed with $n-1$ degrees of freedom.

3* ☐ $(n-1)S^2/\sigma^2$ is χ^2-distributed with $n-1$ degrees of freedom.

4 ☐ S^2 is normal distributed with mean μ and variance σ^2/n^2.

5 ☐ S^2 has the same distribution as $(Z-\sigma^2)/n$, where Z is standard normal distributed.

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

See Section 3.1.6 from beginning and after one page you find the result, that the sampling distribution of the sample variance transformed by multiplying with $(n-1)$ and dividing with σ^2 is χ^2-distributed with $n-1$ degrees of freedom. It is stated in Equation 3-17.

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

Exercise IV

10 people have had their daily energy intake measured (in kJ). The measurements in the sample are shown in the table below:

Energy intake (kJ):	8230	5470	7515	5260	6390	6180	6515	6805	7515	5640

Question IV.1 (6)

What is the median of the sample?

1 ☐ 6390

2 ☐ 6515

3 ☐ (8230+5260)/2

4 ☐ (6390+6180)/2

5* ☐ (6390+6515)/2

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

You can rather easily copy from the pdf into an R script and add some commas between the values, and then:

```
# Read data into R and sort:
x <- sort(c(8230, 5470, 7515, 5260, 6390, 6180, 6515, 6805, 7515, 5640))
x

##  [1] 5260 5470 5640 6180 6390 6515 6805 7515 7515 8230
```

As there are 10 observations, the median is computed as the mean of observations number 5 and 6 after sorting:

$$\frac{(6390+6515)}{2} = 6452.5$$

The result may be verified using R:

```
median(x)

## [1] 6452.5
```

or:

```
quantile(x, prob=0.5, type=2)

##    50%
## 6452.5
```

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

Question IV.2 (7)

The sample mean is $\bar{x} = 6552$, while the sample standard deviation is $s = 975.94$. It is assumed that the daily energy intake may be modelled by a normal distribution, and that the observations are independent and identically distributed. What is the p-value for the t-test that tests the hypothesis that the mean daily energy intake is 7725 kJ?

1 ☐ 0.4

2 ☐ 0.06

3 ☐ 0.04

4 ☐ 0.006

5* ☐ 0.004

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

See Method 3.23. With $\bar{x} = 6552$, $s = 975.94$ and $n = 10$, the observed t-test statistic is calculated as

$$t_{\text{obs}} = \frac{6552 - 7725}{975.94/\sqrt{10}} = -3.8007989$$

and the p-value is thus found as

$$2P(T \leq t_{\text{obs}}) = 2 \cdot 0.0021061 = 0.004\,.$$

$P(T \leq t_{\text{obs}})$ is found in R as:

```r
pt(-3.8007989, 10-1)

## [1] 0.002106097
```

The in-built function could also be used:

```r
t.test(x, mu=7725)

##
##  One Sample t-test
##
## data:  x
## t = -3.8008, df = 9, p-value = 0.004212
## alternative hypothesis: true mean is not equal to 7725
## 95 percent confidence interval:
##  5853.853 7250.147
## sample estimates:
## mean of x
##      6552
```

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

Exercise V

A married couple visits the same restaurant several times a month. Typically, they order a glass of red wine with their food. One day, they decide to complain to the owner. They believe that one of the waiters pours less wine into the glass than what they pay for. Consequently, the owner launches an experiment with three of the restaurant's waiters in order to investigate how much they pour into wine glasses, when they pour using a rule of thumb. Each of the three waiters (here anonymized by A, B, and C) were asked to pour red wine into 20 wine glasses, after which the content in each glass was measured. The data were read into R in two variables: `waiter`, indicating which waiter poured the wine, and `wine`, indicating the amount of wine in the glass (in mL).

The following code was run in R to analyze the data:

```r
anova(lm(wine ~ waiter))

## Analysis of Variance Table
##
## Response: wine
##           Df Sum Sq Mean Sq F value   Pr(>F)
## waiter     2 1043.4  521.71  6.9594 0.001976 **
```

```
## Residuals 57 4273.0   74.97
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question V.1 (8)

What may be concluded from the R output above, when a significance level of 5% is used (both the reasoning and conclusion must be correct)?

1 ☐ As the observed F-test statistic is larger than the 0.95 quantile of the $F(57, 2)$-distribution, there is a significant difference in the expected amount of wine in glasses poured by the three different waiters.

2 ☐ As the p-value is larger than 5%, there is no significant difference in the expected amount of wine in glasses poured by the three different waiters.

3 ☐ As the sum of squared errors, *SSE*, is more than four times the size of the treatment sum of squares, *SS(Tr)*, there is too much noise in the data for it to be meaningful to perform a one-way analysis of variance.

4* ☐ As the observed F-test statistic is larger than the 0.95 quantile of the $F(2, 57)$-distribution, there is a significant difference in the expected amount of wine in glasses poured by the three different waiters.

5 ☐ As the p-value is less than 5%, there is no significant difference in the expected amount of wine in glasses poured by the three different waiters.

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

See Theorem 8.6. From the R-output, it is seen that the relevant F-test statistic is $F_{\text{obs}} = 6.9594$. The 0.95 quantile of the $F(2, 57)$ distribution may be found using R:

```
qf(0.95, df1 = 2, df2 = 57)

## [1] 3.158843
```

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

Question V.2 (9)

Among other things, the owner would like to make a comparison between waiter A (the young waiter, whom the couple complained about) and waiter B (an older waiter with many years of experience in the business). On average, waiter A poured 127 mL of wine into each glass, while

waiter B poured 135 mL. Compute the t-test statistic for the post hoc pairwise hypothesis test which compares the expected amount of wine in glasses poured by waiter A and waiter B.

1 ☐ $t_{obs} = -0.92$

2 ☐ $t_{obs} = -4.13$

3* ☐ $t_{obs} = -2.92$

4 ☐ $t_{obs} = -1.07$

5 ☐ $t_{obs} = -0.11$

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

See Method 8.10. The relevant post hoc t-test statistic is computed as follows:

$$t_{\text{obs}} = \frac{(127 - 135)}{\sqrt{74.97\left(\frac{1}{20} + \frac{1}{20}\right)}} = -2.92$$

So its like the two-sample t.test, except that the estimate of the error variance is taken from the model fitted to all the data $\hat{\sigma}^2 = MSE = \frac{SSE}{n-k}$, i.e. the pooled variance estimate.

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

Question V.3 (10)

In addition to the information in the previous question, it is given that, on average, waiter C poured 136 mL into each glass. Compute the Bonferroni corrected LSD ("least significant difference") value used to perform all possible pairwise comparisons between the three waiters, and determine where there are significant differences (both the LSD value and the conclusion must be correct). Use the significance level $\alpha = 5\%$.

1 ☐ $LSD_{0.05/3} = 7$ mL, so there is a significant difference between the expected amount of wine in glasses poured by waiters B and C, but no significant difference between waiters A and B or between waiters A and C.

2* ☐ $LSD_{0.05/3} = 7$ mL, so there is a significant difference between the expected amount of wine in glasses poured by waiters A and B as well as between waiters A and C, but no significant difference between waiters B and C.

3 ☐ $LSD_{0.05/3} = 4$ mL, so there is a significant difference between the expected amount of wine in glasses poured by waiters A and B and between waiters A and C, but no significant difference between waiters B and C.

4 ☐ $LSD_{0.05/3} = 17$ mL, so there is a significant difference between the expected amount of wine in glasses poured by waiters A and B and between waiters A and C, but no significant difference between waiters B and C.

5 ☐ $LSD_{0.05/3} = 4$ mL, so there is a significant difference between the expected amount of wine in glasses poured by waiters B and C, but no significant difference between waiters A and B or between waiters A and C.

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

See Remark 8.13.

$$LSD_{0.05/3} = t_{1-(0.05/3)/2}\sqrt{2 \cdot MSE/20} = 2.466687 \cdot \sqrt{2 \cdot 74.97/20} = 6.8\,,$$

where $t_{1-(0.05/3)/2} = t_{5.95/6}$ is the $5.95/6 = 0.9916667$ quantile of the t-distribution with $60-3 = 57$ degrees of freedom, found in R as follows:

```
qt(5.95/6, df = 60-3)

## [1] 2.466687
```

So we can use that to determine which of the three waiters will be tested significantly different in two-sample post hoc comparisons. We have information about the average for each waiter:

$$\bar{x}_{\mathrm{A}} = 127 \text{ mL}$$
$$\bar{x}_{\mathrm{B}} = 135 \text{ mL}$$
$$\bar{x}_{\mathrm{C}} = 136 \text{ mL}$$

from which we can see that A is significantly different from B and C, since their differences are higher than 7 mL, and that there is no significant difference between B and C.

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

Exercise VI

A spring is characterized by its spring constant, k. When a spring is stretched, Hooke's law states that

$$F = -k \cdot x\,,$$

where x is the length (in meters) by which the spring is extended, and F is the applied force (in Newtons). The following six observations were made for a given spring:

	1	2	3	4	5	6
x	0.22	0.24	0.26	0.28	0.30	0.32
F	-0.51	-0.85	-0.89	-1.59	-1.97	-2.06

The observations were read into two vectors in R, `x` (length) and `F` (force), respectively, after which the following model was estimated:

```
model1 <- lm(F ~ x)
```

The output from `summary(model1)` is shown below, where some numbers are replaced by letters:

```
##
## Call:
## lm(formula = F ~ x)
##
## Residuals:
##        1        2        3        4        5        6
## -0.04484 -0.04146  0.25365 -0.10667 -0.15758  0.09690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2433     0.5483     A       C    **
## x           -16.8663     2.0148     B       D    **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1686 on 4 degrees of freedom
## Multiple R-squared:  0.946,Adjusted R-squared:  0.9325
## F-statistic: 70.08 on 1 and 4 DF,  p-value: 0.001114
```

Question VI.1 (11)

How may the statistical model corresponding to `model1` be described?

1 ☐ $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where Y_i represents the length by which the spring is extended when the force x_i is applied, and $\varepsilon_1, \ldots, \varepsilon_6$ are assumed to be independent and identically $N(0, \sigma^2)$-distributed.

2 ☐ $Y_i = \beta_1 x_i + \varepsilon_i$, where Y_i represents the force used to extend the spring by the length x_i, and $\varepsilon_1, \ldots, \varepsilon_6$ are assumed to be independent and identically $N(0, 1)$-distributed.

3 ☐ $Y_i = \beta_1 x_i + \varepsilon_i$, where Y_i represents the length by which the spring is extended when the force x_i is applied, and $\varepsilon_1, \ldots, \varepsilon_6$ are assumed to be independent and identically $N(0, 1)$-distributed.

4 ☐ $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where Y_i represents the length by which the spring is extended when the force x_i is applied, and $\varepsilon_1, \ldots, \varepsilon_6$ are assumed to be independent and identically $N(0, 1)$-distributed.

5* ☐ $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where Y_i represents the force used to extend the spring by the length x_i, and $\varepsilon_1, \ldots, \varepsilon_6$ are assumed to be independent and identically $N(0, \sigma^2)$-distributed.

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

See Chapter 5. The R-code `lm(F ~ x)` fits a simple linear regression model, in which `F` (The force) is the dependent variable and `x` (distance) is the explanatory variable. The model is defined in Equation 5-16 and some more information is given in Remark 5.6.

There are $n = 6$ observations in the sample, hence since $i = 1, \ldots, n$, there are: six stochastic variables Y_i, six variables x_i and six stochastic variables ε_i. The i.i.d. assumption is that the six errors:

- all come from the same population, which is normal distributed $N(0, \sigma^2)$
- are drawn independently from the population

The assumption of independence of the errors is actually not trivial! It can be summed up in that the conditions, which leads to "unmodelled" variance in Y_i, must be varied randomly. As an example think of: if another variable (e.g. temperature) actually affected the dependent variables Y_i, and this variable is not measured and thus not included in the model, then the experiment should actually be carried out such that this variable is varied randomly. If not then the sample will be biased and eventually (some of) the conclusions drawn can be affected (estimates, p-values, ...). This basically means, that one should be very careful when designing experiments and making sure that the studied phenomena is not affected by some unmeasured non-random conditions during the experiment...

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

Question VI.2 (12)

Based on the estimated slope in `model1`, give an estimate of the spring constant, k:

1 ☐ 0.5483

2 ☐ 3.2433

3 ☐ 2.0148

4* ☐ 16.8663

5 ☐ 5.2004

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

According to Hooke's law given above, the spring constant corresponds to the estimated slope, but with the opposite sign, i.e. $\hat{k} = -\hat{\beta}_1$.

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

Question VI.3 (13)

It is of interest to test whether the model's intercept is significantly different from zero. Give the relevant test statistic:

1 ☐ -8.371

2* ☐ 5.915

3 ☐ 0.004

4 ☐ 0.548

5 ☐ 0.169

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

The null hypothesis is $H_0 : \beta_0 = 0$, so $\beta_{0,0} = 0$ in Theorem 5.12. Using the R output (and standard notation from the book), the observed t-test statistic may be computed as

$$t_{\text{obs}} = \frac{\hat{\beta}_0}{\hat{\sigma}_{\beta_0}} = \frac{3.2433}{0.5483} = 5.915.$$

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

Question VI.4 (14)

What is the distribution of the test statistic used to test whether the model's slope can be assumed to be zero?

1 ☐ A t-distribution with 6 degrees of freedom.

2 ☐ A standard normal distribution.

3 ☐ An F-distribution with 6 degrees of freedom.

4 ☐ A normal distribution with mean zero and standard deviation 0.1686.

5* ☐ A t-distribution with 4 degrees of freedom.

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

See again Theorem 5.12. Here $n = 6$, so degrees of freedom is $n - 2 = 4$.

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

Question VI.5 (15)

In a simple linear regression like the above, the estimators of the intercept and slope parameters are often correlated. When is this correlation zero?

1 ☐ When the standard deviation of the dependent variable is 1.

2 ☐ When the slope is estimated as zero.

3* ☐ When the average of the explanatory variable is zero.

4 ☐ When the standard deviation of the explanatory variable is 1.

5 ☐ When the average of the dependent variable is zero.

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

According to Theorem 5.8 Equation 5-29, the covariance, and hence the correlation, between the intercept and the slope is zero if the sample mean of the explanatory variable, $\bar{x}$, is zero.

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

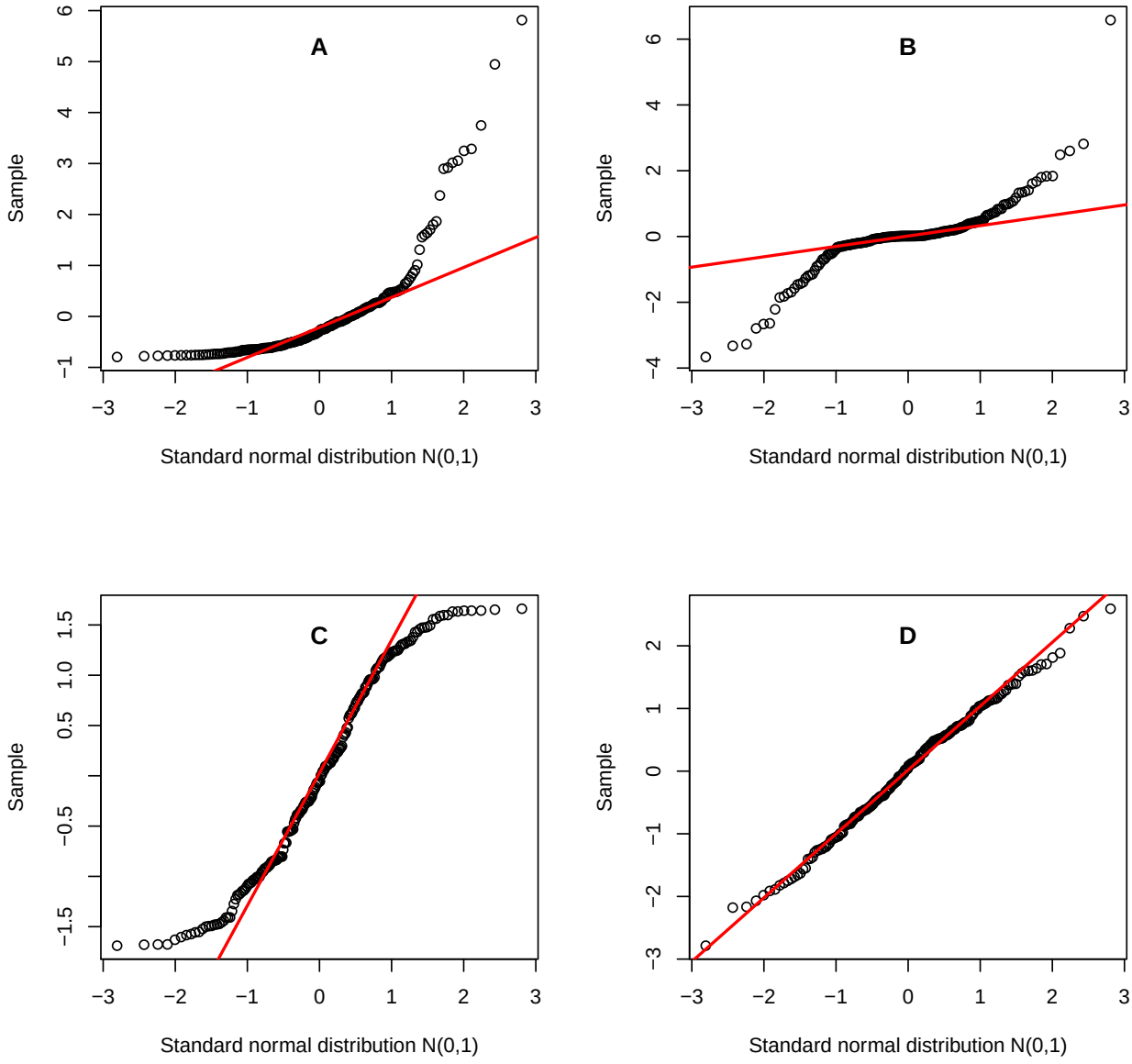
Exercise VII

In order to investigate whether data from a single sample is log-normal distributed, one could compare the data to a normal distribution using a qq-plot. If the data is log-normal distributed there will (typically) be fewer small values and more large values in the data, compared to a normal distribution with the same mean and variance as the sample.

Question VII.1 (16)

Below, four qq-plots are shown in which four different samples with mean 0 and variance 1 are each compared to a standard normal distribution. Let $z_{0.25}$ and $z_{0.75}$ denote the first and third quartile of the standard normal distribution, respectively, while $q_{0.25}$ and $q_{0.75}$ denote the first

and third quartile of the sample. The red line is drawn through the points $(z_{0.25}, q_{0.25})$ and $(z_{0.75}, q_{0.75})$. Which sample fulfills the above description of log-normal distributed data?



1* ☐ A

2 ☐ B

3 ☐ C

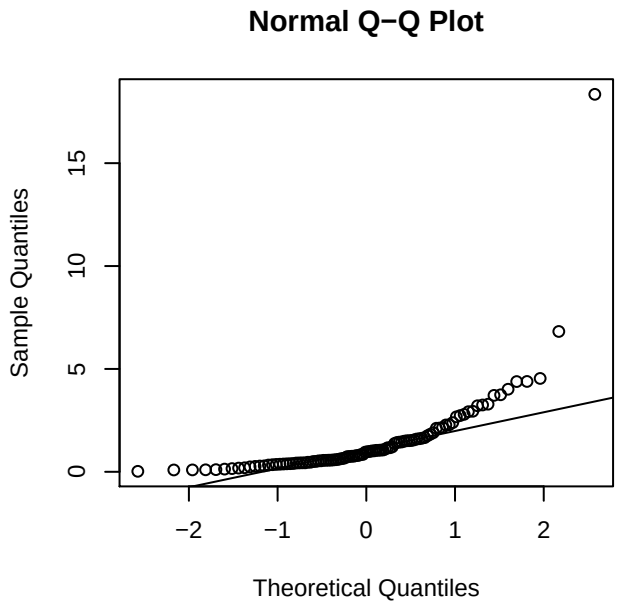
4 ☐ D

5 ☐ None of the above.

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

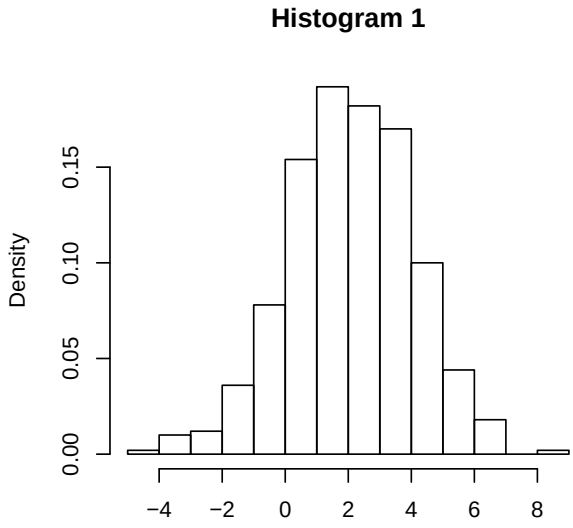
The answer is A. In B, the sample has more small values as well as more large values. The sample in C has fewer small values and fewer large values. The sample in D seems to be normal distributed. Verify the shape of a qq-plot of a log-normal distribution vs. a normal distribution in R:

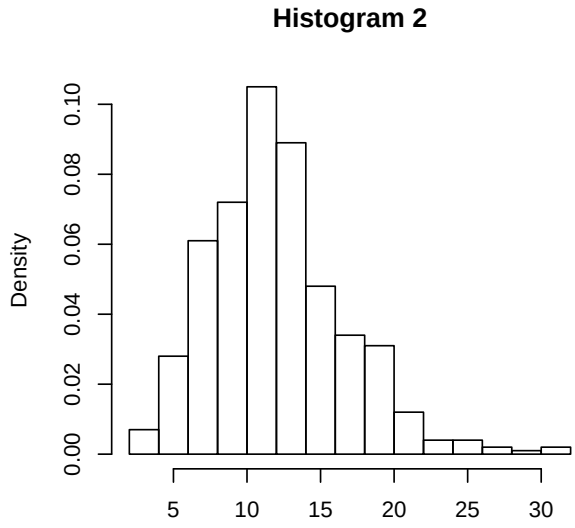
```
x <- rlnorm(100)
qqnorm(x)
qqline(x)
```

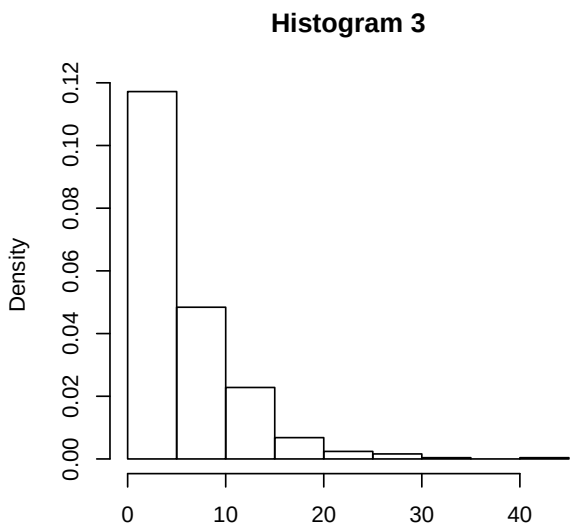


---------- FACIT-END ----------

Exercise VIII







Question VIII.1 (17)

Which distributions are simulated above? ($N(\mu, \sigma^2)$ refers to the normal distribution with mean μ and variance σ^2, χ^2_a to the χ^2 distribution with a degrees of freedom, and $Exp(\beta)$ to the exponential distribution with rate β).

1 ☐ 1: $N(0,4)$, 2: χ^2_{10}, 3: $Exp(1/5)$

2 ☐ 1: χ^2_4, 2: $N(2,4)$, 3: χ^2_1.

3* ☐ 1: $N(2,4)$, 2: χ^2_{12}, 3: $Exp(1/5)$

4 ☐ 1: $N(2,4)$, 2: $Exp(5)$, 3: χ^2_1

5 ☐ 1: $N(2,4)$, 2: χ^2_1, 3: $Exp(1/5)$

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

The distribution in Histogram 1 takes negative values (which the χ^2 distribution doesn't), and it seems symmetric around 2, so based on the available choices, it can only be the $N(2,4)$ distribution. The χ^2_1 distribution has mean 1, and the $Exp(5)$ distribution has mean $1/5$. Thus, based on Histogram 2 (where there isn't even any values in the data which are as small as 1), option 3 is the only possible choice. (This may be further verified by considering the means of some of the other distributions as well).

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

Exercise IX

Two groups of rats are put on a diet while growing up, and their weight gain between day 28 and day 84 is recorded. 10 rats are put on a diet with a high protein content, while 7 rats are put on a diet with a low protein content. The collected data (weight gain in grams) is shown in the table below, with the total weight gain in each group given in the last row:

	High protein content	Low protein content
	134	70
	146	118
	104	101
	119	85
	124	107
	161	132
	107	94
	83	
	113	
	129	
Total	1220	707

Using the numbers in the table, the sample variances in the two groups are calculated to be $s^2_H = 495$ and $s^2_L = 425$, where H and L indicate the groups with high and low protein content, respectively. It is further given that the usual test, for whether the expected weight gain is the same for rats on high and low protein diets, has 13.7 degrees of freedom.

Question IX.1 (18)

Which of the following choices is correct (both statements need to be correct)?

1 ☐ Rats in the low protein diet group gain more weight than rats in the high protein diet group. However, the difference is not statistically significant at the significance level $\alpha = 0.05$.

2 ☐ Rats in the high protein diet group gain more weight than rats in the low protein diet group. The difference is statistically significant at the significance level $\alpha = 0.05$.

3 ☐ Rats in the high protein diet group gain more weight than rats in the low protein diet group. The difference is statistically significant at the significance level $\alpha = 0.01$.

4 ☐ Rats in the low protein diet group gain more weight than rats in the high protein diet group. The difference is statistically significant at the significance level $\alpha = 0.05$.

5* ☐ Rats in the high protein diet group gain more weight than rats in the low protein diet group. However, the difference is not statistically significant at the significance level $\alpha = 0.05$.

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

The estimated difference in expected weight increase is

$$\frac{1220}{10} - \frac{707}{7} = 21\,,$$

that is, the increase is larger in the high protein group. However, the observed t-test statistic is

$$t_{\text{obs}} = \frac{21}{\sqrt{495/10 + 425/7}} = 2.000324$$

and as the 0.975 percentile of the t-distribution with 13.7 degrees of freedom is

```
qt(0.975, df = 13.7)

## [1] 2.149201
```

we conclude that the difference is not significant. This could also be concluded by writing in the values in R and using the `t.test()` function.

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

Exercise X

Statistics Denmark provides data related to Denmark at `www.statistikbanken.dk`, among it data on traffic accidents. The following count data is taken from there:

Year	2010				2017			
Type	All		Alcohol		All		Alcohol	
Zone	City	Rural	City	Rural	City	Rural	City	Rural
Single-vehicle accidents	240	491	107	178	174	340	55	96
Others	1779	988	161	84	1456	819	106	48

Values under "all" count all accidents (including drunk-driving accidents) while numbers under "alcohol" include only drunk-driving accidents.

Question X.1 (19)

Give a 99% confidence interval for the total proportion of drunk driving accidents in 2010, where you use the relevant normal distribution approximation.

1 ☐ $0.848 \pm 2.58\sqrt{\frac{0.848}{3498}}$

2 ☐ $0.152 \pm 2.58\sqrt{\frac{0.848}{3498}}$

3 ☐ $0.848 \pm 1.96\sqrt{\frac{0.152 \cdot 0.848}{3498}}$

4 ☐ $0.848 \pm 2.58\sqrt{\frac{0.152}{3498}}$

5* ☐ $0.152 \pm 2.58\sqrt{\frac{0.152 \cdot 0.848}{3498}}$

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

See Method 7.3. Here, $x = 107 + 178 + 161 + 84 = 530$ and $n = 240 + 491 + 1779 + 988 = 3498$, so

$$\hat{p} = \frac{530}{3498} = 0.152\,, \qquad 1 - \hat{p} = 0.848\,,$$

and $z_{0.995} = 2.58$ is the 0.995 quantile of the standard normal distribution.

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

Question X.2 (20)

Assume that the proportion of drunk-driving accidents in the "single-vehicle accidents" category is representative of the total proportion of drunk-driving. (Thus, data from the "others" category should *not* be used in this question).

Then, using the numbers from the table above and the wording from Table 3.1 of the book, what may be concluded about the difference in drunk driving between the years 2010 and 2017?

1* ☐ There is very strong evidence of a decrease in the proportion of drunk-driving accidents.

2 ☐ There is weak evidence of a decrease in the proportion of drunk-driving accidents.

3 ☐ There is little or no evidence of a difference in the proportion of drunk-driving accidents.

4 ☐ There is weak evidence of an increase in the proportion of drunk-driving accidents.

5 ☐ There is very strong evidence of an increase in the proportion of drunk-driving accidents.

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

See Method 7.18. Here, $x_1 = 107 + 178 = 285$, $n_1 = 240 + 491 = 731$, $x_2 = 55 + 96 = 151$, $n_2 = 174 + 340 = 514$. The test for equality of two proportions can be tested in R using the following:

```
x1 <- 285
n1 <- 731
x2 <- 151
n2 <- 514

prop.test(c(x1,x2), c(n1,n2), correct = FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(x1, x2) out of c(n1, n2)
## X-squared = 12.249, df = 1, p-value = 0.0004656
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.04318181 0.14902331
## sample estimates:
##    prop 1    prop 2
## 0.3898769 0.2937743
```

The estimated proportion of alcohol-related accidents is smaller in 2017 than in 2010, and the small p-value indicates very strong evidence against the hypothesis of no change from 2010 to 2017.

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

Question X.3 (21)

From the same source, there is also data available on the speed limits for the road stretches where the accidents occurred. The following data, describing the number of rural zone accidents at different speed limits in the years 2010 and 2017, were extracted:

	2010	2017
0 to 50 km/h	54	58
50 to 100 km/h	1280	966
100 to 130 km/h	144	135

What is the result of the usual test for no change in the distribution of accidents in the speed limit intervals between the two years (both your conclusion and reasoning must be correct)? Use the significance level $\alpha = 1\%$.

1* ☐ No significant difference is found in the distribution of speed limits between the two years, as the p-value is larger than the significance level.

2 ☐ A significant difference is found in the distribution of speed limits between the two years, as the p-value is larger than the significance level.

3 ☐ A significant difference is found in the distribution of speed limits between the two years, as the p-value is smaller than the significance level.

4 ☐ No significant difference is found in the distribution of speed limits between the two years, as the p-value is smaller than the significance level.

5 ☐ None of the above statements are true.

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

Data is read into R and a χ^2-test is carried out:

```
data <- matrix(c(54, 1280, 144, 58, 966, 135), ncol = 2)
chisq.test(data)

##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 5.8273, df = 2, p-value = 0.05428
```

It shows that the p-value for the test is above 1%, and hence a significant difference is found.

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

Question X.4 (22)

In connection with the usual test for whether the distribution of speed limits is the same in the two years, the following question is asked: What is the estimated proportion of accidents on roads with speed limits from 50 to 100 km/h in 2017 under the null hypothesis?

1 ☐ $(58 + 966 + 135)/(54 + 1280 + 144 + 58 + 966 + 135) = 0.440$

2 ☐ $(966)/(54 + 1280 + 144 + 58 + 966 + 135) = 0.366$

3 ☐ $(966)/(58 + 966 + 135) = 0.833$

4* ☐ $(1280 + 966)/(54 + 1280 + 144 + 58 + 966 + 135) = 0.852$

5 ☐ $(54 + 58 + 144 + 135)/(54 + 1280 + 144 + 58 + 966 + 135) = 0.148$

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

Under the null hypothesis, the proportion of accidents that happen at 50 to 100 km/h roads does not depend on the accident year, then the proportion is estimated using the data from both years: $x_{50-100\text{km/h}} = 1280+966 = 2246$ and $n_{50-100\text{km/h}} = 54+1280+144+58+966+135 = 2637$. Then

$$\hat{p}_{\text{50-100km/h}} = \frac{x_{\text{50-100km/h}}}{n_{\text{50-100km/h}}} = 0.852$$

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

Exercise XI

Below is a sample of 20 independent observations, read into R in the vector `x`:

```
x <-
  c(13, 12, 9, 7, 12, 15, 12, 10, 6, 13, 7, 13, 19, 12, 6, 4, 15, 16, 11, 18)
```

The data do not originate from a known distribution, but we are interested in the population mean and the uncertainty of its estimate.

Question XI.1 (23)

What is the sample mean $\bar{x}$ and variance s^2 (both quantities must be correct)?

1 ☐ $\bar{x} = 11.2$ and $s^2 = 16.7$.

2* ☐ $\bar{x} = 11.5$ and $s^2 = 16.7$.

3 ☐ $\bar{x} = 11.2$ and $s^2 = 4.1$.

4 ☐ $\bar{x} = 11.5$ and $s^2 = 4.1$.

5 ☐ $\bar{x} = 11.5$ and $s^2 = 16.7^2$.

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

Data can be read into R, and sample mean and variance calculated:

```
x <-
  c(13, 12, 9, 7, 12, 15, 12, 10, 6, 13, 7, 13, 19, 12, 6, 4, 15, 16, 11, 18)
mean(x)

## [1] 11.5

var(x)

## [1] 16.68421
```

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

Question XI.2 (24)

Now, we perform a resampling of `x` to get an idea of the uncertainty of the sample mean. We draw 200 resamples with replacement from the 20 observations in `x`, each with sample size 20. Subsequently, the mean of each of the 200 resamples is calculated. The R code for this operation is:

```
apply(replicate(200, sample(x, replace = TRUE)), 2, mean)
```

Below, the 10 largest and 10 smallest sample means of the 200 resamples are shown.

smallest	9.00	9.65	9.65	9.80	9.90	9.95	10.00	10.00	10.00	10.05
largest	12.95	12.95	12.95	13.00	13.05	13.10	13.10	13.10	13.15	13.40

Using the results above and the book's definition of percentiles ("`type = 2`" in R), which of the following is a 95% bootstrap confidence interval for the population mean?

1 ☐ $[10.05, 12.95]$

2 ☐ $[9.00, 13.40]$

3 ☐ $[9.80, 13.10]$

4 ☐ $[9.65, 13.10]$

5* ☐ $[9.925, 13.075]$

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

See Definition 1.7. For $n = 200$, and $p_1 = 0.025$, $p_2 = 0.975$, it holds that $p_1 n = 5$ and $p_2 n = 195$. Then, the relevant 0.25 quantile $q_{0.025}$ is the average of the 5th and 6th ordered averages, while the 0.975 quantile $q_{0.975}$ is average of the 195th and 196th ordered averages:

$$q_{0.025} = \frac{9.90 + 9.95}{2} = 9.925 \quad \text{and} \quad q_{0.975} = \frac{13.05 + 13.10}{2} = 13.075$$

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

Exercise XII

During the preparation for a small festival, the toilet facilities are taken under consideration. Mobile toilets need to be ordered such that the capacity is sufficient, but not too high, since will lead to more cleaning and higher costs.

It is assumed that, on average, 150 guests need to use the toilets every hour, and that their arrival follows a Poisson distribution. In addition, it is assumed that each toilet can serve 20 guests per hour.

Question XII.1 (25)

Suppose that 10 toilets are ordered. What is then the probability that, in a randomly selected hour, the number of guests who arrive at the toilets exceeds the capacity?

1* ☐ 0.0042%

2 ☐ 2.3%

3 ☐ 11%

4 ☐ 24%

5 ☐ 99%

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

Let X represent the number of guests arriving at the toilets in a randomly selected hour, then $X \sim Pois(150)$. The capacity is $10 \cdot 20 = 200$ per hour, hence we need to calculate $P(X > 200) = 1 - P(X \leq 200)$:

```
1 - ppois(200, lambda=150)

## [1] 4.205886e-05
```

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

Question XII.2 (26)

A group of DTU students have decided to help small festivals optimize their logistical conditions. Among other things, the students have collected data on the use of toilets at small festivals. An examination of these data shows that a better model can be made to represent the number of guests who need to use the facilities in a randomly selected hour. This number can be modelled by an exponential distribution with mean $\frac{\text{“number of guests”}}{10}$, where “number of guests” is the total number of guests at the festival. In this question, this new model must be used.

A festival with 1500 guests is now considered. How many toilets should, at least, be ordered to ensure that the probability that not everyone can use the facilities is less than 2% in a randomly selected hour (given as a call in R)? It is still assumed that each toilet can serve 20 guests per hour.

1 ☐ `ppois(20, lambda = 15) * 20`

2 ☐ `qpois(0.98, lambda = 1500/10) / 20`

3 ☐ `qexp(0.98, rate = 10/15)`

4* ☐ `qexp(0.98, rate = 10/1500) / 20`

5 ☐ `qexp(0.98, rate = 10/1500) * 20`

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

Let Y represent the number of guests arriving at the toilets in a randomly selected hour. Then $Y \sim Exp(10/1500)$. We need to find y such that

$$P(Y \leq 20y) \geq 0.98.$$

We can solve $P(Y \leq 20y) = 0.98$ for y by computing

```r
qexp(0.98, rate = 10/1500) / 20

## [1] 29.34017
```

Thus, ordering 30 toilets or more ensures that the probability in focus stays below 2%.

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

Exercise XIII

Below, there's a small sample with 5 independent observations:

Observations:	11.8071067	-1.7913888	-9.1872410	-4.4860901	-0.2324924

Question XIII.1 (27)

Which of the following answer options is the only one that can possibly be correct?

1 ☐ It is impossible that the observations were sampled from a normal distribution with mean 0 and variance 10^2.

2 ☐ It is possible that the observations were sampled from a uniform distribution with parameters -9 and 12.

3* ☐ It is possible that the observations were sampled from a t-distribution with 1 degree of freedom.

4 ☐ It is possible that the observations were sampled from an F-distribution with 1 and 2 degrees of freedom.

5 ☐ It is possible that the observations were sampled from an exponential distribution with rate 1.

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

The F distributions and exponential distributions don't give rise to negative observations, which eliminates options 4 and 5. Likewise, the uniform distribution with parameters -9 and 12 wouldn't yield the observation -9.1872410, which elimates option 2. There is no reason why the observations could not come from the $N(0, 10^2)$ distribution, which elimates option 1. We're then left with option 3: There's no reason why these observations couldn't come from a t-distribution with 1 degree of freedom (like normal distributions, t distributions take both positive and negative values).

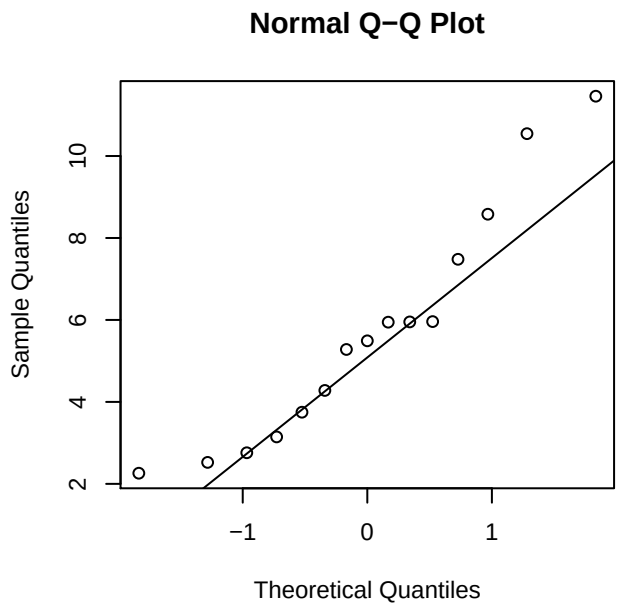
---------- FACIT-END ----------

Exercise XIV

A sample was collected, and the following summary statistics were calculated:

Statistic	Value
$\bar{x}$	5.69
s^2	7.96
Minimum	2.26
Q_1	3.45
Q_2	5.49
Q_3	6.72
Maximum	11.46
n	15

Furthermore, a normal qq-plot was made of the observations:



Question XIV.1 (28)

Which of the following can be concluded using the given information and the book's definition of extreme observations in a sample?

1 ☐ There is one extreme observation in the sample.

2 ☐ There are two extreme observations in the sample.

3 ☐ There are at least three extreme observations in the sample.

4* ☐ There are no extreme observations in the sample.

5 ☐ With the given information, it cannot be concluded whether or not there are extreme observations in the sample.

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

First, note that

$$IQR = Q_3 - Q_1 = 6.72 - 3.45 = 3.27\,,$$

(Definition 1.15) so

$$Q_3 + 1.5 \cdot IQR = 11.625$$
$$Q_1 - 1.5 \cdot IQR = -1.455$$

There are no extreme observations in the sample, as the maximum is smaller than $Q_3 + 1.5 \cdot IQR$ and the minimum is larger than $Q_1 - 1.5 \cdot IQR$.

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

Question XIV.2 (29)

Which of the following conclusions can, with certainty, be drawn about the population using the given information?

1 ☐ The population has no negative values.

2 ☐ The population is normal distributed.

3 ☐ The distribution of the population is left-skewed.

4 ☐ The population has no values above 11.46.

5* ☐ None of the four conclusions above can be drawn.

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

The summary statistics contribute to describing the sample and the data generating distribution, but they don't tell the whole story about the population.

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

Question XIV.3 (30)

Assuming that the observations in the sample are independent and identically normal distributed, what is a correct 95% confidence interval for the population mean?

1 ☐ $5.69 \pm 2.95 \cdot \frac{2.82}{14}$

2 ☐ $5.69 \pm 2.98 \cdot \frac{7.96}{\sqrt{15}}$

3 ☐ $5.69 \pm 1.96 \cdot \frac{2.82}{\sqrt{14}}$

4* ☐ $5.69 \pm 2.14 \cdot \frac{2.82}{\sqrt{15}}$

5 ☐ $5.69 \pm 2.58 \cdot \frac{7.96}{\sqrt{14}}$

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

See Method 3.9:

$$\bar{x} \pm t_{0.975} \cdot \frac{\sqrt{s^2}}{\sqrt{n}} = 5.69 \pm 2.14 \cdot \frac{2.82}{\sqrt{15}}$$

where $t_{0.975}$ is the 0.975 quantile in the t-distribution with 14 degrees of freedom.

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

The exam paper is finished. Have a great Christmas vacation!

Technical University of Denmark *Page 1 of 35 pages.*

Written examination: 27 May 2018

Course name and number: **Introduction to Statistics (02323)**

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

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

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

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

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

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

The exam paper contains 35 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

In order to investigate the download speed at a work station, download times (in seconds) were recorded for 53 files of different sizes (measured in MB). The download times are saved in the vector `time` in R, while the corresponding file sizes are saved in the vector `size`. Furthermore, the following code was executed in R:

```
logtime <- log(time)
model1 <- lm(logtime ~ size)
summary(model1)

##
## Call:
## lm(formula = logtime ~ size)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.571 -0.673  0.132  0.745  2.190
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.109      0.482   -0.23     0.82
## size           0.127      0.023    5.50  1.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.979 on 51 degrees of freedom
## Multiple R-squared:  0.372,Adjusted R-squared:  0.36
## F-statistic: 30.2 on 1 and 51 DF,  p-value: 1.25e-06
```

The statistical model given by `model1` is a linear regression model of the form

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

where ε_i, $i = 1, \ldots, 53$, are assumed to be independent and identically $N(0, \sigma^2)$ distributed.

Question I.1 (1)

Identify the dependent variable and the explanatory variable in the model given by `model1`.

1 ☐ Download time is the dependent variable. File size is the explanatory variable.

2 ☐ File size is the dependent variable. The logarithm of download time is the explanatory variable.

3 ☐ Download time is the explanatory variable. File size is the dependent variable.

4* ☐ File size is the explanatory variable. The logarithm of download time is the dependent variable.

5 ☐ As (the logarithm of) download time depends on file size, both file size and the logarithm of download size are dependent variables. There is no explanatory variable in the model.

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

The variable to the left of the tilde ($\sim$) in the R code corresponds to the dependent variable Y_i, while the variable to the right of the tilde corresponds to the explanatory variable x_i.

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

Question I.2 (2)

Give estimates for the parameters of the model given by `model1`.

1 ☐ $\hat{\beta}_0 = 0.127$, $\hat{\beta}_1 = -0.109$, $\hat{\sigma} = 0.979$

2 ☐ $\hat{\beta}_0 = -0.109$, $\hat{\beta}_1 = 0.127$, $\hat{\sigma} = 0.979^2$

3 ☐ $\hat{\beta}_0 = -0.109$, $\hat{\beta}_1 = 0.127$, $\hat{\sigma} = 0.372$

4 ☐ $\hat{\beta}_0 = 0.127$, $\hat{\beta}_1 = -0.109$, $\hat{\sigma} = 0.979^2$

5* ☐ $\hat{\beta}_0 = -0.109$, $\hat{\beta}_1 = 0.127$, $\hat{\sigma} = 0.979$

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

The estimates of the model intercept, $\hat{\beta}_0$, and slope, $\hat{\beta}_1$, are given in the column `Estimate` in the R output (in the rows `(Intercept)` and `size`, respectively). The estimated standard deviation of the error, $\hat{\sigma}$, is termed `Residual standard error` in the R output.

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

Question I.3 (3)

The following code was also executed in R:

```r
mean(size)

## [1] 20.088

(53-1)*var(size)

## [1] 1807.6
```

Using the model given by `model1` as a starting point, give a 90% prediction interval for the logarithm of the download time for a file of size 17 MB.

1 ☐ $-0.109 + 0.127 \cdot 17 \pm 2.0076 \cdot 0.979 \cdot \sqrt{1 + \frac{1}{53} + \frac{(17-20.088)^2}{1807.6}}$

2* ☐ $17 \cdot 0.127 - 0.109 \pm 0.979 \cdot 1.6753 \cdot \sqrt{1 + \frac{1}{53} + \frac{(20.088-17)^2}{1807.6}}$

3 ☐ $-0.109 + 0.127 \cdot 17 \pm 1.6753 \cdot \sqrt{0.979} \cdot \sqrt{1 + \frac{1}{53} + \frac{(17-20.088)^2}{1807.6}}$

4 ☐ $-0.109 + 17 \cdot 0.127 \pm 1.6753 \cdot 0.979 \cdot \sqrt{\frac{1}{53} + \frac{(17-20.088)^2}{1807.6}}$

5 ☐ $17 \cdot 0.127 + 0.109 \pm 0.979 \cdot 2.0076 \cdot \sqrt{1 + \frac{1}{53} + \frac{(20.088-17)^2}{1807.6}}$

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

Use Method 5.18, equation (5-60). Note that 1.6753 is the 0.95 quantile of the t-distribution with 51 degrees of freedom, `qt(0.95, df = 51)` in R.

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

Question I.4 (4)

Using the model given by `model1` as a starting point, one would like to investigate whether there is a significant linear relationship between the logarithm of download time and file size. Formulate the corresponding statistical null hypothesis that there is no association between the two variables.

1 ☐ $H_0 : \hat{\beta}_1 = 0$

2 ☐ $H_0 : \beta_1 \neq \beta_0$

3* ☐ $H_0 : \beta_1 = 0$

4 ☐ $H_0 : \beta_1 \neq 0$

5 ☐ $H_0 : \hat{\beta}_1 \neq 0$

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

Under the null hypothesis $H_0 : \beta_1 = 0$, the model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ reduces to $Y_i = \beta_0 + \varepsilon$. The latter corresponds to a model for one sample (in which the logarithm of the download time, Y_i, does not depend on file size, x_i). Answer 1 is wrong since we are making a null hypothesis for the true mean and not our estimate.

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

Continue on page 6

Exercise II

Train carriages in a mine are to be loaded with large pieces of blasted rock. Initially, a machine has sorted the pieces into two piles based on their size and weight. The weights of the pieces of rock are assumed to be independent and normally distributed, such that the weight of a randomly extracted piece from pile 1 can be represented by $X_1 \sim N(20, 5^2)$ kg and from pile 2 by $X_2 \sim N(50, 10^2)$ kg.

Question II.1 (5)

What is the probability that a randomly selected piece from pile 1 weighs more than 25 kg?

1* ☐ 15.9 %

2 ☐ 84.1 %

3 ☐ 42.1 %

4 ☐ 57.9 %

5 ☐ None of the values listed.

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

Here,

$$P(X_1 > 25) = 1 - P(X_1 \leq 25) = 1 - F_{X_1}(25)$$

is to be computed, where F_{X_1} denotes the cumulative distribution function of the normal distribution with mean 20 and standard deviation 5. In R, the result may be found as:

```
1 - pnorm(q = 25, mean = 20, sd = 5)
```

```
## [1] 0.1586553
```

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

Question II.2 (6)

Choose a correct statement:

There is a 20% probability of a randomly selected piece of rock from pile 2 being heavier than

1 ☐ 41.6 kg.

2 ☐ 52.5 kg.

3* ☐ 58.4 kg.

4 ☐ 67.4 kg.

5 ☐ 134 kg.

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

Let F_{X_2} be the cumulative distribution function of the normal distribution with mean 50 and standard deviation 10. Here, we need to find x such that

$$P(X_2 > x) = 0.2\,,$$

which corresponds to finding x such that

$$P(X_2 \leq x) = F_{X_2}(x) = 0.8.$$

In R, x may be found as:

```
qnorm(0.8, mean = 50, sd = 10)
```

```
## [1] 58.41621
```

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

Question II.3 (7)

If the train carriages are loaded too heavily, operations must stop. A manual crane must be used to remove pieces of rock from the overloaded carriage, and this procedure is very costly.

Two robotic cranes load the train carriages. One crane takes pieces from pile 1, and the other takes pieces from pile 2. If each crane takes 10 pieces from its pile, and all 20 pieces are loaded into the same (empty) carriage, what is the probability that the total load of this train carriage exceeds 800 kg?

1 ☐ The total weight is $Y \sim N(700, 15625)$, so $P(Y > 800) = 21.2\%$.

2 ☐ The total weight is $Y \sim N(700, 12500)$, so $P(Y > 800) = 18.6\%$.

3 ☐ The total weight is $Y \sim N(700, 2500)$, so $P(Y > 800) = 2.28\%$.

4* ☐ The total weight is $Y \sim N(700, 1250)$, so $P(Y > 800) = 0.234\%$.

5 ☐ The total weight is $Y \sim N(700, 225)$, so $P(Y > 800) = 1.31 \cdot 10^{-9}\%$.

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

Let X_{1i} and X_{2i}, $i = 1, \ldots, 10$, with $X_{1i} \sim N(20, 5^2)$ and $X_{2i} \sim N(50, 10^2)$, be independent random variables which represent the weights of 10 randomly selected pieces of rock from pile 1 and pile 2, respectively. Then, the total load may be represented by

$$Y = \sum_{i=1}^{10} X_{1i} + \sum_{i=1}^{10} X_{2i} \, .$$

According to Theorem 2.40 and Theorem 2.56, Y is normally distributed with

$$\mathrm{E}(Y) = \sum_{i=1}^{10} \mathrm{E}(X_{1i}) + \sum_{i=1}^{10} \mathrm{E}(X_{2i}) = 10 \cdot 20 + 10 \cdot 50 = 700$$
$$\mathrm{V}(Y) = \sum_{i=1}^{10} \mathrm{V}(X_{1i}) + \sum_{i=1}^{10} \mathrm{V}(X_{2i}) = 10 \cdot 5^2 + 10 \cdot 10^2 = 1250$$

so

$$P(Y > 800) = 1 - P(Y \leq 800) = 1 - F_Y(800)$$

where F_Y is the cumulative distribution function for the $N(700, 1250)$ distribution. In R, the result may then be computed as:

```
1 - pnorm(800, mean = 700, sd = sqrt(1250))
```

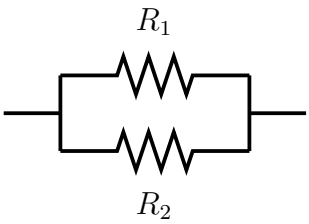
```
## [1] 0.002338867
```

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

Continue on page 9

Exercise III

The resistances in the electrical circuit



are estimated to be $\hat{R}_1 = 2\ \Omega$ and $\hat{R}_2 = 3\ \Omega$, with an estimated standard deviation for the estimators (*standard error*) of $\hat{\sigma}_{\hat{R}_1} = 0.2$ and $\hat{\sigma}_{\hat{R}_2} = 0.5$, respectively. R_1 and R_2 can be assumed independent and normally distributed.

The total resistance through the circuit is given by

$$R = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2}}$$

Question III.1 (8)

Using simulation, determine a 95% confidence interval for the total resistance R. The following R code must be used to get the specified result (and after this code has been executed, only one additional function call in R is needed to get the result):

```
set.seed(7643)
k <- 10000
R1 <- rnorm(k, mean = 2, sd = 0.2)
R2 <- rnorm(k, mean = 3, sd = 0.5)
R <- 1/(1/R1 + 1/R2)
```

1 ☐ [1.11, 1.29]

2* ☐ [0.96, 1.40]

3 ☐ [0.92, 1.47]

4 ☐ [0.82, 1.58]

5 ☐ [0.72, 1.68]

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

We have used the parametric bootstrap simulation approach to error propagation as described in Method 4.7:

```r
set.seed(7643)
k <- 10000
R1 <- rnorm(k, mean = 2, sd = 0.2)
R2 <- rnorm(k, mean = 3, sd = 0.5)
R <- 1/(1/R1 + 1/R2)
quantile(R, c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.9647361 1.4016874
```

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

Continue on page 11

Exercise IV

Students' choice of higher education has great political awareness. The table below is based on numbers from "Universiteternes Statistiske Beredskab", and contains the number of newly enrolled students by discipline for selected years (row and column sums are included in italics)

	Y2012	Y2016	Y2017	*Sum*
Hum	7966	7297	6691	*21954*
Soc	10173	10253	10006	*30432*
Hlth	2789	3137	3157	*9083*
TechNat	8551	10130	10339	*29020*
Sum	*29479*	*30817*	*30193*	*90489*

Question IV.1 (9)

Based on the numbers for 2017, one wants to test the hypothesis that the proportion of students newly enrolled in a technical or natural science bachelor education (TechNat) is 32.0%. The relevant test statistic for this hypothesis, which is assumed to be well approximated by a standard normal distribution, is

1 ☐ $(10253 - 0.68 \cdot 30817)/\sqrt{30817 \cdot 0.32 \cdot 0.68} = -130.70$

2 ☐ $(10339 - 0.32 \cdot 30193)/\sqrt{30193 \cdot 0.32 \cdot 0.32} = 12.18$

3 ☐ $(10130 - 0.32 \cdot 30817)/\sqrt{30817 \cdot 0.32 \cdot 0.68} = 3.28$

4* ☐ $(10339 - 0.32 \cdot 30193)/\sqrt{30193 \cdot 0.32 \cdot 0.68} = 8.36$

5 ☐ $(10339 - 0.68 \cdot 30193)/\sqrt{30193 \cdot 0.32 \cdot 0.32} = -183.30$

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

Using equation (7-16) with $x = 10339$, $n = 30193$, and $p_0 = 0.32$:

```
(10339 - 0.32 * 30193) / sqrt(30193 * 0.32 * 0.68)
```

```
## [1] 8.355261
```

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

Question IV.2 (10)

It is also of interest, whether there was a change from 2016 to 2017 in the proportion of students who were newly enrolled in the humanities (Hum).

Four different tests are performed in R using the following code:

```
  prop.test(x = 6691, n = 30193, p = 7297/30817, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  6691 out of 30193, null probability 7297/30817
## X-squared = 38.5, df = 1, p-value = 5.5e-10
## alternative hypothesis: true p is not equal to 0.23678
## 95 percent confidence interval:
##  0.21696 0.22633
## sample estimates:
##       p
## 0.22161
```

```
prop.test(x = c(7297, 6691), c(30817, 30193), correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(7297, 6691) out of c(30817, 30193)
## X-squared = 19.9, df = 1, p-value = 8.2e-06
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.0085083 0.0218461
## sample estimates:
##  prop 1  prop 2
## 0.23678 0.22161
```

```
binom.test(x = 6691, n = 30193, p = 7297/30817)
```

```
##
##  Exact binomial test
##
## data:  6691 and 30193
## number of successes = 6691, number of trials = 30193, p-value = 4.3e-10
## alternative hypothesis: true probability of success is not equal to 0.23678
## 95 percent confidence interval:
##  0.21693 0.22634
## sample estimates:
## probability of success
##                0.22161
```

```
binom.test(x = 6691, n = 30193, p = (6691+7297)/(30193+30817))

##
##  Exact binomial test
##
## data:  6691 and 30193
## number of successes = 6691, number of trials = 30193, p-value = 0.0015
## alternative hypothesis: true probability of success is not equal to 0.22927
## 95 percent confidence interval:
##  0.21693 0.22634
## sample estimates:
## probability of success
##                0.22161
```

Which line of code performs the desired test?

1 ☐ `binom.test(x = 6691, n = 30193, p = 7297/30817)`

2* ☐ `prop.test(x = c(7297, 6691), c(30817, 30193), correct = FALSE)`

3 ☐ `binom.test(x = 6691, n = 30193, p = (6691+7297)/(30193+30817))`

4 ☐ `prop.test(x = 6691, n = 30193, p = 7297/30817, correct = FALSE)`

5 ☐ None of the four lines of code tests the relevant hypothesis

---------- 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).

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

Question IV.3 (11)

Using a hypothesis test, it is also to be investigated whether the distribution of newly enrolled students across disciplines has changed over the three years for which data is given. The number of degrees of freedom in the relevant distribution of the test statistic is:

1 ☐ 3

2 ☐ 4

3* ☐ 6

4 ☐ 12

5 ☐ 20

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

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

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

Question IV.4 (12)

Assuming independence between year and discipline, the expected number of newly enrolled students in TechNat in the year 2017 is estimated to be

1 ☐ 10339

2* ☐ $29020 \cdot 30193/90489 = 9683$

3 ☐ $(8551 + 10130 + 10339)/3 = 9673$

4 ☐ $10339 \cdot 29020/30193 = 9937$

5 ☐ $10339 \cdot 30193/29020 = 10757$

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

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

$$\frac{\text{column total} \times \text{row total}}{\text{grand total}} = \frac{30193 \times 29020}{90489} = 9683$$

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

Question IV.5 (13)

As the next step in testing whether the distribution across disciplines has changed over the years, the following is given: The test statistic is calculated to be 314.5. 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 numbers provided above cannot be used to argue statistically, whether the distribution across disciplines has changed.

2 ☐ The distribution across disciplines has not changed significantly, as the test statistic is greater than the given 0.95 quantile.

3 ☐ The distribution across disciplines has not changed significantly, as the test statistic is greater than the given 0.975 quantile.

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

5* ☐ The distribution across disciplines has changed significantly, as the test statistic is greater than the given 0.95 quantile.

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

Method 7.22. The test statistic is greater than the given 0.95 quantile, so the null hypothesis of no difference between groups (no change across years) is rejected, and the change is concluded to be significant.

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

Continue on page 16

Exercise V

An experiment was conducted with the purpose of investigating the shelf life of a certain type of medicine. Altogether 26 identical, unopened bottles of medicine with the same production date were used for the experiment. Half of the bottles were stored at room temperature (21 °C), the other half at fridge temperature (5 °C). After 90 days the bottles were opened, and the content of active substance (in mg/ml) in each bottle was measured. The results were read into R in two vectors, `hightemp` (measurements from bottles stored at 21 °C) and `lowtemp` (measurements from bottles stored at 5 °C).

Furthermore, the following code was executed in R:

```
t.test(lowtemp, hightemp)

##
##  Welch Two Sample t-test
##
## data:  lowtemp and hightemp
## t = 5.3, df = 24, p-value = 2e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.4316 3.2592
## sample estimates:
## mean of x mean of y
##    7.4397    5.0943
```

Question V.1 (14)

What may be concluded when the significance level is set to $\alpha = 0.05$?

1* ☐ The mean content of the active substance is significantly larger after storage at fridge temperature than at room temperature. The difference is estimated to be 2.35 mg/ml.

2 ☐ It can be seen from the p-value that there is no significant difference between the mean content of the active substance in bottles stored at room temperature and at fridge temperature, respectively.

3 ☐ The mean content of the active substance is significantly larger after storage at fridge temperature than at room temperature. The difference is estimated to be 5.30 mg/ml.

4 ☐ The mean content of the active substance is significantly less after storage at fridge temperature than at room temperature. The difference is estimated to be 2.35 mg/ml.

5 ☐ The 95% confidence interval contains 2.35. Therefore, there is no significant difference between the mean content of the active substance in bottles which were stored at room temperature and fridge temperature, respectively.

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

The p-value $2 \cdot 10^{-5}$ is much smaller than the significance level $\alpha = 0.05$ (or: 0 is not contained in the 95% confidence interval), so the difference is significant. The mean content of the active substance is estimated to be $7.4397 - 5.0943 = 2.35$ mg/ml larger in bottles which were stored at fridge temperature than in those which were stored at room temperature.

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

Question V.2 (15)

A 99% confidence interval for the difference in the mean content of the active substance between bottles stored at fridge temperature and room temperature may be determined as follows:

1 ☐ $2.3454 \pm \frac{3.2592-2.3454}{2.7969} \cdot 2.0639 = [1.67, 3.02]$

2 ☐ $2.3454 \pm \frac{3.2592-2.3454}{2.4922} \cdot 2.7969 = [1.32, 3.37]$

3 ☐ $\left[1.4316 \cdot \frac{2.7969}{2.0639}, 3.2592 \cdot \frac{2.7969}{2.0639}\right] = [1.94, 4.42]$

4* ☐ $2.3454 \pm \frac{3.2592-2.3454}{2.0639} \cdot 2.7969 = [1.11, 3.58]$

5 ☐ $2.3454 \pm \frac{3.2592-2.3454}{2.0639} \cdot 2.4922 = [1.24, 3.45]$

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

Using the notation from Method 3.47, it may be concluded from the R output that $\bar{x} - \bar{y} = 7.4397 - 5.0943 = 2.3454$, and that the degrees of freedom $\nu = 24$, so that $t_{0.975} = 2.0639$ and $t_{0.995} = 2.7969$ (`qt(0.975, df = 24)` and `qt(0.995, df = 24)`, respectively, in R).

Furthermore, according to the R output, $[1.4316, 3.2592]$ is a 95% confidence interval for the difference in mean content. The only thing in the equation we do not know is $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, but since we have one confidence interval given (95%), we can isolate this term from that equation. Thus, it follows from Method 3.47 that

$$\bar{x} - \bar{y} \pm t_{0.975} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.3454 \pm 2.0639 \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = [1.4316, 3.2592]$$

which may be used to conclude that

$$2.3454 + 2.0639 \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 3.2592 \Leftrightarrow \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \frac{3.2592 - 2.3454}{2.0639}.$$

Now, the 99% confidence interval can be computed using Method 3.47, as well:

$$\bar{x} - \bar{y} \pm t_{0.995} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.3454 \pm 2.7969 \cdot \frac{3.2592 - 2.3454}{2.0639} = [1.11, 3.58]\,.$$

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

Continue on page 19

Exercise VI

In a production process things sometimes go wrong, and a component must be discarded after an inspection. From experience, it is known that there is a 20% probability of a component needing to be discarded. The assessment ("keep" or "discard") for a given component is independent of the assessments for the other components. One can assume that 20 components are produced per day.

Question VI.1 (16)

What is the probability that, on a randomly selected day, no components need to be discarded?

1 ☐ The number of components which need to be discarded is hypergeometrically distributed, and the probability is 0.

2* ☐ The number of components which need to be discarded is binomial distributed, and the probability is 0.0115.

3 ☐ The number of components which need to be discarded is binomial distributed, and the probability is 0.0576.

4 ☐ The number of components which need to be discarded is hypergeometrically distributed, and the probability is 0.0692.

5 ☐ The number of components which need to be discarded is binomial distributed, and the probability is 0.630.

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

The number of components that get the assessment *discard* ("number of successes") out of 20 independently assessed components ("number of independent trials") is binomial distributed with probability $p = 0.2$ and size $n = 20$.

Let X be binomial distributed with probability $p = 0.2$ and size $n = 20$. Then $P(X = 0)$ (or, equivalently, $P(X \leq 0)$) may be computed in R as follows

```
dbinom(0, size = 20, p = 0.2)

## [1] 0.01152922

pbinom(0, size = 20, p = 0.2)

## [1] 0.01152922
```

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

Question VI.2 (17)

A simulation of the number of discarded components per day has been carried out. Let X_i denote the number of discarded components on a randomly selected day i. A sample of $n = 20$ values has been simulated, and is denoted by x_i, $i = 1, \ldots, 20$. Which of the following statements is the only one that can be correct?

1 ☐ The expected number of components to be discarded during one day is $\mu_X = 4.5$. The sample mean of the simulated values was $\hat{\mu}_X = 4.3$.

2* ☐ The expected number of components to be discarded during one day is $\mu_X = 4$. The sample mean of the simulated values was $\hat{\mu}_X = 3.7$.

3 ☐ The expected number of components to be discarded during one day is $\mu_X = 5$. The sample mean of the simulated values was $\hat{\mu}_X = 4.9$.

4 ☐ The expected number of components to be discarded during one day is $\mu_X = 2$. The sample mean of the simulated values was $\hat{\mu}_X = 2.2$.

5 ☐ The expected number of components to be discarded during one day is $\mu_X = 4.25$. The sample mean of the simulated values was $\hat{\mu}_X = 4.25$.

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

Again, let X be binomial distributed with probability $p = 0.2$ and size $n = 20$. See theorem 2.21. The expected number of components to be discarded during one day, μ_X, may then be computed as

$$\mu_X = \mathrm{E}(X) = np = 20 \cdot 0.2 = 4\,,$$

so the answer option with $\mu_X = 4$ is the only one that can be correct.

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

Question VI.3 (18)

An experiment is planned in order to estimate the proportion of components that need to be discarded.

How many days does the experiment need to run in order to obtain a 95% confidence interval for the proportion of components to be discarded, which has an expected width of 5 percentage points?

1 ☐ $n = (1.96 \cdot 0.2/0.05)^2 = 61.5$, i.e. 4 days.

2 ☐ $n = 0.16 \cdot (1.96/0.05)^2 = 246$, i.e. 13 days.

3* ☐ $n = 0.16 \cdot (1.96/0.025)^2 = 983$, i.e. 50 days.

4 ☐ $n = 0.2 \cdot (1.96/0.025)^2 = 1229$, i.e 62 days.

5 ☐ $n = (1.96 \cdot 0.2/0.025)^2 = 3934$, i.e. 197 days.

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

Use Method 7.13 with $p = 0.2$, remembering that the ME is half the expected width of the confidence interval, and using the information that 20 components are produced per day.

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

Continue on page 22

Exercise VII

A factory which produces percussion instruments and accessories produces drumsticks as well. For the purpose of quality control, the lengths (in cm) of 20 drumsticks selected at random were measured. These lengths were read into the vector `length` in R. The lengths can be assumed to be independent and normally distributed with mean μ and variance σ^2.

Question VII.1 (19)

The following commands were executed in R:

```
sum(length)

## [1] 793.1

var(length)

## [1] 0.01099
```

Give an expression for a 95% confidence interval for the mean length of the drumsticks.

1 ☐ $\frac{793.1}{20} \pm 2.093 \cdot \frac{0.01099}{\sqrt{20}}$

2 ☐ $793.1 \pm 2.086 \cdot \frac{0.01099}{\sqrt{20}}$

3* ☐ $\frac{793.1}{20} \pm 2.093 \cdot \frac{\sqrt{0.01099}}{\sqrt{20}}$

4 ☐ $\frac{793.1}{20} \pm 2.086 \cdot \frac{\sqrt{0.01099}}{\sqrt{20}}$

5 ☐ $\frac{793.1}{20} \pm 2.093 \cdot \frac{0.01099}{\sqrt{19}}$

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

Use Method 3.9 / (3-11) with $n = 20$, $\bar{x} = 793.1/20$, $s = \sqrt{0.01099}$. The 0.975 quantile for the t-distribution with 19 degrees of freedom is:

```
qt(0.975, df = 19)

## [1] 2.093024
```

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

Question VII.2 (20)

A t-test is performed in order to investigate whether the mean length of the drumsticks may be assumed to be 39.60 cm. The observed t-test statistic is calculated to be $t_{\text{obs}} = 2.38$. Which of the following is the only correct conclusion?

1 ☐ The mean length is significantly different from 39.60 cm when the significance level $\alpha = 0.01$ is used.

2 ☐ As the p-value is greater than 0.05, the null hypothesis that the mean length is 39.60 cm is rejected, no matter whether the significance level is set to $\alpha = 0.05$ or $\alpha = 0.01$.

3 ☐ As the p-value is less than 0.05, the null hypothesis that the mean length is 39.60 cm is accepted when the significance level is set to $\alpha = 0.05$.

4* ☐ The mean length is significantly different from 39.60 cm when the significance level is set to $\alpha = 0.05$.

5 ☐ As the p-value is greater than 0.05, the null hypothesis that the mean length is 39.60 cm is accepted, no matter whether the significance level $\alpha = 0.05$ or $\alpha = 0.01$ is used.

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

The p-value is

$$p = 2 \cdot P(T > |t_{\text{obs}}|) = 2 \cdot P(T > 2.38) = 2 \cdot (1 - P(T \leq 2.38)) = 0.028$$

where $P(T \leq 2.38)$ is computed in R as `pt(2.38, df = 19)`.

As the p-value lies in between 0.01 and 0.05, the null hypothesis is rejected at significance level $\alpha = 0.05$ (but not at significance level $\alpha = 0.01$).

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

Question VII.3 (21)

In relation to subsequent quality control, plans are made to select a new sample. The sample size must be sufficiently large to detect a difference of 0.5 mm between the mean length of the drumsticks and the desired mean length 39.60 cm with power 90%, at significance level $\alpha = 0.01$. Here, 0.11 is to be used as a guess for the population standard deviation. Use the `power.t.test` function in R to determine the necessary sample size.

1 ☐ $n = 20$

2 ☐ $n = 22$

3 ☐ $n = 146$

4 ☐ $n = 53$

5* ☐ $n = 76$

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

```
power.t.test(power = 0.9, delta = 0.05, sd = 0.11,
             sig.level = 0.01, type = "one.sample")

##
##      One-sample t test power calculation
##
##               n = 75.36328
##           delta = 0.05
##              sd = 0.11
##       sig.level = 0.01
##           power = 0.9
##     alternative = two.sided
```

See example 3.67 for examples of power.t.test.

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

Continue on page 25

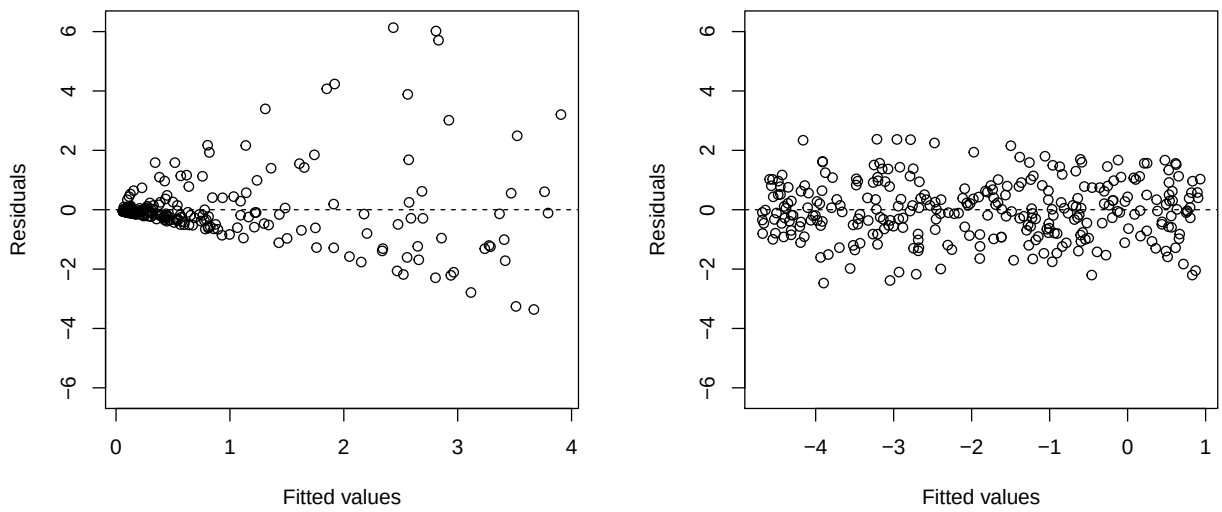
Exercise VIII

Two quantitative variables have been read into R as `x` and `y`. One would like to describe the relationship between these two variables using a linear regression model. To this end, two different linear regression models have been estimated in R, see the R code below.

```
logx = log(x)
logy = log(y)
model1 <- lm(y ~ x)
model2 <- lm(logy ~ logx)
```

Question VIII.1 (22)

Plots of the residuals against the fitted values for `model1` (left) and `model2` (right), respectively, are shown below. On the basis of these plots, would one prefer to analyse the data using the statistical model given by `model1` or the one given by `model2`? (Both the conclusion and reasoning must be valid).



1 ☐ There are clear linear associations between the residuals and the fitted values in the plot to the left, while no linear association is seen in the plot to the right. Thus, one would prefer to use `model1`.

2 ☐ The assumption of variance homogeneity is clearly not satisfied for `model2`, while the assumption seems reasonable for `model1`. Thus, one would prefer to use `model1`.

3 ☐ In the plot to the left, a lot of the fitted values lie in the interval [0,1], while they are better spread out over the whole x axis in the plot to the right. Thus, one would prefer to use `model2`.

4* ☐ The assumption of variance homogeneity is clearly not satisfied for `model1`, while the assumption seems reasonable for `model2`. Thus, one would prefer to use `model2`.

5 ☐ The residuals in the figure to the right are obviously uniformly distributed in an interval around 0 (thus not normally distributed), while the residuals in the figure to the left might well be normally distributed. Thus, one would prefer to use `model1`.

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

In the plot to the left, it is clear that the variance of the residuals increases with the fitted values, revealing that the assumption of variance homogeneity does not hold for `model1`. In the plot to the right, the variance of the residuals seems to be quite constant across the fitted values, indicating that the assumption of variance homogenity should be ok for `model2`.

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

Continue on page 27

Exercise IX

Many factors affect the indoor climate of a building. One of the most common measures for the quality of the indoor climate is the level of CO_2. If there is insufficient ventilation, the CO_2 level becomes too high, which, among other things, decreases peoples' ability to concentrate. In new buildings with classrooms, the CO_2 level may not exceed 1000 ppm - in outdoor air there is around 400 ppm (before the industrial revolution it was around 280 ppm!).

In a study of the indoor climate in classrooms, samples of the CO_2 level were taken from two different classrooms. Both samples consist of one-hour average values measured over a period of 2 months. Only values where people were present in the classroom have been included in the samples. The observations for room 1 and room 2, respectively, were loaded into R in the vectors `room1CO2` and `room2CO2`.

Question IX.1 (23)

The following code was run in R:

```
length(room1CO2)

## [1] 304

length(room2CO2)

## [1] 252

sum((room1CO2 - mean(room1CO2))^2)

## [1] 131606104

(length(room2CO2)-1)*var(room2CO2)

## [1] 12775276
```

Determine the sample standard deviation for room 1 (s_1) and room 2 (s_2), respectively.

1* ☐ $s_1 = 659.0475$ and $s_2 = 225.6048$

2 ☐ $s_1 = 434343.6$ and $s_2 = 50897.51$

3 ☐ $s_1 = 657.9626$ and $s_2 = 225.1567$

4 ☐ $s_1 = 11471.97$ and $s_2 = 3574.252$

5 ☐ None of the four answers above can be correct.

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

See definition 1.11.

$$s_1 = \sqrt{\frac{131606104}{304-1}} = 659.0475$$
$$s_2 = \sqrt{\frac{12775276}{252-1}} = 225.6048$$

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

Question IX.2 (24)

In addition, the following has been run in R:

```
Q3 <- function(x){ quantile(x, 0.75) }

simSamples1 <- replicate(10000, sample(room1CO2, replace = TRUE))
simSamples2 <- replicate(10000, sample(room2CO2, replace = TRUE))

simQ3s1 <- apply(simSamples1, 2, Q3)
simQ3s2 <- apply(simSamples2, 2, Q3)
simQ3sdiff <- simQ3s1 - simQ3s2

quantile(simQ3s1, c(0, 0.025, 0.05, 0.95, 0.975, 1))

##       0%     2.5%       5%      95%    97.5%     100%
## 1417.896 1562.332 1583.146 1833.104 1838.021 1953.792

quantile(simQ3s2, c(0, 0.025, 0.05, 0.95, 0.975, 1))

##       0%     2.5%       5%      95%    97.5%     100%
## 772.0833 827.5000 831.1042 916.4583 920.5000 966.6667

quantile(simQ3sdiff, c(0, 0.025, 0.05, 0.95, 0.975, 1))

##        0%      2.5%        5%       95%     97.5%      100%
##  534.6458  685.8297  712.4562  976.1042  991.6250 1093.4167
```

Use this R output to determine a 95% confidence interval for the difference between the 0.75 quantiles for the CO_2 level in room 1 and 2.

1 ☐ $[916, 1833]$

2 ☐ $[828, 921]$

3* ☐ $[686, 992]$

4 ☐ $[1556 - 828, 1838 - 921] = [728, 917]$

5 ☐ $[828, 1838]$

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

The 95% bootstrap confidence interval is determined by the 0.025 and 0.975 quantiles of the simulated differences (`simQ3sdiff`).

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

Continue on page 30

Exercise X

The table below shows the average yield (measured in hkg/acres) for 5 crops (Crop 1-5) in Denmark in the years 2014-2017.

	2014	2015	2016	2017	*Average*
Crop 1	79	80	73	83	*78.75*
Crop 2	46	48	47	52	*48.25*
Crop 3	64	63	57	66	*62.50*
Crop 4	66	68	62	68	*66.00*
Crop 5	57	60	55	58	*57.50*
Average	*62.40*	*63.80*	*58.80*	*65.40*	*62.60*

In addition to the row and column averages given in the table (in italics), it is given that $SS(\text{Year}) = 118.8$ and $SST = 2172.8$.

In this exercise, the 20 average crop yields in the table are considered to be observations from 20 different randomly selected fields. It is assumed that within each year, there is no difference between the expected yields from the five crops. The analysis must therefore be carried out as if the same crop was sown on all the fields (and the information about crop type should not be used in the exercise). We use a model of the form

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

with $\sum \alpha_i = 0$, where $\varepsilon_{ij} \sim N(0, \sigma^2)$ and independent.

Question X.1 (25)

Let α_1 describe the effect of year 2014 on the expected yield. Give the estimate of α_1.

1 ☐ $\hat{\alpha}_1 = 78.75 - 62.40 = 16.35$

2* ☐ $\hat{\alpha}_1 = 62.40 - 62.60 = -0.20$

3 ☐ $\hat{\alpha}_1 = 78.75$

4 ☐ $\hat{\alpha}_1 = 62.60$

5 ☐ $\hat{\alpha}_1 = 62.40$

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

As stated in equation 8-4 $\hat{\alpha}_1$ is computed as the average yield in the year 2014 (62.40) minus the overall average yield (62.60).

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

Question X.2 (26)

Set the significance level to $\alpha = 0.05$. Give the critical value for the usual test used to investigate whether the expected crop yield differs between years.

1 ☐ 3.73

2* ☐ 3.24

3 ☐ 26.30

4 ☐ 3.49

5 ☐ 2.96

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

Theorem 8.6. The F-test statistic is evaluated using the F distribution with

$$(k-1, n-k) = (4-1, 20-4) = (3, 16)$$

degrees of freedom, and the critical value is the 0.95 quantile of this distribution:

```
qf(0.95, df1 = 3, df2 = 16)
```

```
## [1] 3.238872
```

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

Question X.3 (27)

Give the estimate of σ^2.

1 ☐ $\hat{\sigma}^2 = 2054$

2 ☐ $\hat{\sigma}^2 = 3.058$

3 ☐ $\hat{\sigma}^2 = 36.7$

4 ☐ $\hat{\sigma}^2 = 29.7$

5* ☐ $\hat{\sigma}^2 = 128.375$

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

As can be read in chapter 8.2.2, MSE and MST are both central estimators for the the variance, but that MSE holds true both if the null-hypothesis is rejected or not, so in this case we will use this as the estimate.

$$SSE = SST - SS(\text{Year}) = 2172.8 - 118.8 = 2054$$

and

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-k} = \frac{2054}{16} = 128.375$$

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

Continue on page 33

Exercise XI

An experiment was carried out, and the following sample was collected. Here, the sample is sorted in increasing order and loaded into R. A built-in R function has also been run on the sample values. Note that some values in the output are replaced with a letter.

```
x <- c(-6.5, -6.5, -5.2, -4.9, -4.8, -2.9, -2.6, -2.0, -1.9, -1.8, -1.5,
       -0.1, 1.9, 2.1, 2.6, 5.2, 8.0)

t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = -1.24, df = 16, p-value = A
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -3.330  0.871
## sample estimates:
## mean of x
##     B
```

Question XI.1 (28)

What is the median of the sample (using the book's definition of the sample median)?

1 ☐ $Q_2 = q_{0.5} = -2.0$

2* ☐ $Q_2 = q_{0.5} = -1.9$

3 ☐ $Q_2 = q_{0.5} = -1.85$

4 ☐ $Q_2 = q_{0.5} = -1.8$

5 ☐ $Q_2 = q_{0.5} = 0$

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

The book's definition of the sample median corresponds to R's `type = 2` median (Or one could follow definition 1.5 and calculate it by hand):

```
x <- c(-6.5, -6.5, -5.2, -4.9, -4.8, -2.9, -2.6, -2.0, -1.9, -1.8, -1.5,
       -0.1, 1.9, 2.1, 2.6, 5.2, 8.0)
median(x, type = 2)

## [1] -1.9
```

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

Question XI.2 (29)

When testing the null hypothesis for the mean

$$H_0 : \mu = 10$$

at significance level $\alpha = 0.05$, what is the conclusion based on the calculations carried out in R?

1 ☐ The null hypothesis is accepted and it can be concluded that the mean is not equal to 10.

2 ☐ The null hypothesis is rejected and it can be concluded that the mean is significantly greater than 10.

3 ☐ The null hypothesis is accepted.

4* ☐ The null hypothesis is rejected and it can be concluded that the mean is significantly less than 10.

5 ☐ None of the above conclusions can be drawn based on the calculations carried out.

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

10 is not included in the 95% confidence interval. Therefore we reject that the true $\mu = 10$. Since the confidence interval contains values smaller than 10, we conclude that the mean is significantly smaller than 10.

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

Question XI.3 (30)

What is the sample mean $\bar{x}$?

1 ☐ $(-1.9 - 1.8)/2 = -1.85$

2 ☐ $(-1.9 - 1.8)/15 = -0.25$

3 ☐ $(8.0 - 6.5)/2 = 0.75$

4 ☐ $-1.24/2 = -0.62$

5* ☐ $(0.871 - 3.330)/2 = -1.23$

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

Simply the middle of the confidence interval: $(0.871 - 3.330)/2 = -1.23$

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

The exam is finished. Have a great summer!

Technical University of Denmark *Page 1 of 40 pages.*

Written examination: 14. August 2019

Course name and number: **Introduction to Statistics (02323)**

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 11 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	3	1	4	5	2	4	3	3	1	4

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

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

The exam paper contains 40 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

Assume that $X_1, \ldots, X_{25}$ are independent random variables, which are normal distributed with $N(5, 2^2)$.

Question I.1 (1)

Which of the following values has the property: The probability that X_1 is lower than this value is 15% (remember that the answer can be rounded)?

1 ☐ -0.85

2 ☐ 0.85

3* ☐ 2.93

4 ☐ 3.93

5 ☐ 5.43

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

Simply the 15% quantile in the distribution, hence

```
qnorm(0.15, mean=5, sd=2)
```

```
2.93
```

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

Question I.2 (2)

What is the probability that the sample mean $\bar{X} = \frac{1}{25}\sum_{i=1}^{25} X_i$ is greater than 4.5?

1* ☐ $P(\bar{X} > 4.5) = 0.89$

2 ☐ $P(\bar{X} > 4.5) = 0.85$

3 ☐ $P(\bar{X} > 4.5) = 2.05 \times 10^{-10}$

4 ☐ $P(\bar{X} > 4.5) = 0.18$

5 ☐ $P(\bar{X} > 4.5) = 0.55$

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

$\overline{X} \sim N(5, 4/25)$ (Theorem 3.3). The probability is calculated with R by

```
1-pnorm(4.5, mean=5, sd=2/sqrt(25))

## [1] 0.89
```

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

Continue on page 4

Exercise II

Given Lambert Beer's law the absorbance of light through a liquid solution can be calculated as

$$A = \gamma \cdot l \cdot c$$

where γ is a constant, l the path length through the liquid and c the concentration of solution.

Question II.1 (3)

Given that the standard deviation of the path length σ_l and the standard deviation of the concentration σ_c are known, the standard deviation of the absorbance can be approximated by which of the following formulas?

1 ☐ $(\frac{\partial A}{\partial c})^2\sigma_l^2 + (\frac{\partial A}{\partial l})^2\sigma_c^2$

2 ☐ $\sqrt{(\frac{\partial A}{\partial c})^2\sigma_l^2 + (\frac{\partial A}{\partial l})^2\sigma_c^2}$

3 ☐ $(\frac{\partial A}{\partial l})^2\sigma_l^2 + (\frac{\partial A}{\partial c})^2\sigma_c^2$

4* ☐ $\sqrt{(\frac{\partial A}{\partial l})^2\sigma_l^2 + (\frac{\partial A}{\partial c})^2\sigma_c^2}$

5 ☐ $\sqrt{(\frac{\partial A}{\partial c})^2\sigma_l^2} + \sqrt{(\frac{\partial A}{\partial l})^2\sigma_c^2}$

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

See Method 4.3. We take the square root, since we are trying to estimate the standard deviation.

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

Question II.2 (4)

In an experiment the mean path length is determined to be 1 cm with a standard deviation of 0.1 cm. The average concentration is determined to be 0.65 M with a standard deviation of 0.09 M. γ is given as 0.3 $M^{-1}cm^{-1}$. Which of the following simulations can be used to determine the standard deviation of the absorbance?

1 ☐
```
k = 10000
e = 0.3
l = rnorm(k, 1, 0.1)
c = rnorm(k, 0.65, 0.09)
A = e*l*c
var(A)
```

2 ☐
```
e = 0.3
l = rnorm(1, 1, 0.1^2)
c = rnorm(1, 0.65, 0.09^2)
A = e*l*c
sd(A)
```

3 ☐
```
e = 0.3
l = rnorm(1, 1, 0.1^2)
c = rnorm(1, 0.65, 0.09^2)
A = e*l*c
var(A)
```

4 ☐
```
k = 10000
e = 0.3
l = rnorm(k, 1, 0.1^2)
c = rnorm(k, 0.65, 0.09^2)
A = e*l*c
sd(A)
```

5* ☐
```
k = 10000
e = 0.3
l = rnorm(k, 1, 0.1)
c = rnorm(k, 0.65, 0.09)
A = e*l*c
sd(A)
```

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

To get a good estimate of the standard deviation we need to do a large number of simulations and therefore answer 2 and 3 are wrong. In R the rnorm function needs the number of simulations, the mean and the standard deviation (and not the variance), which makes answer 4 wrong too. This leaves us with answer 1 and 5 left, but since we are estimating the standard deviation and not the variance we use sd(A) rather than var(A).

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

Question II.3 (5)

In the question above a random sample from a normal distribution was simulated using the command `rnorm`. Which of the commands below can be used to simulate a random sample from the standard normal distribution of length `n`?

1 ☐
```r
pnorm(runif(n))
```

2* ☐
```r
qnorm(runif(n))
```

3 ☐
```r
dnorm(runif(n))
```

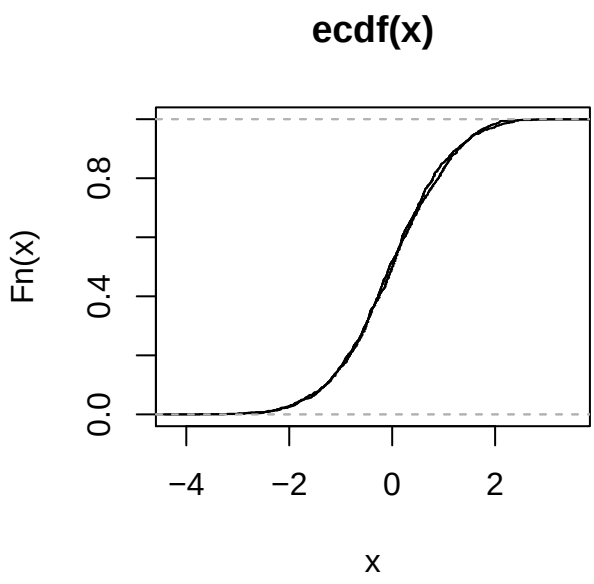
4 ☐
```r
qnorm(punif(n))
```

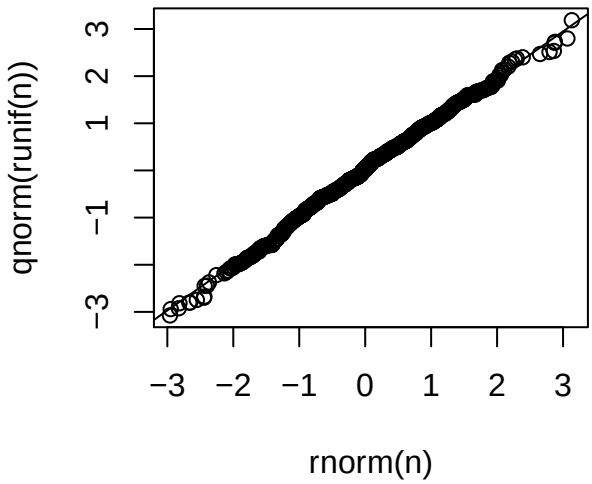
5 ☐
```r
pexp(n)
```

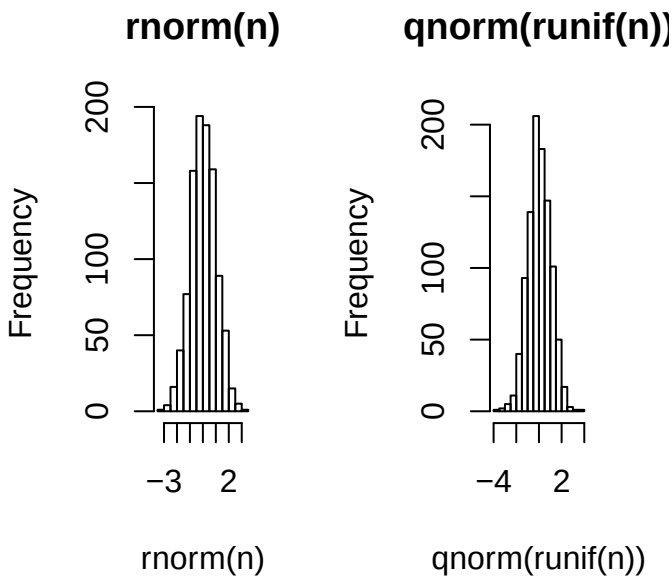
---------- FACIT-BEGIN ----------

See Example 2.52. Example with 1000 observations:

```r
n <- 1000
plot.ecdf(rnorm(n))
plot.ecdf(qnorm(runif(n)), add=TRUE)

qqplot(rnorm(n), qnorm(runif(n)))
qqline(rnorm(n), qnorm(runif(n)))

## Warning in if (datax) {:  the condition has length > 1 and only the first element
will be used

par(mfrow=c(1,2))
hist(rnorm(n), main="rnorm(n)")
hist(qnorm(runif(n)), main='qnorm(runif(n))')
```







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

Continue on page 8

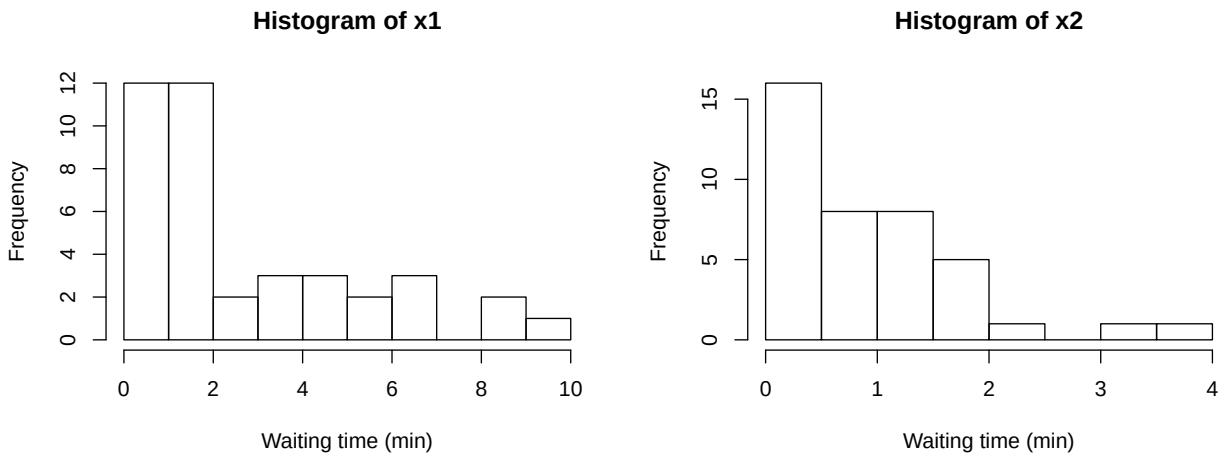
Exercise III

The human ressource department of a supermarket chain is interested in comparing waiting times for customers in two local shops. The waiting times (in minutes) of 40 customers have been measured in the two shops during an afternoon from 4 PM to 5 PM.

Let $X_{1,i}$ represent the i'th observed waiting time in Store 1. It can be assumed to follow an exponential distribution $X_{1,i} \sim Exp(\lambda_1)$ where $i = 1, \ldots, 40$.

Let $X_{2,i}$ represent the i'th observed waiting time in Store 2. It can be assumed to follow an exponential distribution $X_{2,i} \sim Exp(\lambda_2)$ where $i = 1, \ldots, 40$.

The data from each sample is stored in `x1` and `x2`, respectively, and a histogram of each sample is plotted below:



The average waiting times (in minutes) for the two shops are:

```
mean(x1)

## [1] 2.76

mean(x2)

## [1] 0.897
```

Question III.1 (6)

Estimate the rate parameters λ_1 and λ_2. The rates should be calculated in customers per hour (h^{-1}).

1 ☐ $\hat{\lambda}_1 = 2.76\,h^{-1}$ and $\hat{\lambda}_2 = 0.90\,h^{-1}$

2 ☐ $\hat{\lambda}_1 = 0.36\,h^{-1}$ and $\hat{\lambda}_2 = 1.11\,h^{-1}$

3 ☐ $\hat{\lambda}_1 = 19.54\,h^{-1}$ and $\hat{\lambda}_2 = 25.45\,h^{-1}$

4* ☐ $\hat{\lambda}_1 = 21.74\,h^{-1}$ and $\hat{\lambda}_2 = 66.89\,h^{-1}$

5 ☐ It is not possible to estimate λ_1 and λ_2.

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

See Theorem 2.49. The rate parameters can be estimated using $\hat{\lambda} = \frac{1}{\bar{x}}$. To obtain the rate parameter in h^{-1} the result has to be multiplied by 60.

$\hat{\lambda}_1 = \frac{1}{2.760\,\mathrm{min}} \cdot 60\ \mathrm{min\ h^{-1}} = 21.74\,\mathrm{h^{-1}}$ and $\hat{\lambda}_1 = \frac{1}{0.897\,\mathrm{min}} \cdot 60\ \mathrm{min\ h^{-1}} = 66.86\,\mathrm{h^{-1}}$

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

Question III.2 (7)

At two other shops similar samples were collected. The rates were estimated to $\hat{\lambda}_1 = 0.23$ min^{-1} for the first shop and $\hat{\lambda}_2 = 0.39$ min^{-1} for the second.

To find out if there was a significant difference in mean waiting time at two shops the following calculations was carried out in R:

```
k <- 10000
simX1_samples <- replicate(k, rexp(40, 0.23))
simX2_samples <- replicate(k, rexp(40, 0.39))
sim_dif_means <- apply(simX1_samples, 2, mean) - apply(simX2_samples, 2, mean)

quantile(sim_dif_means, c(0.005, 0.995))

##   0.5%  99.5%
## -0.121  3.955

quantile(sim_dif_means, c(0.025, 0.975))

##  2.5% 97.5%
##  0.30  3.42

quantile(sim_dif_means, c(0.05, 0.95))

##    5%   95%
## 0.547 3.156
```

Which of the following statements is correct?

1 ☐ Non-parametric bootstrapping was carried out. The 95% confidence interval is [-0.121, 3.955] and contains zero, hence the mean waiting times are significantly different.

2 ☐ Non-parametric bootstrapping was carried out. The 95% confidence interval is [-0.121, 3.955] and contains zero, hence the mean waiting times are NOT significantly different.

3* ☐ Parametric bootstrapping was carried out. The 95% confidence interval is [0.30, 3.42] and doesn't contain zero, hence the mean waiting times are significantly different.

4 ☐ Parametric bootstrapping was carried out. The 95% confidence interval is [0.30, 3.42] and doesn't contain zero, hence the mean waiting times are NOT significantly different.

5 ☐ Parametric bootstrapping was carried out. The 95% confidence interval is [0.547, 3.156] and doesn't contain zero, hence the mean waiting times are NOT significantly different.

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

The simulation makes an assumption about the underlying exponential distribution, hence parametric bootstrapping is applied. In the correct answer the confidence interval matches the requested one. The confidence interval doesn't contain zero, hence the null hypothesis of the mean waiting times being equal for the two shops can be rejected at the 5% significance level.

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

Continue on page 12

Exercise IV

This exercise is about quality control in a company which produces hard disk drives for NAS ("Network Attached Storage"). The company would like to investigate the probability that a certain type of hard disk drive breaks down within the first three years of "typical use". The company chooses a random sample of 950 hard disk drives from their production line. They ask the customers who buy these drives to report it if a drive fails within the first three years of use. All the NAS hard disk drives are assumed to have the same probability p of failing within the first three years, and they are assumed to fail independently of each other.

Question IV.1 (8)

It was reported that altogether 92 of the hard disk drives failed within the first three years of their lifetime. Give the estimated standard error, $\hat{\sigma}_{\hat{p}}$, for the estimated proportion of hard disk drives which break down within the first three years.

1 ☐ $\hat{\sigma}_{\hat{p}} = 9.2 \cdot 10^{-5}$

2 ☐ $\hat{\sigma}_{\hat{p}} = 0.0031$

3* ☐ $\hat{\sigma}_{\hat{p}} = 0.0096$

4 ☐ $\hat{\sigma}_{\hat{p}} = 0.087$

5 ☐ $\hat{\sigma}_{\hat{p}} = 0.30$

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

The estimated proportion of hard disk drives which break down within three years is

$$\hat{p} = \frac{92}{950}$$

so it follows from equation (7-6) that the standard error may be estimated as

$$\hat{\sigma}_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{92/950 \cdot (1 - 92/950)/950} = 0.0096\,.$$

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

Question IV.2 (9)

The company aims for 90% of their NAS hard disk drives to have a lifetime which exceeds three years. Using a statistical test, they would like to investigate whether they live up to this goal. Which statistical null hypothesis is then relevant to test?

1* ☐ $H_0 : \ p = 0.1$

2 ☐ $H_0 : \ p = 0.9$

3 ☐ $H_0 : \ p \neq 0.1$

4 ☐ $H_0 : \ p \neq 0.9$

5 ☐ None of the above hypotheses are applicable.

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

The company's aim stated above corresponds to 10% of the hard disk drives breaking down within the first three years of use or, put differently, each hard disk drive having a $p = 0.1$ probability of failing within this time period.

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

(The exercise text is continued)
Now, the company would like to compare the lifetime of their special NAS hard disk drives to the lifetime of regular hard disk drives (when these are used in a NAS setup). To this end, they present the following contingency table, which also includes data for the lifetime of 650 regular hard disk drives:

	NAS HDD	Regular HDD	Total
$<$ 1 year	10	7	17
1-2 years	33	45	78
2-3 years	49	69	118
$>$ 3 years	858	529	1387
Total	950	650	1600

This table summarizes how many of a given type of hard disk drive that failed within a certain age interval. For example, one can read from this table that 69 out of 650 regular hard disk drives broke down after 2-3 years of use. These data are to be used in the rest of the questions in this exercise.

Question IV.3 (10)

The company would like to investigate whether the two types of hard disk drives have the same probability of failing within the first three years of their lifetime. Which of the following snippets of R code carries out the relevant statistical test?

1 ☐ `prop.test(x = c(49, 69), n = c(950, 650), correct = FALSE)`

2 ☐ `prop.test(x = c(49, 69), n = c(858, 529), correct = FALSE)`

3 ☐ `prop.test(x = c(92, 950), n = c(121, 650), correct = FALSE)`

4* ☐ `prop.test(x = c(92, 121), n = c(950, 650), correct = FALSE)`

5 ☐ None of the above.

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

See, e.g., the R code in Example 7.19. Note that $10 + 33 + 49 = 92$ out of 950 NAS hard disk drives failed within the first three years of use, while the same was true for $7 + 45 + 69 = 121$ of 650 regular hard disk drives.

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

Question IV.4 (11)

The company could also have chosen to investigate whether the distribution of the number of drive failures in the four age intervals differs for the two types of hard disk drives. Under the corresponding null hypothesis H_0, what is the number of regular hard disk drives which are expected to fail after 1-2 years?

1 ☐ 29

2 ☐ 33

3 ☐ 39

4 ☐ 45

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

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

Following equation 7-53 the correct answer is:

$$\frac{\textit{2nd row total} \cdot \textit{2nd column total}}{\textit{grand total}} = \frac{78 \cdot 650}{1600} = 31.6875 \approx 32$$

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

Question IV.5 (12)

Suppose that the company actually carries out a χ^2-test to investigate whether the distribution of the number of drive failures in the four age intervals differs for the two types of hard disk drives. How many degrees of freedom does the χ^2-distribution, which is used in this test, have?

1 ☐ 1 degree of freedom

2 ☐ 2 degree of freedom

3* ☐ 3 degree of freedom

4 ☐ 6 degree of freedom

5 ☐ 9 degree of freedom

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

See Method 7.22. The contingency table has $r = 4$ rows and $c = 2$ columns, so the distribution in question has

$$(r-1)(c-1) = (4-1)(2-1) = 3$$

degrees of freedom.

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

Continue on page 17

Exercise V

On a small island it is known that the rate of blackouts in the electrical system is one per week. Define the random variable X which denotes the number of blackouts for some randomly chosen week. The number of blackouts per week is assumed to follow a poisson distribution.

Question V.1 (13)

What is the variance of X?

1 ☐ $\sigma^2 = \frac{1}{7}$

2 ☐ $\sigma^2 = 0.368$

3* ☐ $\sigma^2 = 1$

4 ☐ $\sigma^2 = 2.72$

5 ☐ $\sigma^2 = 7$

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

Use Theorem 2.28 to get

$$\sigma^2 = \lambda = 1$$

where lambda is the rate per week.

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

Question V.2 (14)

What is the probability that there will be no blackout on a randomly chosen day?

1 ☐ 0.13

2 ☐ 0.24

3 ☐ 0.53

4 ☐ 0.76

5* ☐ 0.87

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

First scale the rate from per week to per day (see equation 2-33)

$$\lambda_{\text{day}} = \frac{\lambda_{\text{week}}}{7} = \frac{1}{7}$$

and then calculate the probability of no events by

```
dpois(0, 1/7)
```

```
## [1] 0.8668779
```

```
ppois(0, 1/7)
```

```
## [1] 0.8668779
```

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

Continue on page 19

Exercise VI

You would like to compare 5 groups with 6 observations in each. You will do this by making a one-way analysis of variance and test the hypothesis that all groups have same mean value. The observations are assumed to be independent of each other. The test statistic for this test is 4.30.

Question VI.1 (15)

What is the p-value for this test?

1* ☐ 0.009

2 ☐ 0.0002

3 ☐ 0.03

4 ☐ 0.00001

5 ☐ 0.05

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

See theorem 8.6. In this exercise $k = 5$ and $n = 30$) so the calculation becomes:

```
1 - pf(4.30, 5-1, 30-5)

## [1] 0.008766031
```

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

Continue on page 20

Exercise VII

The analysis of variance results from a one-way analysis of variance are:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatm	2	1.19	A	B	C
Residuals	9	4.53	D		

Question VII.1 (16)

As you can see some numbers are missing. The number replaced by B is:

1 ☐ 0.26

2* ☐ 1.18

3 ☐ 1.53

4 ☐ 2.20

5 ☐ 7.40

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

Following Theorem 8.6, we can calculate the F-statistic (since we have both SS(Tr), SSE and the degrees of freedom from the table):

```
(1.19/2) / (4.53/9)

## [1] 1.182119
```

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

Question VII.2 (17)

It is stated that there were equally many observations in each group. How many observations were there in one of the groups?

1 ☐ 2

2 ☐ 3

3* ☐ 4

4 ☐ 5

5 ☐ 9

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

From table 8.2.2 it is seen that the residuals degrees of freedom is $9 = n - k$. And we also know k since treatment degrees of freedom is $2 = k - 1$. This gives us that there are 3 groups and 12 observations in total. So answer 3 is correct.

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

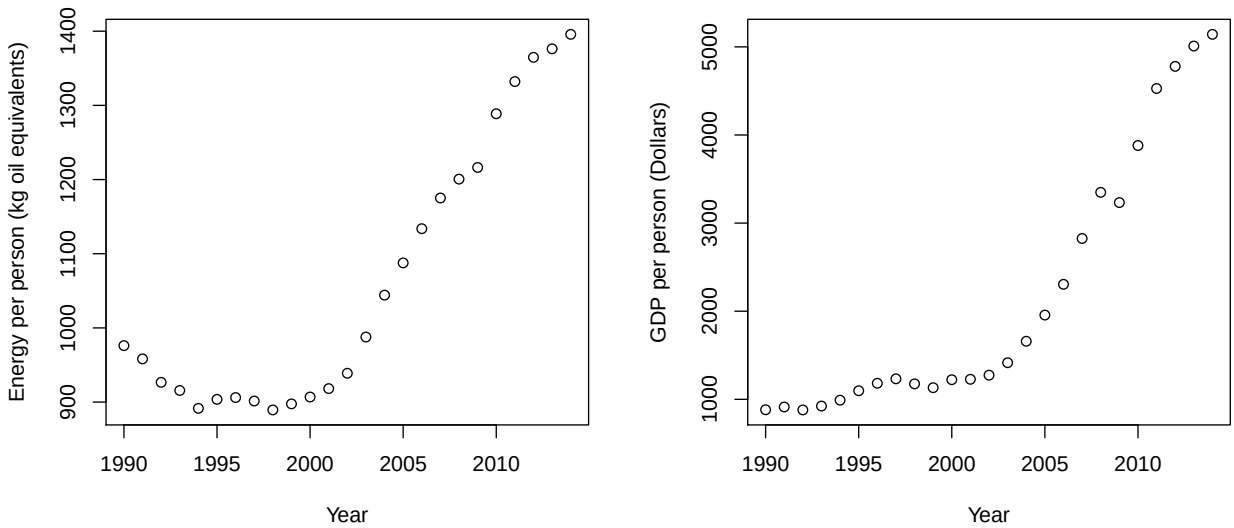
Continue on page 22

Exercise VIII

It is an engineering challenge to develop the technology that can cover the world's energy demand in a sustainable way. Considering The World Bank's population forecasts for 2050 one will reach the result that if everyone in 30 years should have the same energy demand, as the rich countries have now, then the energy demand will triple compared to 2014.

This exercise uses data retrieved from The World Bank, which categorizes the world's countries into the categories: low, middle and high income countries. The development of middle income countries is very important for the development of the world energy demand.

The following plot shows the Energy Consumption and Gross National Product (GDP) per year per person for middle income countries from 1990 to 2014:



The data consists of the plotted annual values stored in the vectors: `year` is the year, `energy` is energy demand and `gdp` is GDP. Only this data is used, thus all conclusions in the exercise apply only to middle income countries in this particular period.

First four summary statistics are calculated:

```
c(mean(energy), mean(gdp))

## [1] 1061 2169

c(sd(energy), sd(gdp))

## [1]  179 1465
```

Thereafter two different simple linear regression models are estimated:

```
summary(lm(energy ~ year))

##
## Call:
## lm(formula = energy ~ year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -122.49  -60.45    3.37   74.54  174.70
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42299.35    4669.35   -9.06  4.8e-09 ***
## year            21.66       2.33    9.29  3.0e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.1 on 23 degrees of freedom
## Multiple R-squared:  0.789,Adjusted R-squared:  0.78
## F-statistic: 86.2 on 1 and 23 DF,  p-value: 3.03e-09

summary(lm(energy ~ gdp))

##
## Call:
## lm(formula = energy ~ gdp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52.82 -29.23  -9.45  27.37  69.09
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.01e+02   1.35e+01    59.5   <2e-16 ***
## gdp         1.20e-01   5.18e-03    23.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.2 on 23 degrees of freedom
## Multiple R-squared:  0.959,Adjusted R-squared:  0.957
## F-statistic:  536 on 1 and 23 DF,  p-value: <2e-16
```

Question VIII.1 (18)

According to these results, what is estimated mean annual increase in energy demand in the period (in "kg oil equivalents" per year)?

1 ☐ 0.120

2 ☐ 2.33

3 ☐ 3.37

4* ☐ 21.7

5 ☐ 801

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

In the first model energy is modelled with year as the explanatory variable. We can read the estimate of the slope from the output from R to be 21.66, so answer 4 is correct.

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

Question VIII.2 (19)

What is the calculated correlation between the energy demand and the GDP?

1 ☐ 0.83

2 ☐ 0.93

3* ☐ 0.98

4 ☐ 1.93

5 ☐ This cannot be calculated with the information given in the exercise.

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

Equation 5-80

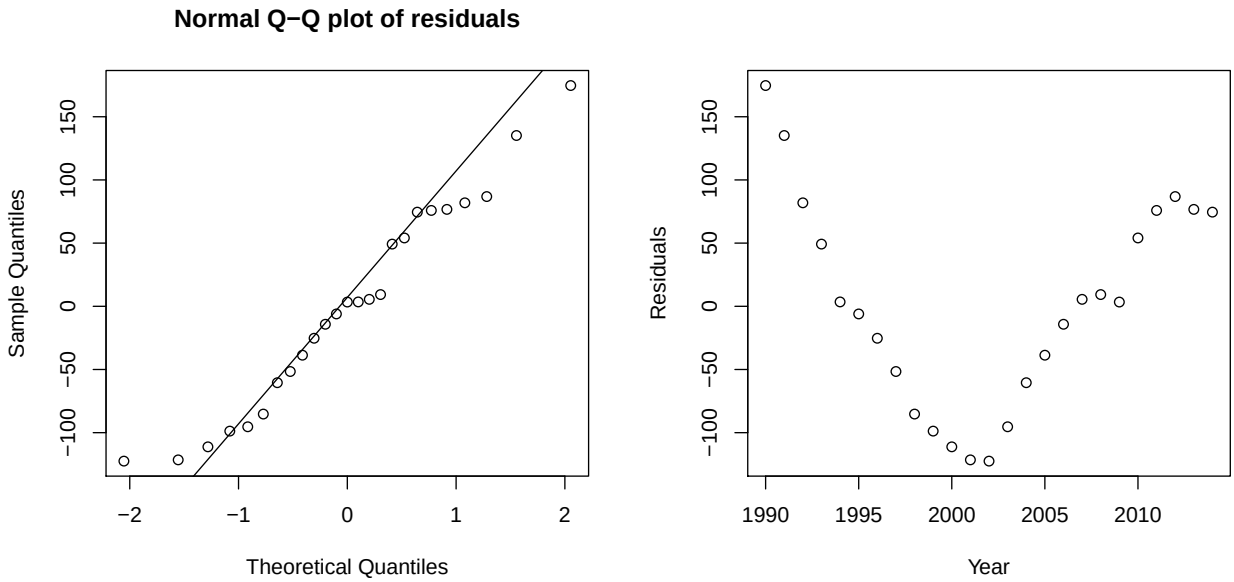
```
1465 / 179 * 0.12

## [1] 0.982

sd(gdp) / sd(energy) * lm(energy ~ gdp)$coef[2]
```

```
##   gdp
## 0.979

cor(energy, gdp)

## [1] 0.979
```

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

Question VIII.3 (20)

The following two plots are generated for the residual analysis of the estimated model between the energy demand and the year:



Which of the following conclusions is most appropriate based on the two plots above (both the conclusion and the argument must be correct)?

1 ☐ The assumption of independent errors should be rejected, since the distribution of the residuals appears to be heavily right skewed.

2 ☐ The assumption of independent errors should be rejected, since the distribution of the residuals appears to be heavily left skewed.

3 ☐ The assumption of independent errors should be rejected, since a clear linear relation can be seen between the residuals and the years.

4* ☐ The assumption of independent errors should be rejected, since a clear non-linear relation can be seen between the residuals and the years.

5 ☐ None of the above conclusions with their associated argument are correct.

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

The V-curve on the scatterplot of the residuals vs the year that there is some non-linear relation between the residuals and the year and therefore we should not assume that the residuals are independent.

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

Question VIII.4 (21)

Are there, according to the book's definition, any extreme observations in the sample consisting of the residuals from the estimated model between the energy demand and the year (both conclusion of argument must be correct)?

1 ☐ Yes, since $-262.9 < 122.5$ and $174.7 < 277.0$.

2* ☐ No, since $-262.9 < 122.5$ and $174.7 < 277.0$.

3 ☐ Yes, since $135.0 < 297.2$.

4 ☐ No, since $135.0 < 297.2$.

5 ☐ Yes, since $0.5 < 0.789$.

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

We find the 1st quartile (Q_1, which is the 25% quantile) and the 3rd quantile, from the print of `summary(lm(energy ~ year))` under `Residuals` (`Q1` and `Q3`), and then

```
IQR <- 74.54 - (-60.45)

1.5 * IQR + 74.54

## [1] 277
```

which is higher than the highest residual at 174.70 (see either the plot of at `Max` in the summary).

Similarly in the low end

```
-60.45 - 1.5 * IQR

## [1] -263
```

is lower than the lowest residuals (`Min`) at -122.49.

Hence no extreme observations.

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

Question VIII.5 (22)

The model is now extended to a multiple linear regression model, using both the year and the GDP as explanatory variables.

The following result is obtained by estimating the model:

```
summary(lm(energy ~ year + gdp))

##
## Call:
## lm(formula = energy ~ year + gdp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52.87 -29.05  -9.28  27.18  69.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.24e+02   4.98e+03    0.13     0.90
## year        8.93e-02   2.50e+00    0.04     0.97
## gdp         1.19e-01   1.26e-02    9.52    3e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38 on 22 degrees of freedom
## Multiple R-squared:  0.959,Adjusted R-squared:  0.955
## F-statistic:  256 on 2 and 22 DF,  p-value: 5.75e-16
```

When comparing the result from the model with only the year as explanatory variable (from the start of the exercise) and the result of the model with both the year and GDP, the following "absurd" conclusion can be drawn for the hypothesis of a dependence between year and energy demand:

There is very strong evidence of the hypothesis when the year alone is used as explanatory variable, while there is little or no evidence when both year and GDP are used.

However, this result is by no means absurd statistically, as it often can occur if the following is true:

1 ☐ GDP is decreasing in the period.

2 ☐ There is a relatively high non-linear relationship between the year and the energy demand in the observed data.

3 ☐ There is a relatively high non-linear relationship between the year and the GDP demand in the observed data.

4 ☐ There is a relatively high correlation between the year and the energy demand in the observed data.

5* ☐ There is a relatively high correlation between the year and the GDP demand in the observed data.

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

This is an example of collinearity, where the explanatory variables are highly correlated, see Section 6.3 in the book.

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

Continue on page 30

Exercise IX

In a study of two types of pig feeds, 20 pigs was divided into two (smaller) groups (x: Group 1 with 8 and Group 2 with 12 pigs). Those two groups received from the age of 3 months until they were slaughtered (6 months) each a different type of feed. The table below shows the pigs weight when slaughtered (kg):

x	113.3	117.9	111.9	109.6	109.6	111.5	97.8	103.3				
y	110.7	108.3	110.6	106.7	109.7	107.5	105.9	111.0	99.9	110.2	99.4	103.6

The following was calculated $\bar{x} = 109.4$, $\bar{y} = 107.0$, $s_x^2 = 6.2^2$ and $s_y^2 = 4.1^2$. It can be assumed that the weight when they were slaugtered followed a normal distribution in each group. Further, the pooled variance was calculated to $s_p^2 = 5.0^2$.

Question IX.1 (23)

What is the 95% confidence interval for the mean weight of the pigs from Group 1 when slaughtered?

1* ☐ $[104.2, 114.6]$

2 ☐ $[105.2, 113.6]$

3 ☐ $[107.6, 111.2]$

4 ☐ $[101.7, 117.1]$

5 ☐ $[106.6, 112.2]$

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

This is a standard confidence interval for one sample (see Method 3.9).

$$\bar{x} \pm t_{1-0.95/2} * \frac{s}{\sqrt{n}}$$

In R:

```
mean_x <- 109.4
sigma_x <- 6.2
n <- 8
mean_x + c(-1, 1) * qt(0.975, n-1) * (sigma_x / sqrt(n))

## [1] 104.2 114.6
```

Or just type everything into R

```r
x <- c(113.3, 117.9, 111.9, 109.6, 109.6, 111.5, 97.8, 103.3)
t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 50, df = 7, p-value = 3e-10
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  104.2 114.6
## sample estimates:
## mean of x
##     109.4
```

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

Question IX.2 (24)

A 99% confidence interval for the variance of the weight in Group 1 is wanted. How is this calculated correctly?

1* ☐ $\left[\frac{7 \cdot 6.2^2}{20.3}, \frac{7 \cdot 6.2^2}{1.0}\right]$

2 ☐ $\left[\frac{8 \cdot 6.2}{20.3}, \frac{8 \cdot 6.2}{1.0}\right]$

3 ☐ $\left[\frac{9 \cdot 6.2}{20.3}, \frac{9 \cdot 6.2}{1.0}\right]$

4 ☐ $\left[\frac{8 \cdot 6.2^2}{20.3}, \frac{8 \cdot 6.2^2}{1.0}\right]$

5 ☐ $\left[\frac{7 \cdot 6.2}{20.3}, \frac{7 \cdot 6.2}{1.0}\right]$

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

See Method 3.19.

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

The chi-squared quantiles can be found in R as

```r
qchisq(0.995, 8-1)

## [1] 20.28

qchisq(0.005, 8-1)

## [1] 0.9893
```

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

Question IX.3 (25)

When testing for the difference in mean slaughter weight between Group 1 and Group 2, what is the result of the usual Welch test statistics?

1* ☐ $|t_{\text{obs}}| = 0.96$

2 ☐ $|t_{\text{obs}}| = 1.0$

3 ☐ $|t_{\text{obs}}| = 2.6$

4 ☐ $|t_{\text{obs}}| = 49.8$

5 ☐ $|t_{\text{obs}}| = 90.8$

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

One could either use the equation given in method 3.49 the test-statistic and use the numbers given in the exercise or type everything into R:

```
x <- c(113.3, 117.9, 111.9, 109.6, 109.6, 111.5, 97.8, 103.3)
y <- c(110.7, 108.3, 110.6, 106.7, 109.7, 107.5, 105.9, 111.0, 99.9, 110.2, 99.4, 103.6)
t.test(x, y)

##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 1, df = 11, p-value = 0.4
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.1  7.9
## sample estimates:
## mean of x mean of y
##       109       107
```

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

Question IX.4 (26)

If, in a new experiment, it is wanted to obtain a strength of 80% to be able to detect one difference of 4 kg between the two groups of on a confidence level of 99%, and the weighted variance is used as a guess of the population's variance, how many pigs should at least be included in this experiment?

1 ☐ 22

2 ☐ 42

3 ☐ 52

4* ☐ 78

5 ☐ 104

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

See example 3.67. The number of pigs in each group can be calculated by entering the 4 other values:

```
power.t.test(delta=4, sd=5.0, sig.level=0.01, power=0.8)$n

## [1] 38.19
```

which rounded up means that there must be 39 pigs in each group and thus in total there must be 78 pigs included in the experiment.

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

Continue on page 35

Exercise X

The following sample has been collected and sorted:

1	2	3	4	5	6	7	8	9	10	11	12
0.4	1.0	1.4	2.0	2.5	2.7	3.0	3.3	4.2	4.5	6.5	7.6

Question X.1 (27)

What is the median of the sample?

1 ☐ 2.6

2 ☐ 2.7

3* ☐ 2.85

4 ☐ 3.0

5 ☐ 3.3

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

The easiest is probably just to type in the vector in R and as for the median. Alternative one can do it by hand as described in Definition 1.5.

```
median(x, type=2)

## [1] 2.85
```

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

Question X.2 (28)

If the sample is stored in the vektor `x` in R, which of the following calls is calculating the standard deviation of the sample?

1* ☐ `sqrt(sum((x-mean(x))^2)/(length(x)-1))`

2 ☐ `sd(x)*length(x)`

3 ☐ `sqrt(sd(x))`

4 ☐ `var(x)^2/length(x)`

5 ☐ `sd(x)^2/length(x)`

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

See Definition 1.11. Alternatively you can try in R and see which gives you the correct result:

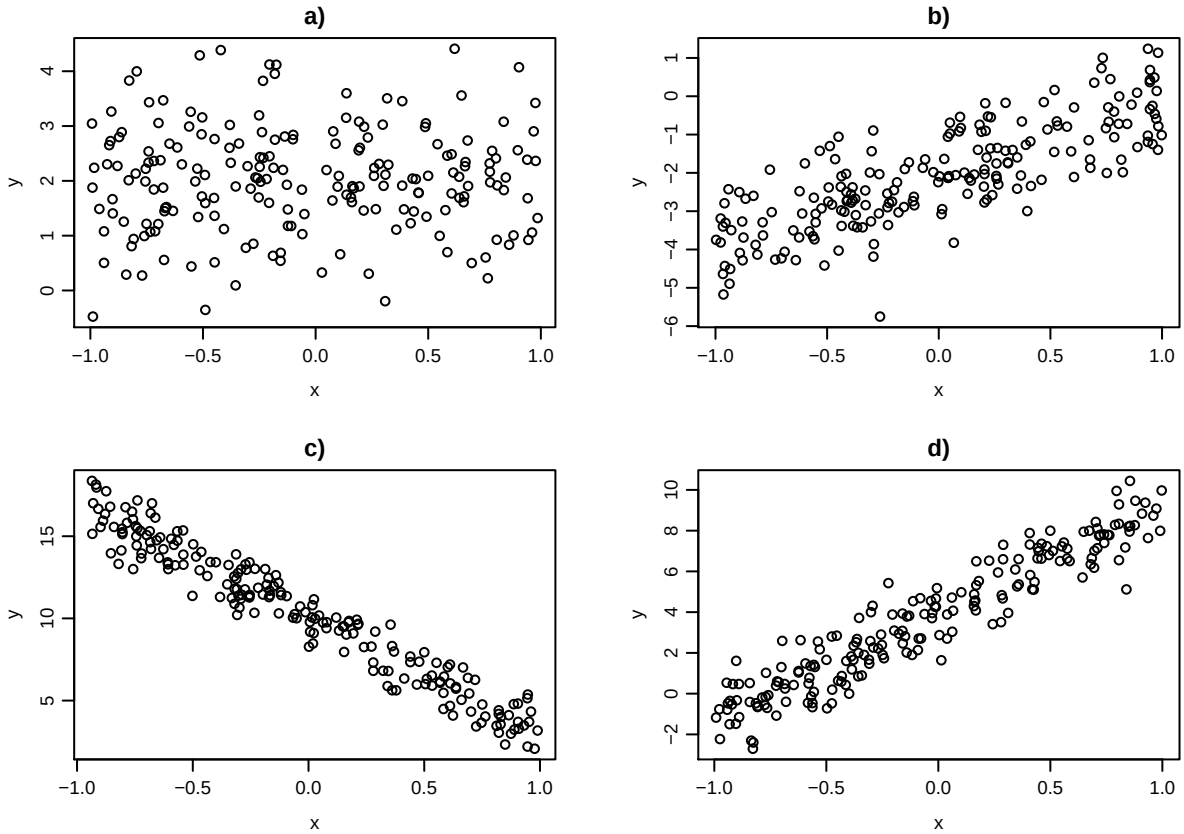
```
sd(x)

## [1] 2.15

sqrt(sum((x-mean(x))^2)/(length(x)-1))

## [1] 2.15
```

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

Continue on page 37

Exercise XI

Below are four scatter plots of y and x observations:



Question XI.1 (29)

Which four correlation coefficients (in the order: `a), b), c), d)`) fits best with the observations in the figure?

1 ☐ 0.02, 0.79, 0.95, -0.97

2 ☐ 0.02, 0.95, 0.79, -0.97

3 ☐ -0.97, 0.02, 0.79, 0.95

4 ☐ 0.02, 0.95, -0.97, 0.79

5* ☐ 0.02, 0.79, -0.97, 0.95

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

The a) plot has close to zero correlation, since almost no linear dependence is seen. So from the possible values in the answers its 0.02.

The b) its positive, but lower than d), so must be 0.79.

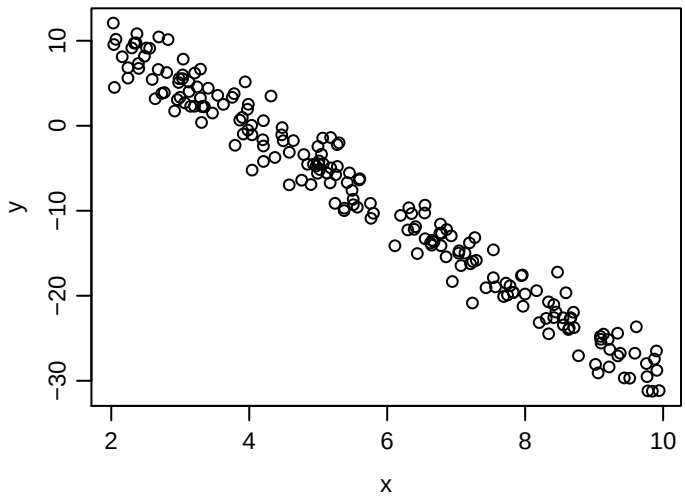
The c) its negative, so must be -0.97.

The d) its positive and stronger than b), so must be 0.95.

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

Question XI.2 (30)

Another sample of x and y data is plotted below:



A linear regression is carried out on the values in the plot with the R code

```
summary(lm(y ~ x))
```

and the result for the coefficients estimates from the summary is:

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)      A       0.0716   138.3   <2e-16 ***
## x                B       0.1236   -54.9   <2e-16 ***
```

The estimated values of the coefficients have been replaced with letters.

Which of the following answers is the only which is not with very unlikely?

1 ☐ A is 10 and B is -2.

2* ☐ A is 20 and B is -5.

3 ☐ A is 4 and B is -5.

4 ☐ A is 4 and B is -2.

5 ☐ A is 10 and B is -8.

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

A is the intercept with the y axis. It looks like this will be around 20 (notice that the x axis starts at 2 and not 0). In the same manner it can be seen that everytime x increases by 2, y decreases by approximately 10, so the slope must be around -5 (which also matches with the intercept at 20).

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

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

Technical University of Denmark *Page 1 of 31 pages.*

Written examination: 15. December 2019

Course name and number: **Introduction to Statistics (02323)**

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

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

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

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

Exercise	VI.1	VI.2	VII.1	VII.2	VII.3	VII.4	VII.5	VII.6	VIII.1	VIII.2
Question	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Answer	1	3	5	2	1	1	4	4	2	4

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

The exam paper contains 31 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 biodynamic farm wants to degrade its biomass residues into bio liquid to be used for renewable energy production. In an experiment, the farmers used 10 liter reaction containers to assess the efficiency of the biomass conversion. Varying amounts of an enzymatic cocktail were added to each of the containers, and the mixtures were left for three days of reaction time. Afterwards, the volumes of produced bio liquid were determined.

A simple linear regression model of the form $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ was established, in order to investigate the relationship between the amount of enzyme added (`enzyme`, in ml) and the bio liquid yield (`liquid`, in dl). The R output from fitting the model can be seen below:

```
##
## Call:
## lm(formula = liquid ~ enzyme)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8103 -4.3885 -0.0775  4.3672  9.2489
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.325      2.291   4.070 0.000653 ***
## enzyme         1.956      0.196   9.982 5.42e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.438 on 19 degrees of freedom
## Multiple R-squared:  0.8398,Adjusted R-squared:  0.8314
## F-statistic: 99.64 on 1 and 19 DF,  p-value: 5.419e-09
```

Question I.1 (1)

Given the R output above, what is the sample size n?

1 ☐ 20

2 ☐ 19

3* ☐ 21

4 ☐ 1

5 ☐ The sample size cannot be determined from this R output.

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

For a simple linear regression model the degrees of freedom are given by $df = n - 2$. Using the R output above, we can therefore conclude that the sample size is $n = 21$.

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

Question I.2 (2)

In the experiment, the average amount of enzymatic cocktail used in a reaction container was $\bar{x} = 10$ ml. Compute the average bio liquid yield, $\bar{y}$.

1* ☐ $\bar{y} = 28.9$ dl

2 ☐ $\bar{y} = 9.3$ dl

3 ☐ $\bar{y} = 2.0$ dl

4 ☐ $\bar{y} = 19.6$ dl

5 ☐ $\bar{y} = 24.1$ dl

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

It follows from (5-10) that $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, so the average bio liquid yield may be computed as

$$\bar{y} = 9.325 + 1.956 \cdot 10 = 28.9\,\text{dl}$$

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

Question I.3 (3)

Which of the statements below does not represent a necessary assumption for a simple linear regression model?

1 ☐ The errors ε_i are independent.

2 ☐ The errors ε_i are identically distributed.

3* ☐ The outcomes Y_i are identically distributed.

4 ☐ The outcomes Y_i are independent.

5 ☐ The outcomes Y_i and the errors ε_i have the same variance.

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

See, e.g., the first section of Chapter 5. The expected outcome $E(Y_i)$ depends on the value of x_i so the distribution of Y_i depends on the value of x_i (whenever $\beta_1 \neq 0$).

The model can also be written as

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

for $i = 1, \ldots, n$ with the assumption that the outcomes Y_i are independent.

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

Continue on page 5

Exercise II

In connection with the examination in an introductory statistics course, one wants to examine whether students, who have been enrolled in the study program for one year, perform differently than students who have been enrolled for two years. The exam score is calculated as a number between -30 and 150 by the rules:

- there are 30 questions in total,
- -1 point is given for a wrong answer,
- 5 points are given for a correct answer,
- only one answer can be given to each question.

Two samples consisting of exam scores have been collected randomly from the students: One from students who are in their first year (`x`), and one from students who are in their second year (`y`).

The samples each contains 50 observations, and their means are $\bar{x} = 84.0$ and $\bar{y} = 86.6$, respectively. The following simulations and calculations are carried out in R:

```
k <- 10000

simxsamples <- replicate(k, sample(x, replace = TRUE))
simysamples <- replicate(k, sample(y, replace = TRUE))
simmeandifs <- apply(simxsamples, 2, mean) - apply(simysamples, 2, mean)

quantile(simmeandifs, c(0.05, 0.95))

##     5%    95%
## -15.12   9.87

quantile(simmeandifs, c(0.025, 0.975))

##   2.5%  97.5%
## -17.26  12.42

quantile(simmeandifs, c(0.005, 0.995))

##   0.5%  99.5%
## -22.04  17.32
```

Question II.1 (4)

The null hypothesis

$$H_0 : \mu_X = \mu_Y$$

is to be tested at significance level $\alpha = 5\%$, without making assumptions about the distribution of the scores in the two samples. Which of the following answers is correct? (Both the conclusions and argument must hold).

1* ☐ The null hypothesis is not rejected, as $0 \in [-17.26, 12.42]$. Hence, a significant difference cannot be detected.

2 ☐ The null hypothesis is not rejected, as $2.6 \in [-15.12, 9.87]$. Hence, a significant difference cannot be detected.

3 ☐ The null hypothesis is rejected, as $0 \in [-17.26, 12.42]$. Hence, it can be established that students in their first year perform better that students in their second year.

4 ☐ The null hypothesis is rejected, as $0 \notin [-22.04, 17.32]$. Hence, it can be established that students in their first year perform better that students in their second year.

5 ☐ The null hypothesis is not rejected, as $2.6 \in [-22.04, 17.32]$. Hence, it can be established that students in their second year perform better that students in their first year.

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

As 0 is contained in the 95% bootstrap confidence interval, we must accept the hypothesis $\mu_X - \mu_Y = 0$.

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

Continue on page 7

Exercise III

In a hospital, a group of patients are randomly selected. They receive a questionnaire about the hospital's service, both when they are admitted and when they leave the hospital. In both questionnaires, the patients are asked to indicate their satisfaction with the hospital's service on a continuous scale from 0 to 1. Subsequent analysis of data reveals that both series of measurements of service satisfaction can be assumed to be normally distributed. Which of the following 5 tests is most suitable for a comparison of the service assessment upon hospitalization and when leaving the hospital?

Question III.1 (5)

1 ☐ A χ^2-test in a contingency table

2 ☐ A one-way analysis of variance

3 ☐ A t-test with two independent samples

4* ☐ A paired t-test

5 ☐ A regression analysis

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

See Section 3.2.3 for an explanation of the paired t-test and when it is applicable.

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

Continue on page 8

Exercise IV

Question IV.1 (6)

Assume that the random variable $X \in [0, 1]$ follows a distribution with density function $f(x) = 2x$ for $x \in [0, 1]$, and thus has the distribution function $F(x) = x^2$. Which of the following pieces of R code simulates outcomes of the random variable X?

1 ☐ `2 * runif(k)`

2 ☐ `rchisq(k, df = 1)`

3 ☐ `runif(k)^2`

4 ☐ `rchisq(k, df = k - 1)`

5* ☐ `sqrt(runif(k))`

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

See Theorem 2.51, and note that $F^{-1}(u) = \sqrt{u}$ is the inverse of $F(x) = x^2$ on the interval $[0, 1]$.

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

Continue on page 9

Exercise V

A plastic manufacturer wants to determine if there is a difference in the quality of plastic produced with materials from different suppliers (`Supplier`). In the production, a particular measured variable Y (`y`) is known to determine the quality of the produced plastic. Higher values of Y indicate higher quality of the produced plastic. The table below shows values of Y collected from separate production runs with materials from 5 different suppliers. Subsequently, output from the analysis that was run in R by the company's engineers is shown.

Supplier A	Supplier B	Supplier C	Supplier D	Supplier E
9.9	8.7	8.3	10.4	7.7
10.5	10.3	10.7	12.1	11.7
8.2	6.1	8.7	11.5	10.1
7.7	7.6	9.5	11.2	9.0

```
anova(lm(y ~ Supplier))

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value Pr(>F)
## Supplier   4   20.9    5.23    2.77  0.066 .
## Residuals 15   28.3    1.89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question V.1 (7)

Given the model used in the analysis, what is the estimate of the expected value $\mathrm{E}(Y_{\mathrm{D},i})$ for supplier D?

1 ☐ 10.3

2 ☐ 10.5

3 ☐ 10.8

4* ☐ 11.3

5 ☐ 11.5

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

The model is a one-way ANOVA, so the estimate in question is simply the average of the observations in the "supplier D" group. In R:

```r
mean(c(10.4, 12.1, 11.5, 11.2))

## [1] 11.3
```

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

Question V.2 (8)

Which of the following answers most accurately describes the hypothesis tested in the R output above?

1 ☐ The hypothesis $\alpha_i = 1$ for all $i = 1, 2, 3, 4, 5$, in a model of the form $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$.

2* ☐ The hypothesis $\alpha_i = 0$ for all $i = 1, 2, 3, 4, 5$, in a model of the form $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$.

3 ☐ The hypothesis $\beta_0 = 1$ in a model of the form $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

4 ☐ The hypothesis $\beta_1 = 0$ in a model of the form $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

5 ☐ None of the above answers describe the hypothesis that is tested.

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

The model is a one-way ANOVA, and the test investigates whether there is a significant difference in mean between the groups. Since α_i is the difference from the global mean (the mean without any grouping), then the null hypothesis is

$$H_0 : \alpha_i = 0 \text{ for all } i$$

See the section around Theorem 8.6.

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

Question V.3 (9)

Use the significance level $\alpha = 5\%$. Is there a significant difference in the quality of the plastic produced with the materials from the 5 suppliers? (Both the conclusion and argument must hold).

1* ☐ A significant difference in quality cannot be detected, as the p-value is above the significance level.

2 ☐ A significant difference in quality can be detected, as the p-value is under the significance level.

3 ☐ A significant difference in quality cannot be detected, as the p-value is under the significance level.

4 ☐ A significant difference in quality can be detected, as the p-value is above the significance level.

5 ☐ None of the above conclusions are correct.

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

The relevant p-value is 0.066 (found in the R output), which is larger than the significance level of 0.05. Thus, no significant difference is established. See the ANOVA table and the example below Theorem 8.6.

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

Question V.4 (10)

How much of the total variation cannot be explained by the model?

1 ☐ $\frac{16.7}{28.3+20.9} = 33.9\%$

2 ☐ $\frac{20.9}{28.3+20.9} = 42.5\%$

3 ☐ 6.6%

4* ☐ $\frac{28.3}{28.3+20.9} = 57.5\%$

5 ☐ $\frac{1.89}{1.89+5.23} = 26.5\%$

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

See Chapter 8 for the notation. $SST = 28.3+20.9$ expresses the total variation, while $SS(Tr) = 20.9$ represents the variation explained by the model, and $SSE = 28.3$ describes the remaining (unexplained) variation.

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

Continue on page 12

Exercise VI

The value of X has been measured for 5 individuals in Group 1 and 10 individuals in Group 2, respectively. It can be assumed that the observations in both groups are normally distributed and that all the observations are mutually independent. The variances in the two groups are allowed to be different. One would like to test the hypothesis that the two groups have the same mean (against the alternative that the means are different). The test is performed at a 5% significance level.

Question VI.1 (11)

By a comparison with the usual test statistic, which of the following quantiles can easily be used in order to determine whether there is a significant difference between the two means?

1* ☐ The 0.025 quantile of the relevant t-distribution.

2 ☐ The 0.05 quantile of the relevant t-distribution.

3 ☐ The 0.95 quantile of the standard normal distribution.

4 ☐ The 0.90 quantile of the standard normal distribution.

5 ☐ The 0.50 quantile of the standard normal distribution.

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

See Method 3.51. The test is two-sided, so with a significance level of 5%, the difference between means is significant if the t-test statistic falls outside the interval $[t_{0.025}; t_{0.975}]$. Here, $t_{0.025}$ and $t_{0.975}$ respectively denote the 0.025 and 0.975 quantiles of the relevant t-distribution. As all t-distributions are symmetric (i.e. $-t_{0.025} = t_{0.975}$), one can simply check whether the test statistic falls outside the interval $[t_{0.025}; -t_{0.025}]$.

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

Question VI.2 (12)

The sample mean and standard deviation in Group 1 are, $\bar{x}_1 = 1.99$ and $s_1 = 0.58$, while the corresponding numbers for Group 2 are, $\bar{x}_2 = 1.14$ and $s_2 = 0.84$. The variances in the two groups are assumed to be different. In this case, the test statistic for the above test is:

1 ☐ $t_{\text{obs}} = 4.3$

2 ☐ $t_{\text{obs}} = 1.9$

3* ☐ $t_{\text{obs}} = 2.3$

4 ☐ $t_{\text{obs}} = 6.2$

5 ☐ None of the above possibilities.

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

See Method 3.49. The Welch two-sample t-test statistic may be computed as:

$$t_{\text{obs}} = \frac{1.99 - 1.14}{\sqrt{0.58^2/5 + 0.84^2/10}} = 2.3\,.$$

In R:

```
x1 <- 1.99; x2 <- 1.14
s1 <- 0.58; s2 <- 0.84
n1 <- 5; n2 <- 10
(x1-x2)/sqrt(s1^2/n1+s2^2/n2)

## [1] 2.289451
```

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

Continue on page 14

Exercise VII

A research project involves collecting insects by driving predefined trips with a net on the roof of a car. After the trip, the collected insects are sent to the university for counting.

Question VII.1 (13)

Which of the following distributions is presumably best for describing the number of insects in a net?

1 ☐ An exponential distribution

2 ☐ A binomial distribution

3 ☐ A normal distribution

4 ☐ A hypergeometric distribution

5* ☐ A Poisson distribution

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

See Section 2.3.3. The Poisson distribution may be used to describe the probability of a given number of insects being collected during a pre-specified trip, under certain assumptions.

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

Question VII.2 (14)

A total of four trips are planned on the same stretch of road, and it is assumed that the variance of the number of insects, σ^2, is the same on each of the four trips. Furthermore, the results of the four trips are assumed to be independent. What is the variance of the total number of insects captured on the four trips?

1 ☐ $16\sigma^2$

2* ☐ $4\sigma^2$

3 ☐ $\sigma^2/4$

4 ☐ 4σ

5 ☐ $\sigma^2/2$

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

See Theorem 2.56. Let X_i describe the number of insects caught during Trip i ($i = 1, 2, 3, 4$) and X describe the total over all four trips. Then

$$V(X) = V(X_1 + X_2 + X_3 + X_4) = V(X_1) + V(X_2) + V(X_3) + V(X_4) = 4\sigma^2 .$$

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

The four planned trips are carried out, and the captured insects divided into two types: small and large insects. The result of the counting is shown in the contingency table below.

	Trip 1	Trip 2	Trip 3	Trip 4	Total
Small insects	178	242	126	87	633
Large insects	26	59	30	8	123
Total	204	301	156	95	756

Question VII.3 (15)

Looking at the overall result (all four trips combined), which of the following is a 95% confidence interval for the proportion of large insects?

1* ☐ $[0.14; 0.19]$

2 ☐ $[0.81; 0.87]$

3 ☐ $[0.16; 0.23]$

4 ☐ $[0.09; 0.23]$

5 ☐ $[0.81; 0.86]$

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

See Method 7.3. The total number of observed insects is $n = 756$, and the estimated proportion of large insects is

$$\hat{p} = \frac{123}{756} = 0.1626984 .$$

With $z_{0.975} = 1.959964$ being the 0.975 quantile of the standard normal distribution, the confidence interval is given as

$$0.1626984 \pm 1.959964 \cdot \sqrt{\frac{0.1626984 \cdot (1 - 0.1626984)}{756}} = [0.14, 0.19] .$$

In R:

```r
p.hat <- 123/756
n <- 756
p.hat + c(-1,1) * qnorm(0.975)*sqrt(p.hat*(1-p.hat)/n)

## [1] 0.1363885 0.1890083
```

or see `95 percent confidence interval` in the output below.

```r
prop.test(x = 123, n = 756, correct = FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  123 out of 756, null probability 0.5
## X-squared = 344.05, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1381050 0.1907024
## sample estimates:
##         p
## 0.1626984
```

(Note that both methods give the same result when rounded to the correct number of decimals).

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

Question VII.4 (16)

There are special reasons for examining whether the proportion of large insects can be assumed to be the same on Trip 1 and Trip 2. What is the p-value and the conclusion at significance level $\alpha = 5\%$, for a test investigating whether the proportion of large insects differs between Trip 1 and Trip 2?

1* ☐ The p-value is 0.043, and a difference can therefore be established.

2 ☐ The p-value is 0.03 and a difference can therefore be established.

3 ☐ The p-value is 0.060 and a difference can therefore be established.

4 ☐ The p-value is 0.043 and therefore no difference can be established.

5 ☐ The p-value is 0.060 and therefore no difference can be established.

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

See Method 7.18. In R:

```
p.hat <- (26+59)/(204+301)
p1.hat <- 26/204
p2.hat <- 59/301
z <- (p1.hat-p2.hat)/sqrt(p.hat*(1-p.hat)*(1/204+1/301))
2*(1-pnorm(abs(z)))

## [1] 0.04331396
```

or

```
prop.test(c(26,59), c(204,301), correct = FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(26, 59) out of c(204, 301)
## X-squared = 4.0831, df = 1, p-value = 0.04331
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.132635303 -0.004489314
## sample estimates:
##    prop 1    prop 2
## 0.1274510 0.1960133
```

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

In the following questions, we look at data from all four trips, in order to test whether the distribution between large and small insects can be assumed to be the same on all trips.

Question VII.5 (17)

In order to perform the statistical test, the expected number of insects in each cell, under the null hypothesis, must be calculated. What is the expected number of large insects on Trip 3?

1 ☐ 130.6

2 ☐ 4.9

3 ☐ 105.5

4* ☐ 25.4

5 ☐ 32.1

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

See, e.g., Example 7.23.

$$e_{23} = \text{third column total} \cdot \frac{\text{second row total}}{\text{grand total}} = 156 \cdot \frac{123}{756} = 25.4\,.$$

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

Question VII.6 (18)

The usual test statistic for examining the difference between the distribution of the number of insects on the four trips is calculated to be 9.6127. What is the p-value and the corresponding conclusion at significance level $\alpha = 5\%$?

1 ☐ The p-value is 0.022, so no difference can be detected.

2 ☐ The p-value is 0.087, hence there is a difference.

3 ☐ The p-value is 0.087, so no difference can be detected.

4* ☐ The p-value is 0.022, hence there is a difference.

5 ☐ The p-value is 0.045, hence there is a difference.

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

See Method 7.22. Here, $r = 2$ and $c = 4$, so the χ^2-distribution with $(r-1)\cdot(c-1) = 3$ degrees of freedom should be used. The p-value may be computed "by hand" as

```
1 - pchisq(9.6127, df = 3)

## [1] 0.02216216
```

or found in the following output:

```
M <- matrix(c(178, 26, 242, 59, 126, 30, 87, 8), nrow = 2)
chisq.test(M)
```

```
##
##  Pearson's Chi-squared test
##
## data:  M
## X-squared = 9.6127, df = 3, p-value = 0.02216
```

As the p-value is smaller than 0.05, there is a significant difference between the four trips.

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

Continue on page 20

Exercise VIII

The amount of detergent necessary for washing laundry typically depends on several factors. In this context, the relationship between washing efficiency (`efficiency`), water hardness (`hardness`), and the amount of detergent used (`detergent`) is to be investigated using the following multiple linear regression model:

$$efficiency_i = \beta_0 + \beta_1 \cdot hardness_i + \beta_2 \cdot detergent_i + \varepsilon_i \,,$$

where the ε_i are independent and $N(0, \sigma^2)$-distributed. R output from the model is shown below:

```
##
## Call:
## lm(formula = efficiency ~ hardness + detergent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9022 -1.4491 -0.5854  1.4225  5.3286
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.5892     3.6590  -0.434   0.6695
## hardness     -2.1981     0.8958  -2.454   0.0252 *
## detergent     3.0239     0.4961   6.095 1.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.496 on 17 degrees of freedom
## Multiple R-squared:  0.7322,Adjusted R-squared:  0.7006
## F-statistic: 23.23 on 2 and 17 DF,  p-value: 1.371e-05
```

Question VIII.1 (19)

Look at the R output above. Which of the following statements is correct, given a significance level of $\alpha = 1\%$?

1 ☐ Water hardness appears to have a significant effect on washing efficiency, while the amount of detergent used does not.

2* ☐ The effect of water hardness on washing efficiency is not significant, because the p-value is greater than 0.01.

3 ☐ Both water hardness and the amount of detergent used are significant, because the p-values are less than 0.05.

4 ☐ Neither water hardness nor the amount of detergent used appear to be significant, because the p-values are less than 0.05.

5 ☐ The model intercept is significant, because the p-value of 0.6695 is greater than 0.01.

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

Given the significance level, p-values have to be smaller than 0.01 for significance.

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

Question VIII.2 (20)

Look at the same R output above. What effect does an increase of two units of detergent have on expected washing efficiency? Assume water hardness to be constant.

1 ☐ The expected washing efficiency increases by 3.02 units.

2 ☐ The expected washing efficiency decreases by 2.20 units.

3 ☐ The expected washing efficiency decreases by 4.40 units.

4* ☐ The expected washing efficiency increases by 6.05 units.

5 ☐ The expected washing efficiency remains constant.

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

The estimated slope for detergent is $\beta_2 = 3.0239$. A two unit increase of detergent will lead to an expected increase of washing efficiency of $2 \cdot 3.0239 = 6.0478$ (hardness is kept constant).

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

Question VIII.3 (21)

Give an estimate of the variance σ^2 based on the R output above.

1 ☐ $\hat{\sigma}^2 = 23.23$

2 ☐ $\hat{\sigma}^2 = 0.7006$

3 ☐ $\hat{\sigma}^2 = 2.496$

4 ☐ $\hat{\sigma}^2 = 0.7322$

5* ☐ $\hat{\sigma}^2 = 6.230$

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

In the R-output, `Residual standard error` is an estimate of the standard deviation σ, so

$$\hat{\sigma}^2 = 2.496^2 = 6.230016\,.$$

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

Continue on page 23

Exercise IX

The temperature in a refrigerator was measured at 12 o'clock on randomly selected days during the month of July. The following observations were measured (in degrees celsius) and loaded into R in the vector `x`:

```
x <- c(6.5, 5.7, 1.2, 0.2, 7.0, 3.3)
```

The observations are assumed to be normally distributed and mutually independent.

Question IX.1 (22)

Compute the usual test statistic for testing the hypothesis that the mean is 3.0 degrees.

1* ☐ $t_{\text{obs}} = 0.84$

2 ☐ $t_{\text{obs}} = 0.20$

3 ☐ $t_{\text{obs}} = 2.41$

4 ☐ $t_{\text{obs}} = 3.01$

5 ☐ $t_{\text{obs}} = 1.99$

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

See Equation (3-21). In R:

```
x <- c(6.5, 5.7, 1.2, 0.2, 7.0, 3.3)
(mean(x)-3)/(sd(x)/sqrt(6))

## [1] 0.8420842
```

or

```
t.test(x, mu = 3)

##
##  One Sample t-test
##
## data:  x
## t = 0.84208, df = 5, p-value = 0.4382
## alternative hypothesis: true mean is not equal to 3
## 95 percent confidence interval:
```

```
##  0.9815683 6.9850984
## sample estimates:
## mean of x
##  3.983333
```

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

Question IX.2 (23)

Determine a 90% confidence interval for the variance of the refrigerator temperature.

1 ☐ $[2.9, 51.5]$

2 ☐ $[3.2, 49.2]$

3* ☐ $[3.7, 35.7]$

4 ☐ $[3.9, 33.7]$

5 ☐ $[4.1, 31.7]$

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

See Method 3.19. In R:

```
(6-1)*var(x)/qchisq(c(0.95, 0.05), df = 6-1)

## [1]  3.695257 35.712948
```

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

Continue on page 25

Exercise X

The fuel consumption for two different tractors was analysed for 8 different work tasks and the following observations were obtained (in litres per hectare). The fuel consumption can be assumed normal distributed for each tractor:

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8
Tractor A	10.8	8.2	8.7	12	6.2	11.2	8.6	5.5
Tractor B	8.4	8.1	9.4	12.9	10.1	10.4	10.2	11.8

Question X.1 (24)

Which of the following methods is the best for analysing if there is a difference in the two tractors fuel consumption for the tasks carried out?

1 ☐ Test in a multiple linear regression model.

2 ☐ χ^2-test in a contingency table.

3 ☐ Two-sample t-test in a non-paired setup.

4 ☐ One-way analysis of variance.

5* ☐ Two-sample t-test in a paired setup.

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

The setup is similar to the example in Section 3.2.3. In the setup the "treatment" is the tractor (sleeping medicine in the example) and it was applied to each task once (person in the example).

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

Question X.2 (25)

The median fuel consumption (in litres per hectare) for Tractor A is:

1 ☐ 8.7

2 ☐ 12

3 ☐ 8.9

4* ☐ 8.65

5 ☐ 12

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

See Definition 1.5 and Example 1.6. Hence, the values must be sorted and the since n is an even number, the median is the average of the two middle numbers.

In R, see the Example 1.22, it can be calculated by:

```
x <- c(10.8, 8.2, 8.7, 12.0, 6.2, 11.2, 8.6, 5.5)
median(x)

## [1] 8.65
```

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

Continue on page 27

Exercise XI

Let $X \sim N(0, \sigma^2)$ and define the random variable Y by $Y = e^X$.

Question XI.1 (26)

What is $P(Y > 1)$?

1 ☐ 0.84

2* ☐ 0.5

3 ☐ 0.025

4 ☐ 0.16

5 ☐ 0.95

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

As $e^X > 1$ if and only if $X > 0$, then

$$P(Y > 1) = P(X > 0) = 0.5\,.$$

The last equality holds because any normal distribution is symmetric around its mean.

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

Question XI.2 (27)

What is the variance of Y?

1 ☐ $e^{\sigma^2/2}$

2 ☐ $e^{2+\sigma^2}(e^{1/2} - 1)$

3* ☐ $e^{\sigma^2}(e^{\sigma^2} - 1)$

4 ☐ $e^{2+\sigma^2}(e^{\sigma^2} - 1)$

5 ☐ e^{σ^2}

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

See Section 2.5.3 on the log-normal distribution, from which it follows that $Y \sim LN(0, \sigma^2)$. Then, the result follows from Theorem 2.47.

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

Continue on page 28

Exercise XII

A manufacturer would like to examine the quality of its production facilities. A random sample, consisting of observed times between the production of faulty elements, was collected from the production plant. The values are in hours and loaded into R with the following code:

```
x <- c(39.5, 59.7, 42.1, 13, 3.6, 10.9, 61.6, 1, 17.8, 5,
       24.3, 21, 4.2, 21.1, 78.9, 11.1, 6.6, 0.3, 9.2, 10.4)
```

Question XII.1 (28)

Use the book's definition of sample quantiles to determine the *IQR* (*"Inter Quartile Range"*) of the sample.

1 ☐ $IQR = 69.6$

2* ☐ $IQR = 26.1$

3 ☐ $IQR = 58.35$

4 ☐ $IQR = 6.25$

5 ☐ $IQR = 16.05$

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

The *IQR* is the difference between the 0.25 and 0.75 sample quantiles, here computed using Definition 1.7:

```
quantile(x, 0.75, type = 2) - quantile(x, 0.25, type = 2)

##  75%
## 26.1
```

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

Question XII.2 (29)

It has been decided that the plant must be stopped and repaired if the time between the faults becomes too short. To avoid making assumptions regarding the distribution of the time between faults, one would like to construct a non-parametric 95% bootstrap confidence interval for the median. Which of the following R codes determines this interval correctly?

1 ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
quantile(apply(simsamples, 2, mean), c(0.05, 0.95))
```

2 ☐
```
simsamples <- replicate(10000, sample(x, replace = FALSE))
quantile(apply(simsamples, 2, mean), c(0.025, 0.975))
```

3 ☐
```
simsamples <- replicate(10000, sample(x, replace = FALSE))
quantile(apply(simsamples, 2, median), c(0.05, 0.95))
```

4* ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
quantile(apply(simsamples, 2, median), c(0.025, 0.975))
```

5 ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
quantile(apply(simsamples, 2, median), c(0.005, 0.995))
```

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

In order to obtain the desired interval, a large number of medians must be simulated. To make it non-parametric correctly the draw from the sample must be with replacement, hence `replace = TRUE`.

Subsequently, the endpoints of the confidence interval are chosen as the 0.025 and 0.975 sample quantiles of the simulated medians.

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

Question XII.3 (30)

After a repair of the plant, a new sample is collected and entered into R with the code below:

```
y <- c(15.3, 28.2, 53.3, 42, 28.5, 45.3, 40.3, 32.3, 81.1, 29.3,
       82.9, 38.7, 131.5, 24.7, 5.7, 104.3, 30, 31.8, 46.9, 34.9)
```

Subsequently, the following simulations and calculations are carried out:

```
simXsamples <- replicate(10000, rexp(length(x), 1/mean(x)))
simYsamples <- replicate(10000, rexp(length(y), 1/mean(y)))
simDiff <- apply(simXsamples, 2, median) - apply(simYsamples, 2, median)

quantile(simDiff, c(0.005,0.995))

##       0.5%      99.5%
## -50.595024   7.893082

quantile(simDiff, c(0.025,0.975))
```

```
##       2.5%      97.5%
## -42.009475   2.646692

quantile(simDiff, c(0.05,0.95))

##           5%          95%
## -37.50105759  -0.01677407
```

Which of the following conclusions is correct based on the R output in this question?

1 ☐ At $\alpha = 1\%$ significance level it may be concluded that there is a significant difference in medians, when no assumptions are made about the distributions of the times.

2* ☐ At $\alpha = 5\%$ significance level it may be concluded that there is no significant difference in medians, under the assumption that the times in both samples are exponentially distributed.

3 ☐ At $\alpha = 10\%$ significance level it may be concluded that there is no significant difference in means, under the assumption that the times in both samples are exponentially distributed.

4 ☐ At $\alpha = 10\%$ significance level it may be concluded that there is a significant difference in means, under the assumption that the times in both samples are normally distributed.

5 ☐ At $\alpha = 1\%$ significance level it may be concluded that there is a significant difference in means, when no assumptions are made about the distributions of the times.

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

The simulations assume that the observations in both samples are exponentially distributed. As the 95% parametric bootstrap confidence interval for the difference between the medians contains 0, no significant difference is established.

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

The exam paper is finished. Have a great Christmas vacation!

Technical University of Denmark *Page 1 of 34 pages.*

Written examination: 26. May 2019

Course name and number: **Introduction to Statistics (02323)**

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 11 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	III.1	III.2	III.3	IV.1	IV.2	V.1
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer	4	2	4	5	1	5	2	4	4	4

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

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

The exam paper contains 34 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

In a cola tasting experiment there are 4 glasses with cola. Each glass contains either regular cola or light cola. You know that there are two glasses of each. A taster randomly chooses two glasses.

Question I.1 (1)

What is the probability that she gets regular cola in one of the glasses and light cola in the other?

1 ☐ 1/4

2 ☐ 1/3

3 ☐ 1/2

4* ☐ 2/3

5 ☐ 3/4

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

It is drawing without replacement, i.e. the hypergeometric distribution:

```
## We use
dhyper(x=1, m=2, n=2, k=2)

## [1] 0.6666667

## Or we can do it another way and calculate it directly. The probability of getting
(1/2*1/3)

## [1] 0.1666667

## thus the probability of getting two times cola zero is the same.
(1/2*1/3)

## [1] 0.1666667

## Hence not getting either 2 cola or 2 zeros (and thus getting one of each) is
1 - 2 * (1/2*1/3)

## [1] 0.6666667
```

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

Question I.2 (2)

In another experiment, a glass of regular cola and a glass of light cola are given to each of 25 tasters. They are told to taste and answer if they think that there is a difference between the cola in the glasses. The answers are independent of each other.

From experience, one knows that it can be assumed that there is $p = 0.8$ probability that a taster can taste the difference between regular and light. Let X denote the number of the 25 tasters who say there is a difference. What will be the variance of X?

1 ☐ $\mathrm{V}(X) = 5$

2* ☐ $\mathrm{V}(X) = 4$

3 ☐ $\mathrm{V}(X) = 3$

4 ☐ $\mathrm{V}(X) = 2$

5 ☐ $\mathrm{V}(X) = 1$

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

In this setup it is "drawing" with replacement, hence X follows a binomial distribution

$$X \sim B(n = 25, p = 0.8)$$

and we can use Theorem 2.21 to find the variance

```
n <- 25
p <- 0.8
n*p*(1-p)

## [1] 4
```

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

Continue on page 4

Exercise II

10 women measured their morning temperature on both July 1st and December 1st. From the measurements, one would like to investigate whether there is a difference in the morning temperature for women in the summer compared to the winter. It can be assumed that the summer measurements are normally distributed and that the winter measurements are normally distributed.

Question II.1 (3)

Which analysis will be most appropriate?

1 ☐ Test for the difference between two proportions

2 ☐ Regression analysis

3 ☐ (Un-paired) t-test

4* ☐ Paired t-test

5 ☐ Test in the binomial distribution

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

This is a paired setup, since for each individual there exists two observations of the same variable at two different times. Hence the two samples of each 10 temperatures can be paired via the women.

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

Question II.2 (4)

When the test was carried out a p-value of 0.4 was obtained. This means that:

1 ☐ There is a 40% probability that there is a difference between the morning temperature in the summer compared to the winter.

2 ☐ There is a 0.4% probability that there is a difference between the morning temperature in the summer compared to the winter.

3 ☐ The hypothesis cannot be tested.

4 ☐ There is definitely a difference between the morning temperature in the summer compared to the winter.

5* ☐ Under the null hypothesis, the probability of obtaining a value of the test statistic which is less extreme, than the value obtained, is 0.6.

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

This derived from the definition of the p-value in Definition 3.22.

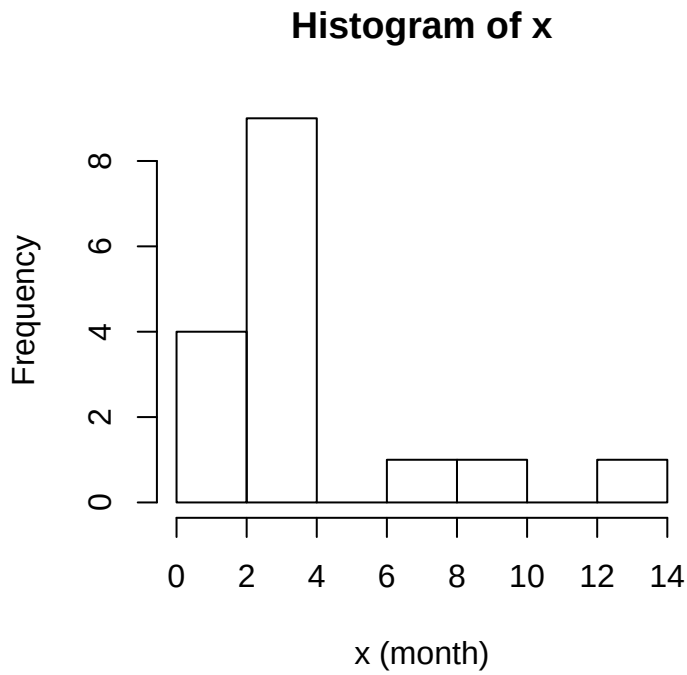
---------- FACIT-END ----------

Continue on page 6

Exercise III

A company has purchased a new 3D printer technology and they want to investigate whether it can be used to make components that are durable enough to be included in a specific product.

An experiment has been carried out where components, printed with the new technology, have been used in a batch of test products. These products have then been subjected to a test that determines their lifetime. It is assumed that the lifetime follows an exponential distribution, so let $X \sim Exp(\lambda)$ denote the lifetime in months. A sample has been collected for $n = 16$ products. A histogram of the sample is:



The observed life times has been saved in the vector `x` and the following R code is run:

```
## Number of simulations
k <- 10000
nx <- length(x)
## Simulate k times
simxsamples <- replicate(k, rexp(nx, 1/mean(x)))
## Calculate the sample mean
simmeans <- apply(simxsamples, 2, mean)
## Quantiles of the means
quantile(simmeans, c(0.005,0.995))

##  0.5% 99.5%
##  1.70  6.42

quantile(simmeans, c(0.025,0.975))
```

```
##  2.5% 97.5%
##  2.07  5.68

quantile(simmeans, c(0.05,0.95))

##   5%  95%
## 2.26 5.26
```

Question III.1 (5)

It was pre-planned to investigate whether it can be shown, at significance level $\alpha = 1\%$, that the average lifetime m_X is over 2 months for the components.

Can this be concluded on the basis of the collected sample and the calculations above (both conclusion and argument must be correct)?

1* ☐ Since 2 is contained in the calculated 99% confidence interval it cannot be concluded.

2 ☐ Since 2 is not contained in the calculated 99% confidence interval it can be concluded.

3 ☐ Since 2 is contained in the calculated 95% confidence interval it cannot be concluded.

4 ☐ Since 2 is not contained in the calculated 95% confidence interval it can be concluded.

5 ☐ With the given information it is not possible to answer this question.

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

To find the 99% confidence interval we find the 0.005 and 0.995 quantiles of the simulated means (the first of the three quantile calculations above). Since 2 is contained in this interval which is [1.70, 6.42], we cannot conclude with $\alpha = 1\%$ that the true average lifetime is not 2.

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

Question III.2 (6)

What is the sample mean of the collected sample?

1 ☐ $\bar{x} = 3.40$

2 ☐ $\bar{x} = 3.76$

3 ☐ $\bar{x} = 3.875$

4 ☐ $\bar{x} = 4.06$

5* ☐ With the given information it is not possible to answer this question.

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

One approach is to take the value between the confidence intervals. But for this to be the mean, we need to assume that the means are normally distributed (so the quantiles of the means are symmetrical), and this is only true if nx is big enough so it satisfies the central limit theorem and we don't know if this is the case. Therefore we cannot find the mean of the sample.

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

Question III.3 (7)

A new sample of lifetimes has been collected where a new material has been used to print the components. They are subsequently subjected to the same tests and the observed lifetimes are stored in the vector `y`. There are $n_Y = 17$ observations in the new sample.

The following R code is run afterwards:

```
## Number of simulations
k <- 10000
nx <- length(x)
ny <- length(y)
## Simulate k times
simxsamples <- replicate(k, rexp(nx, 1/mean(x)))
simysamples <- replicate(k, rexp(ny, 1/mean(y)))
## Calculate the simulated statistics
simdifmeans <- apply(simysamples, 2, mean) - apply(simxsamples, 2, mean)
simdifmedians <- apply(simysamples, 2, median) - apply(simxsamples, 2, median)
## Quantiles of the simulated statistics
quantile(simdifmeans, c(0.025,0.975))
```

```
##  2.5% 97.5%
## 0.733 9.443
```

```
quantile(simdifmeans, c(0.05,0.95))
```

```
##   5%  95%
## 1.30 8.59
```

```
quantile(simdifmedians, c(0.025,0.975))
```

```
##   2.5%  97.5%
## -0.428  8.265
```

```
quantile(simdifmedians, c(0.05,0.95))
```

```
##     5%    95%
## 0.0837 7.3868
```

Which of the following conclusions can be drawn on the basis of these calculations?

1 ☐ At $\alpha = 5\%$ significance level it can be concluded that the 50% quantile of the product lifetime is higher with components of the new material.

2* ☐ At $\alpha = 10\%$ significance level it can be concluded that the 50% quantile of the product lifetime is higher with components of the new material.

3 ☐ At $\alpha = 5\%$ significance level it can be concluded that there is at least 50% probability that the product lifetime is higher with components of the new material.

4 ☐ At $\alpha = 10\%$ significance level it can be concluded that there is at least 50% probability that the product lifetime is higher with components of the new material.

5 ☐ With the given information no conclusions can be drawn.

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

The 50% quantile is the same as the median. Therefore we are using the simdifmedians.From the two last calculations in the R-code it can be seen that the 95% confidence overlaps with 0, but the 90% confidence interval does not. Therefore, with $\alpha = 5\%$, we cannot conclude that there is a difference between the medians, but at $\alpha = 10\%$ we can, so answer 2 is correct.

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

Continue on page 10

Exercise IV

Assume that X is normally distributed with mean 10 and variance 4, Y is normally distributed with mean 20 and variance 25, and X and Y are independent.

Question IV.1 (8)

Then $2Y - 2X + 4$ has the variance:

1 ☐ 36

2 ☐ 58

3 ☐ 84

4* ☐ 116

5 ☐ None of the values above.

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

Use the variance identities in Theorem 2.54 and 2.56 to get

$$\mathrm{V}(2Y - 2X + 4) = 4\,\mathrm{V}(Y) + 4\,\mathrm{V}(X) = 4 \cdot 4 + 4 \cdot 25 = 116.$$

Or you can simulate it:

```
k <- 100000
x <- rnorm(k, 10, sqrt(4))
y <- rnorm(k, 20, sqrt(25))
z <- 2*y - 2*x + 4
var(z)

## [1] 116.0609
```

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

Question IV.2 (9)

What is the standard deviation of $f(X, Y) = 2Y^2 + X^3/3$ (tip: if you solve this using simulation, remember to have many repetitions and choose the answer with the result being approx. $\pm$ 10 from the stated number in the answer)?

1 ☐ $\sigma_{f(X,Y)} \approx 100$

2 ☐ $\sigma_{f(X,Y)} \approx 250$

3 ☐ $\sigma_{f(X,Y)} \approx 350$

4* ☐ $\sigma_{f(X,Y)} \approx 450$

5 ☐ $\sigma_{f(X,Y)} \approx 5 \cdot 10^4$

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

Solve it using simulation (as presented in beginning of Chapter 4)

```
k <- 1000000
x <- rnorm(k, 10, sqrt(4))
y <- rnorm(k, 20, sqrt(25))
sd(2*y^2 + x^3/3)

## [1] 460.3031
```

or use the linear approximation with the error propagation formula in Method 4.3

$$\begin{aligned}
\sigma^2_{f(X,Y)} &= (\frac{\partial f}{\partial x})^2\sigma_x^2 + (\frac{\partial f}{\partial y})^2\sigma_y^2 \\
&= (3X^2/3)^2 \cdot 4 + (4Y)^2 \cdot 25 \\
&= X^4 \cdot 4 + 16Y^2 \cdot 25 \\
&= 10^4 \cdot 4 + 16 \cdot 20^2 \cdot 25
\end{aligned}$$

```
  sqrt(10^4 * 4 + 16 * 20^2 * 25)

## [1] 447.2136
```

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

Continue on page 12

Exercise V

The association between pressure (p) and depth (h) in an open liquid container may be described theoretically by the equation

$$p = p_0 + \rho g h \,,$$

where p_0 is atmospheric pressure, ρ is the density of the liquid, and g is the acceleration due to gravity. An experiment was conducted with the purpose of determining the density of a special liquid. 10 measurements of depth (in m) and pressure (in Pa) were conducted in this liquid, and the results were assigned to two vectors in R, `depth` and `pressure`, respectively. Furthermore, the following R code was run:

```
model1 <- lm(pressure ~ depth)
summary(model1)

##
## Call:
## lm(formula = pressure ~ depth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -119166  -73422   30513   53635  124689
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 1.019e+08  5.867e+04 1737.529  < 2e-16 ***
## depth       5.031e+03  9.455e+02    5.321 0.000711 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85880 on 8 degrees of freedom
## Multiple R-squared:  0.7797,Adjusted R-squared:  0.7521
## F-statistic: 28.31 on 1 and 8 DF,  p-value: 0.0007105
```

Question V.1 (10)

Give the estimate of the atmospheric pressure during the experiment:

1 ☐ $5.031 \cdot 10^3$ Pa

2 ☐ $5.867 \cdot 10^4$ Pa

3 ☐ $9.455 \cdot 10^7$ Pa

4* ☐ $1.019 \cdot 10^8$ Pa

5 ☐ $1.025 \cdot 10^8$ Pa

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

Estimated atmospheric pressure corresponds to the model intercept. It can be seen from the theoretical equation that the atmospheric pressure is the bias term corresponding to our normal β_0.

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

Question V.2 (11)

One would like to test the hypothesis that the expected atmospheric pressure is $1.005 \cdot 10^8$ Pa under the experimental conditions. Give the usual test statistic used to test this hypothesis:

1 ☐ $t_{\text{obs}} = 1738$

2 ☐ $t_{\text{obs}} = 5.321$

3* ☐ $t_{\text{obs}} = 23.86$

4 ☐ $t_{\text{obs}} = 28.31$

5 ☐ $t_{\text{obs}} = 0.000711$

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

Method 5.14 for the intercept parameter β_0 with $\hat{\beta}_0 = 1.019 \cdot 10^8$, $\beta_{0,0} = 1.005 \cdot 10^8$ and $\hat{\sigma}_{\beta_0} = 5.867 \cdot 10^4$. Then

$$t_{\text{obs}} = \frac{1.019 \cdot 10^8 - 1.005 \cdot 10^8}{5.867 \cdot 10^4} = 23.86$$

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

Question V.3 (12)

Give a 95% confidence interval for the parameter which describes the association between depth and pressure:

1 ☐ $1.019 \cdot 10^8 \pm 2.306 \cdot 85880/(10-2)$

2 ☐ $1.019 \cdot 10^8 \pm 2.306 \cdot 85880$

3 ☐ $5031 \pm 2.306 \cdot 85880$

4 ☐ $1.019 \cdot 10^8 \pm 2.306 \cdot 5.867 \cdot 10^4$

5* ☐ $5031 \pm 2.306 \cdot 945.5$

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

Method 5.15 with $\hat{\beta}_1 = 5031$, $\hat{\sigma}_{\beta_1} = 945.5$ (both from the R-output) and $t_{0.975}$ found using

```
qt(0.975, df = 10-2)

## [1] 2.306004
```

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

Question V.4 (13)

Give an estimate of the density of the liquid during the experiment, when the acceleration due to gravity, g, is 9.82 N/kg:

1* ☐ 512 kg/m^3

2 ☐ 1004 kg/m^3

3 ☐ 307 kg/m^3

4 ☐ 802 kg/m^3

5 ☐ 610 kg/m^3

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

The model slope is $\beta_1 = \rho g$, so

$$\hat{\rho} = \frac{\hat{\beta}_1}{g} = \frac{5031}{9.82} = 512$$

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

Continue on page 16

Exercise VI

A sample was taken with independent observations from a normally distributed population. One would like to test the hypothesis that the mean is zero against the alternative, that it is different from zero. The test statistic for the test follows a t-distribution. A p-value of 0.001 was obtained.

Question VI.1 (14)

What is then known about the 99% confidence interval for the mean?

1 ☐ It contains zero.

2* ☐ It does not contain zero.

3 ☐ It contains zero, but not the estimate of the mean.

4 ☐ There is not enough information to know anything specific about the confidence interval.

5 ☐ It contains 0.01.

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

It is a one-sample t-test. We use Theorem 3.33 and we can see that testing the null hypothesis $H_0 : \mu = 0$ and rejecting it on a significance level $\alpha = 0.01$ means that the 0 is not contained in the 99% confidence interval.

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

Question VI.2 (15)

If there were $n = 20$ observations in the sample, what do we then know about the observed test statistic?

1 ☐ $t_{\text{obs}} = -1.33$ or $t_{\text{obs}} = 1.33$

2 ☐ $t_{\text{obs}} = -1.73$ or $t_{\text{obs}} = 1.73$

3 ☐ $t_{\text{obs}} = -3.55$ or $t_{\text{obs}} = 3.55$

4 ☐ $t_{\text{obs}} = -3.58$ or $t_{\text{obs}} = 3.58$

5* ☐ $t_{\text{obs}} = -3.88$ or $t_{\text{obs}} = 3.88$

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

It is a one-sample t-test and since had have a p-value of 0.001, hence

$$\begin{aligned} 2 \cdot P(T > t_{\text{obs}}) &= 0.001 \Leftrightarrow \\ P(T > t_{\text{obs}}) &= 0.001/2 = 0.0005 \end{aligned}$$

we need to find the 0.05% or 99.95% quantile. So

```
qt(0.001/2, df=19)

## [1] -3.883406
```

or

```
qt(1 - 0.001/2, df=19)

## [1] 3.883406
```

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

Continue on page 18

Exercise VII

The Danish Veterinary and Food Administration wants to reduce the proportion of resistant bacteria in pigs intestinal flora, as they pose a human risk. qPCR is one microbiological method to count the number of specific genes in a faeces sample. Below is the count of three genes: 16S, which is a reference gene, and two genes that encode resistance to tetracycline (tetO and tetM). Four samples were taken at different times (first Sample 1, then 2, 3 and finally 4) on the same farm and the researchers want to investigate whether changes have occurred.

	16S	tetO	tetM	Sum
Sample 1	4675	171	76	4922
Sample 2	2222	95	1	2318
Sample 3	2750	49	2	2801
Sample 4	2040	47	1	2088
Sum	11687	362	80	12129

A χ^2-test should be carried out to determine if the proportion of resistant genes has changed over time.

Question VII.1 (16)

The degrees of freedom in this test is:

1 ☐ 8

2 ☐ 12

3* ☐ 6

4 ☐ 9

5 ☐ It doesn't make sense to do a χ^2-test, when two of the observations are 1.

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

It is the χ^2-test in Method 7.22 which is used. The degrees of freedom is:

$$(r-1)(c-1) = 3 * 2 = 6$$

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

Question VII.2 (17)

Under the null hypothesis what is the expected number of tetM copies in Sample 4?

1 ☐ 20

2 ☐ 1

3* ☐ 13.77

4 ☐ 26.10

5 ☐ 696

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

The expected number in the cell is found as in equation 7-53 by

$$\frac{columntotal * rowtotal}{grandtotal} = 80 * 2088/12129 = 13.77$$

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

Question VII.3 (18)

The test statistic turns out to be 132.3. The relevant p-value is found using which of the following calls in R?

1 ☐ `1 - dchisq(132.3, df=6)`

2* ☐ `1 - pchisq(132.3, df=6)`

3 ☐ `qchisq(132.3, df=6)`

4 ☐ `pchisq(132.3, df=6)`

5 ☐ `qchisq(1/132.3, df=6)`

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

See Method 7.22. We need to find $P(\chi^2 > \chi^2_{\text{obs}})$, which can by done either by

```r
1 - pchisq(132.3, df=6)
```

```
## [1] 0
```

which, since the value is below the machine precision is zero, actually by

```r
pchisq(132.3, df=6, lower.tail = FALSE)
```

```
## [1] 4.212873e-26
```

it can be calculated, but in practice it so small that is doesn't make much difference, but maybe nicer to be led to think that it is exactly zero.

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

Question VII.4 (19)

It has previously been planned to investigate whether the occurrence of tetO has changed between the first sample and fourth sample. For reasons not explained here, the observations of tetM should not be considered in this test. The following code has been run with the associated code output:

```
prop.test(x=c(171, 47), n=c(4675+171, 2040+47), correct=FALSE, conf.level=0.95)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(171, 47) out of c(4675 + 171, 2040 + 47)
## X-squared = 7.8067, df = 1, p-value = 0.005205
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.004550394 0.020982546
## sample estimates:
##     prop 1     prop 2
## 0.03528683 0.02252036
```

The usual $\alpha = 0.05$ significance level is used. What is the conclusion (both the conclusion and the argumentation must be correct)?

1 ☐ No significant change has been detected, since $0.02098 < 0.02252$.

2 ☐ A significant change has been detected, since $0.0052 < 0.95$, but it is not possible to conclude if the occurrence has increased or decreased.

3 ☐ A significant change has been detected, since $0.0052 < 0.05$, and the occurrence of tetO has increased.

4 ☐ A significant change has been detected, since $0.0052 < 0.95$, and the occurrence of tetO has increased.

5* ☐ A significant change has been detected, since $0.0052 < 0.05$, and the occurrence of tetO has decreased.

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

We compare the p-value (0.0052) with the significance level of 0.05. Since $P - value < \alpha$ there is a significant change. If we look at the sample estimates we see that the estimate has changed from 0.035 in the first sample to 0.023 in the fourth sample. So answer is correct.

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

Continue on page 22

Exercise VIII

The IQ of a randomly selected individual is modeled by a normally distributed random variable. 50% of the population have an IQ over 100 (and 50% have an IQ below 100). Suppose 68% of the population have an IQ in the range of 85-115.

Question VIII.1 (20)

What percentage of the population have an IQ of at least 140 and is thus considered geniuses according to this model?

1 ☐ 0.01%

2 ☐ 1%

3 ☐ 4%

4* ☐ 0.4%

5 ☐ 0.06%

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

Use need to use the standardized normal distribution. Theorem 2.43. First we find out how many percent have $IQ < 85$:

```
## The quantile at 85
50-68/2

## [1] 16
```

We find out that this is 16% of the population. We now need to find the 0.16 quantile in the std. normal distribution:

```
## The 0.16 quantile in the std. norm.
qnorm(0.16)

## [1] -0.9944579
```

We can then use this in the equation to find the standard deviation of the IQ distribution:

$$Z = \frac{X-\mu}{\sigma} \iff \sigma = \frac{X-\mu}{Z} = \frac{85-100}{-0.994}$$

```r
## The standard deviation in the IQ distribution
sigma <- (85-100) / qnorm(0.16)
sigma

## [1] 15.0836
```

And we now have the mean and the standard deviation of the IQ distribution and like usual we can find the proportion of people with $IQ > 140$ as:

```r
## The proportion of geniueses
1 - pnorm(140, mean=100, sd=sigma)

## [1] 0.004002158
```

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

Continue on page 25

Exercise IX

The data below have been collected from two groups:

Group 1: 10.5, 9.3, 10.7, 10.8, 11.2

Group 2: 8.9, 9.5, 10.2, 9.8, 10.3

All measurements are assumed to be taken independent. The Group 1 measurements are believed to originate from a normal distribution, and the measurements in Group 2 are assumed to originate from a normal distribution. In addition, it is assumed that the variances in the two normal distributions are identical.

Question IX.1 (21)

What is the sample mean of the Group 2 sample?

1* ☐ 9.74

2 ☐ 9.8

3 ☐ 10.2

4 ☐ 10.31

5 ☐ 48.5

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

Simply calculate the sample mean of the values from Group 2 in R:

```
y <- c(8.9, 9.5, 10.2, 9.8, 10.3)
mean(y)

## [1] 9.74
```

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

Question IX.2 (22)

What will be the numerical value of the test statistic for the usual test of the hypothesis that there is no difference in mean of the two groups?

1 ☐ 0.8

2 ☐ 1.04

3* ☐ 1.86

4 ☐ 2.19

5 ☐ 2.55

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

This is a two-sample t-test, hence either use the formulas in Method 3.47 or maybe easier do it in R by:

```
x <- c(10.5, 9.3, 10.7, 10.8, 11.2)
y <- c(8.9, 9.5, 10.2, 9.8, 10.3)
## Same result of tobs if the variance is pooled
t.test(x, y)

##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 1.8564, df = 7.601, p-value = 0.1024
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1927507  1.7127507
## sample estimates:
## mean of x mean of y
##     10.50      9.74

# Or we can specify the variances as being equal (in which case the pooled variance wi
t.test(x, y, var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  x and y
## t = 1.8564, df = 8, p-value = 0.1005
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1840545  1.7040545
## sample estimates:
## mean of x mean of y
##     10.50      9.74
```

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

Question IX.3 (23)

What is the 90% confidence interval for the mean in Group 1?

1 ☐ [9.61, 11.39]

2 ☐ [9.32, 11.68]

3 ☐ [8.92, 12.03]

4 ☐ [9.87, 12.03]

5* ☐ None of the intervals above are correct.

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

Use the formula in Method 3.9 or do the calculation in R by:

```
x <- c(10.5, 9.3, 10.7, 10.8, 11.2)
t.test(x, conf.level=0.9)

##
##  One Sample t-test
##
## data:  x
## t = 32.717, df = 4, p-value = 5.204e-06
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##   9.815813 11.184187
## sample estimates:
## mean of x
##      10.5
```

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

Question IX.4 (24)

A new experiment must be designed in order to achieve a greater power of the statistical test for the mean values. There is still an equal number of observations in each group. The researchers want to have 99% power to discover a difference in mean of at least 1 between the two groups, at significance level 1%. As a guess of the variance, the pooled variance estimate from the two samples are used.

What is the minimum number of observations needed from each group in order for the above requirements to be fulfilled?

1 ☐ At least 4

2 ☐ At least 6

3 ☐ At least 12

4 ☐ At least 18

5* ☐ At least 22

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

First use Method 3.52 as done in Example 2.85 to calculate the pooled variance estimate by:

```
x <- c(10.5, 9.3, 10.7, 10.8, 11.2)
y <- c(8.9, 9.5, 10.2, 9.8, 10.3)
n1 <- length(x)
n2 <- length(y)
varp <- ((n1-1)*var(x)+(n2-1)*var(y)) / (n1+n2-2)
```

Then use this to calculate the needed sample size by inserting the 4 out of 5 needed values and the R calculate the sample size, see Section 3.3.3:

```
power.t.test(delta=1, sd=sqrt(varp), sig.level=0.01, power=0.99)
```

```
##
##      Two-sample t test power calculation
##
##               n = 21.87928
##           delta = 1
##              sd = 0.6473021
##       sig.level = 0.01
##           power = 0.99
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

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

Continue on page 29

Exercise X

How much clothes a person wears (the clothing level) has a large influence on the level of comfort in offices. In the table below samples from three rooms of the average clothing level (on a scale 0 to 1) are presented:

	Room 1	Room 2	Room 3
	0.43	0.56	0.38
	0.36	0.71	0.39
	0.41	0.20	0.48
	0.42	0.57	0.52
	0.41	0.69	0.23
	0.54	0.55	0.37
	0.61	0.78	0.60
	0.53	0.42	0.46
	0.49	0.42	0.44
	0.69	0.59	0.44
Means	0.49	0.55	0.43

As an initial analysis, a one-way analysis of variance, with room as explanatory factor, is carried out. The result is shown in the R output below (where significant codes have been removed and some numbers are replaced by letters):

```
anova(lm(clo ~ room, data=Data))

## Analysis of Variance Table
##
## Response: clo
##           Df  Sum Sq  Mean Sq F value Pr(>F)
## room       2 0.06963 0.034813     A   0.1385
## Residuals 27 0.44147 0.016351
```

Question X.1 (25)

What is the value of A (rounded) in the R output above?

1 ☐ $A = 1.07$

2 ☐ $A = 2.00$

3* ☐ $A = 2.13$

4 ☐ $A = 4.00$

5 ☐ $A = 4.26$

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

See theorem 8.6.

```
((0.06963)/2) / (0.44147/27)

## [1] 2.129261

## or simply
0.034813 / 0.016351

## [1] 2.129105
```

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

Question X.2 (26)

What is the conclusion (at significance level $\alpha = 0.05$) about the difference in mean clothing level between the three rooms (both the conclusion and the argument must be correct)?

1 ☐ There is a significant difference since $0.016351 < 0.05$.

2 ☐ There is not a significant difference since $0.016351 < 0.05$.

3 ☐ There is a significant difference since $0.1385 > 0.05$.

4* ☐ There is not a significant difference since $0.1385 > 0.05$.

5 ☐ There is a significant difference since $0.034813 < 0.05$.

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

By examining the p-value which can be read as 0.1385 from the R output, it is clear that this is lower than the significance level of $\alpha = 0.05$ and therefore there is no significant difference.

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

Question X.3 (27)

What is a pre-planned 95% confidence interval for the difference between the mean value in Room 1 and Room 2 (i.e. it was planned to make this confidence interval only, before the sample was collected)?

1 ☐ $[0.12, 0.45]$

2 ☐ $[0.03, 0.25]$

3 ☐ $[-0.17, 0.09]$

4* ☐ $[-0.06, 0.18]$

5 ☐ $[-0.30, 0.42]$

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

We use Method 8.9 to calculate the single pre-planned post-hoc 95% confidence interval

```
0.55-0.49+c(-1,1)*qt(0.975,df=27)*sqrt(0.016351*2/10)
```

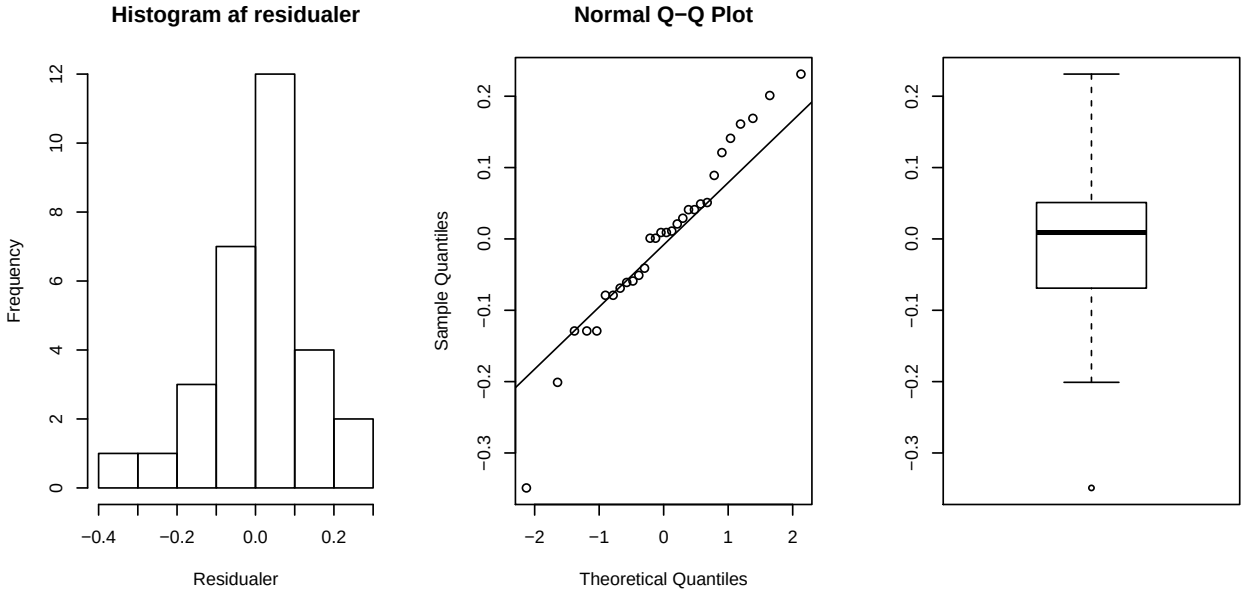
```
## [1] -0.05733529  0.17733529
```

The degrees of freedom $n - k = 27$ and the $\frac{SSE}{n-k} = MSE = 0.016351$ we get from the ANOVA table printed in the R result.

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

Question X.4 (28)

The following histogram, normal qq-plot and box-plot are of the residuals:



What can rightly be judged based on these plot from the books definition of outliers?

1 ☐ That it is clear that the distribution of residuals is left-skewed.

2 ☐ That the residuals appears normally distributed without any outliers.

3* ☐ That the residuals appears normally distributed, though with a single outlier.

4 ☐ That it is clear that the distribution of residuals is right-skewed.

5 ☐ That the residuals do not follow a normal distribution.

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

From the histogram it is not clear to see that the distribution is left-skewed. From the QQ-plot it is seen that the points approximately follow the line, except for the lowest point. And this can also be concluded to be an outlier from the boxplot.

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

Continue on page 33

Exercise XI

The following sample has been sorted:

$$10, 25, 25, 36, 37, 41, 54, 64, 68, 83$$

Question XI.1 (29)

What is the median of the sample?

1 ☐ 37

2 ☐ 38

3* ☐ 39

4 ☐ 40

5 ☐ 41

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

```
quantile(x, type=2)

##   0%  25%  50%  75% 100%
##   10   25   39   64   83
```

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

Question XI.2 (30)

What is the sample variance?

1 ☐ $\mathrm{V}(x) = 22.60$

2* ☐ $\mathrm{V}(x) = 510.7$

3 ☐ $\mathrm{V}(x) = 1521$

4 ☐ $\mathrm{V}(x) = 1962$

5 ☐ $\mathrm{V}(x) = 2052$

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

```r
var(x)
```

```
## [1] 510.6778
```

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

The exam is finished. Have a great summer!

Technical University of Denmark *Page 1 of 38 pages.*

Written examination: 19. Dec 2020

Course name and number: **Introduction to Statistics (02323)**

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

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

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

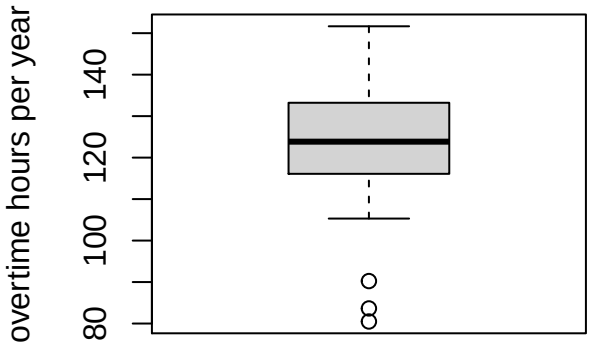
The exam paper contains 38 pages.

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

A city department has introduced a quality improvement program which allows employees to get credit for overtime hours when attending meetings. The total number of overtime hours per year for 36 employees is visualized in the boxplot below.



Question I.1 (1)

Which of the following statements is correct?

1 ☐ $IQR = Q1 - Q3 \approx 17$ hours.

2* ☐ $IQR = Q3 - Q1 \approx 17$ hours.

3 ☐ $IQR = Q4 - Q1 \approx 48$ hours.

4 ☐ The IQR cannot be determined because the boxplot contains three outliers.

5 ☐ $IQR = Q3 - Q1 \approx 48$ hours.

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

Using the boxplot above we can find $Q3 \approx 133$ and $Q1 \approx 116$, hence $IQR = Q3 - Q1 \approx 17$

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

Continue on page 4

Exercise II

The table below shows the number of persons tested positive for coronavirus that were admitted to hospitals in Denmark on 3 different dates during the spring of 2020. Furthermore, the table shows the numbers of those persons that were also in an intensive care unit (ICU).

Date	ICU	Admitted
April 30	62	255
April 10	113	433
March 20	37	153

Question II.1 (2)

Based on the numbers above, what is the usual 95% confidence interval for the probability that, given you are admitted, you are also in an intensive care unit? Assume that the model assumptions are fulfilled.

1 ☐ $[0.72, 0.78]$

2* ☐ $[0.22, 0.28]$

3 ☐ $[0.18, 0.22]$

4 ☐ $[0.16, 0.35]$

5 ☐ $[0.12, 0.28]$

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

The best estimate is to pool the observations from the three dates (this is what we must do if told nothing else about changing conditions etc.):

```
icu <- c(62,113,37)
n <- c(255,433,153)
ph <- sum(icu)/(sum(n))
ph + c(-1,1) * qnorm(0.975) * sqrt(ph*(1-ph)/sum(n))

## [1] 0.2227350 0.2814268

# The result from the built-in function is slightly different (it's using the t-distr
prop.test(sum(icu), sum(n), correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  sum(icu) out of sum(n), null probability 0.5
## X-squared = 206.76, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2239073 0.2825089
## sample estimates:
##         p
## 0.2520809
```

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

Question II.2 (3)

In order to investigate the development over time, the numbers from April 30th and March 20th are now compared. With the null hypothesis that the proportions of patients in ICU are equal on the two dates, what is the p-value and the conclusion given a significance level $\alpha = 0.05$?

1 ☐ p-value=0.476 and the difference is significant.

2 ☐ p-value=0.029 and the difference is not significant.

3* ☐ p-value=0.976 and the difference is not significant.

4 ☐ p-value=0.024 and the difference is significant.

5 ☐ p-value=0.060 and the difference is not significant.

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

```
## Q2: Compare two prop
phs <- icu[c(1,3)] / n[c(1,3)]
ph <- sum(icu[c(1,3)])/sum(n[c(1,3)])

(z <- diff(phs)/sqrt(ph*(1-ph)*(1/n[1]+1/n[3])))

## [1] -0.02981863

2*(1-pnorm(abs(z)))

## [1] 0.9762117
```

```r
# or with the build in function
prop.test(icu[c(1,3)], n[c(1,3)], correct=FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  icu[c(1, 3)] out of n[c(1, 3)]
## X-squared = 0.00088915, df = 1, p-value = 0.9762
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.08457431  0.08718868
## sample estimates:
##    prop 1    prop 2
## 0.2431373 0.2418301
```

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

Question II.3 (4)

The distribution of patients across different regions is now investigated. The table below shows the number of persons admitted to hospital on different dates in the 5 regions of Denmark, we assume here that the same person is not admitted on more than 1 date.

Date	Nordjylland	Midtjylland	Syddanmark	Hovedstaden	Sjælland	All DK
April 30	13	33	12	144	53	255
April 16	21	54	35	183	60	353
April 2	32	77	85	251	86	531
March 18	10	16	12	64	27	129
Total	76	180	144	642	226	1268

We will now investigate if the proportion of admitted patients in the different regions is the same over time (the null hypothesis) or if it changes. Formally, this can be written as

$$\mathrm{H}_0: \quad p_{ij} = p_i$$

for all i.

Under the null hypothesis, what is the contribution to the test-statistics for "Nordjylland" on March 18?

1 ☐ 7.73

2 ☐ 0.59

3 ☐ 5.14

4* ☐ 0.67

5 ☐ 10

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

Under the null hypothesis we assume that the proportion is equal for each row across the groups. Then we can calculate the expected count in the cell and then the contribution to the χ^2 statistic:

```
(e <- 129/1268*76)

## [1] 7.731861

(10-e)^2 / e

## [1] 0.6653577
```

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

Question II.4 (5)

The test statistics is calculated to $\chi^2_{obs} = 29$. Given a significance level $\alpha = 0.05$, what is the p-value and conclusion for the corresponding hypothesis test? (Both argument and conclusion must be correct)

1 ☐ p-value=0.0012 and there is a significant difference

2 ☐ p-value=0.0099 and there is not a significant difference

3 ☐ p-value=0.024 and there is a significant difference

4 ☐ p-value=0.088 and there is not a significant difference

5* ☐ p-value=0.0039 and there is a significant difference

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

```
1 - pchisq(29, df = 12)

## [1] 0.00393999
```

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

Question II.5 (6)

If we on a given day assume that 4% of the population is infected with a virus, how many people should then be tested at random in order to get a margin of error on maximum 1% using significance level $\alpha = 0.05$?

1 ☐ 1039

2* ☐ 1476

3 ☐ 369

4 ☐ 9603

5 ☐ 6764

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

```
0.04 * 0.96 * (qnorm(0.975)/0.01)^2
```

```
## [1] 1475.12
```

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

Continue on page 9

Exercise III

The 2008-09 nine-month academic salary for Professors in a given U.S. college is to be assessed. The data includes salaries of 125 male Professors working in applied departments (in US dollars). It is of interest to find out if the salary depends on the years of work since obtaining a Ph.D. degree and years of service.

Question III.1 (7)

An initial multiple linear regression model was established. The model summary is given below. Assume that the model assumptions are fulfilled!

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd + yrs.service, data = sal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -72479 -20472   -288  16051  92778
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   130213.8     6956.5  18.718   <2e-16 ***
## yrs.since.phd   -304.2      430.1  -0.707    0.481
## yrs.service      529.3      378.5   1.398    0.165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26450 on 122 degrees of freedom
## Multiple R-squared:  0.02085,Adjusted R-squared:  0.004803
## F-statistic: 1.299 on 2 and 122 DF,  p-value: 0.2765
```

Which of the following statements is correct given a significance level $\alpha = 0.05$? (Both conclusion and argument must be correct)

1 ☐ The Professor `salary` depends on `yrs.since.phd` and `yrs.service` because both p-values are greater than 0.05.

2 ☐ The Professor `salary` does NOT depend on `yrs.since.phd` and `yrs.service` because both p-values are greater than 0.025.

3 ☐ We are not given sufficient information to make a conclusion about the relation between Professor `salary` and `yrs.since.phd` and `yrs.service`.

4 ☐ The Professor `salary` depends on `yrs.since.phd` and `yrs.service` because the respective p-values are less than 0.5.

5* ☐ The Professor `salary` does NOT depend on `yrs.since.phd` and `yrs.service` because both p-values are greater than 0.05.

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

The p-values for `yrs.since.phd` and `yrs.service` are both greater than the significance level $\alpha = 0.05$ (Note: *alpha* is not 0.025). We can therefore state that there is no significant correlation between Professor `salary` and `yrs.since.phd` and `yrs.service`. This is equivalent to stating that the Professor `salary` does not depend on `yrs.since.phd` and `yrs.service`.

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

Question III.2 (8)

Backwards model selection was performed for the multiple linear regression model above, resulting in the following R output:

```
##
## Call:
## lm(formula = salary ~ yrs.service, data = sal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73189 -20581     29  15226  92951
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 126901.7     5133.7  24.719   <2e-16 ***
## yrs.service    307.7      212.0   1.451    0.149
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26400 on 123 degrees of freedom
## Multiple R-squared:  0.01684,Adjusted R-squared:  0.008846
## F-statistic: 2.107 on 1 and 123 DF,  p-value: 0.1492
```

```
##
## Call:
## lm(formula = salary ~ 1, data = sal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65959 -19018   -693  16858  98027
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   133518       2372    56.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26510 on 124 degrees of freedom
```

Which R code results in the correct 95% confidence interval for the mean of the Professor salary?

1 ☐
```
133518 + c(-1, 1) * qt(0.95, 124) * 2372
```

2 ☐
```
133518 + c(-1, 1) * qt(0.975, 123) * 2372
```

3* ☐
```
133518 + c(-1, 1) * qt(0.975, 124) * 2372
```

4 ☐
```
126902 + c(-1, 1) * qt(0.975, 124) * 5134
```

5 ☐
```
130214 + c(-1, 1) * qt(0.95, 124) * 6957
```

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

The correct 95% confidence interval is found using the fully reduced model. `yrs.since.phd` and `yrs.service` were not significant and were removed step-wise. Hence, the mean Professor salary can be found using the following R-command:

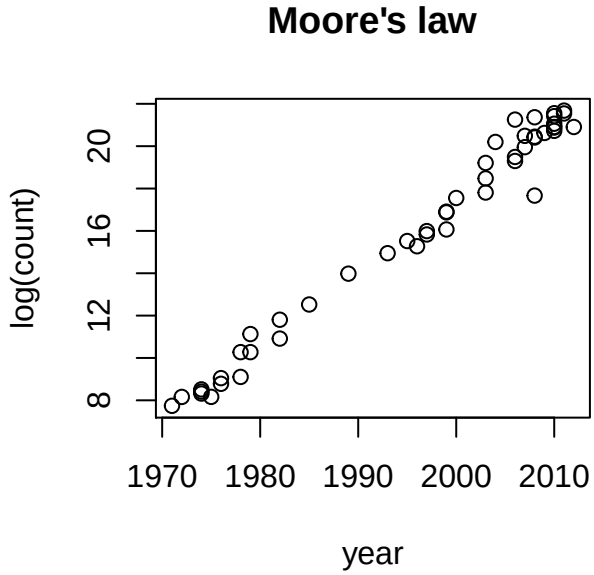
```
133518 + c(-1, 1) * qt(0.975, 124) * 2372

## [1] 128823.1 138212.9
```

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

Continue on page 13

Exercise IV

Moore's law is about the observation that the number of transistors in a dense integrated circuit doubles about every two years. The observation is named after Gordon Moore, the co-founder of Fairchild Semiconductor. In the figure below the transistor count has been transformed using the natural logarithm and plotted against year.




```
##
## Call:
## lm(formula = log(count) ~ year, data = moore)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60701 -0.26843 -0.01245  0.35038  1.67737
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.786e+02  1.414e+01  -48.01        ?
## year         3.481e-01  7.083e-03       ?   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6762 on 46 degrees of freedom
## Multiple R-squared:  0.9813,Adjusted R-squared:  0.9809
## F-statistic:  2415 on 1 and 46 DF,  p-value: < 2.2e-16
```

Question IV.1 (9)

Calculate the test statistic which is missing in the model summary above (missing values have been replaced by question marks in the table above). Which of the following answers is correct?

1 ☐ $t_{\text{obs}} = 0.02$

2 ☐ $t_{\text{obs}} = 12.25$

3 ☐ $t_{\text{obs}} = 0.49$

4* ☐ $t_{\text{obs}} = 49.15$

5 ☐ $t_{\text{obs}} = 12.49$

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

The test statistic, t_{obs}, describes how many standard errors the estimated slope $\hat{\beta}_{year}$ is away from the hypothesized slope $\beta_{year,0} = 0$ and can be calculated using the formula:

$t_{obs} = \frac{\hat{\beta}_{year} - \beta_{year,0}}{\hat{\sigma}_{year}}$, where $\hat{\sigma}_{year}$ is the standard error for the slope of `year`.

Using the model summary we obtain:

$t_{obs} = \frac{3.481 \cdot 10^{-1}}{7.083 \cdot 10^{-3}} = 49.15$

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

Question IV.2 (10)

We want to test the hypothesis $H_0 : \beta_0 = 0$, where β_0 represents the model intercept. Which of the following statements is correct (given $\alpha = 0.05$)? (Both argument and conclusion must be correct!)

1 ☐ We compare the absolute value of the corresponding test statistic $|t_{obs}| = 48.01$ with the critical t-value, $t_{crit} = 1.96$. We reject H_0 because $|t_{obs}| > t_{crit}$.

2* ☐ We compare the absolute value of the corresponding test statistic $|t_{obs}| = 48.01$ with the critical t-value, $t_{crit} = 2.01$. We reject H_0 because $|t_{obs}| > t_{crit}$.

3 ☐ We compare the absolute value of the corresponding test statistic $|t_{obs}| = 48.01$ with the critical t-value, $t_{crit} = 1.68$. We reject H_0 because $|t_{obs}| > t_{crit}$.

4 ☐ We compare the absolute value of the corresponding test statistic $|t_{obs}| = 48.01$ with the critical t-value, $t_{crit} = 2.01$. We accept H_0 because $|t_{obs}| > t_{crit}$.

5 ☐ We compare the absolute value of the corresponding test statistic $|t_{obs}| = 48.01$ with the critical t-value, $t_{crit} = 1.96$. We accept H_0 because $|t_{obs}| > t_{crit}$.

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

We reject H_0 because $|t_{obs}| > t_{crit}$

```
(t_crit <- qt(0.975, df = 46))

## [1] 2.012896
```

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

Question IV.3 (11)

According to the linear model above, what is the expected transistor count increase from 2010 to 2015?

1 ☐ $\ln(-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015) - \ln(-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010)$

2 ☐ $e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015 + 6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010}$

3* ☐ $e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015} - e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010}$

4 ☐ $\ln(-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015 + 6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010)$

5 ☐ $e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010} - e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015}$

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

We use the estimated linear model parameters as shown in the summary above to find expected transistor counts for years 2015 and 2010. These expected counts are on the natural logarithmic scale and require back-transformation. After we have performed the back-transformation we can subtract the expected count of 2010 from the expected count of 2015.

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

Continue on page 16

Exercise V

Question V.1 (12)

One is interested in determining the density of a liquid. To do so, the mass, m, and the volume, V, of the liquid are measured. The density of the liquid is given by

$$\rho = \frac{m}{V}$$

What is the precision (standard deviation, σ_ρ) of the determined density if the mass and the volume can be measured with a precision $\sigma_m = 0.2$ and $\sigma_V = 0.4$, respectively? Assume that mass and volume measurements are independent and normally distributed.

1 ☐ $\sigma_\rho \approx \frac{1}{V^2}(0.2^2 + \frac{0.4^2 m^2}{V^2})$

2* ☐ $\sigma_\rho \approx \sqrt{\frac{1}{V^2}(0.2^2 + \frac{0.4^2 m^2}{V^2})}$

3 ☐ $\sigma_\rho \approx \frac{1}{V^2}(0.4^2 + \frac{0.2^2 m^2}{V^2})$

4 ☐ $\sigma_\rho \approx \frac{0.4^2}{V^2} + \frac{0.2^2 m^2}{V^4}$

5 ☐ $\sigma_\rho \approx \sqrt{\frac{0.4^2}{V^2} + \frac{0.2^2 m^2}{V^4}}$

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

We can find the precision σ_ρ using the error approximation rule for non-linear functions (see slides week 7).

$\sigma_\rho^2 \approx (\frac{\partial \rho}{\partial m})^2 \sigma_m^2 + (\frac{\partial \rho}{\partial V})^2 \sigma_V^2$

$\sigma_\rho^2 \approx \frac{1}{V^2}\sigma_m^2 + \frac{m^2}{V^4}\sigma_V^2$

$\sigma_\rho^2 \approx \frac{1}{V^2}(\sigma_m^2 + \frac{\sigma_V^2 m^2}{V^2})$

$\sigma_\rho \approx \sqrt{\frac{1}{V^2}(\sigma_m^2 + \frac{\sigma_V^2 m^2}{V^2})}$

$\sigma_\rho \approx \sqrt{\frac{1}{V^2}(0.2^2 + \frac{0.4^2 m^2}{V^2})}$

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

Question V.2 (13)

Let X_i be a random variable. The following code is run in R to draw 100 random numbers X_i from a given distribution.

```
x <- rnorm(100)^2 + rnorm(100)^2 + rnorm(100)^2
```

Which of the following statements is correct?

1 ☐ X_i follows a χ^2-distribution with 1 degree of freedom.

2 ☐ X_i follows a standard normal distribution with mean 0 and variance 1.

3 ☐ X_i follows a χ^2-distribution with 2 degrees of freedom.

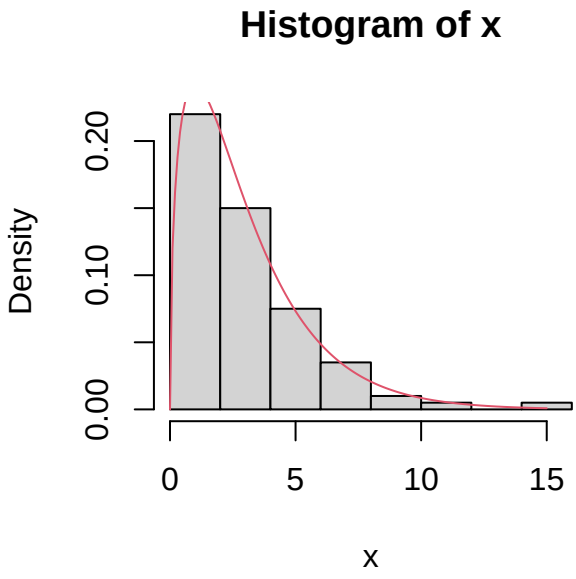
4* ☐ X_i follows a χ^2-distribution with 3 degrees of freedom.

5 ☐ X_i follows a normal distribution with mean 0 and variance 3.

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

We draw 100 standard normal numbers three times, and for each element we sum them, so we according to 2.79 we have 3 df. Check it by:

```
x <- rnorm(100)^2 + rnorm(100)^2 + rnorm(100)^2
length(x)

## [1] 100

hist(x, prob=TRUE)
xseq <- seq(0,15,by=0.1)
lines(xseq, dchisq(xseq, df=3), col=2)
```



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

Question V.3 (14)

Which of the following R commands is drawing 10 random numbers from an exponential distribution?

1* ☐ `replicate(10, rexp(1, 2))`

2 ☐ `pexp(seq(0.1, 1, length.out=10), 2)`

3 ☐ `qexp(seq(0.1, 1, 0.1), 2)`

4 ☐ `rep(dexp(10, 2), 10)`

5 ☐ None of the above. The exponential distribution requires a second parameter, which is missing in all of the above

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

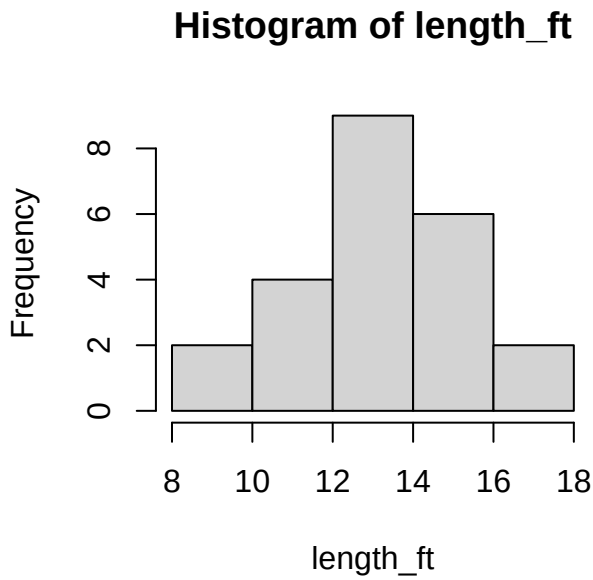
`rexp(1, 2)` draws 1 number from an exponential distribution with mean = 2. The `replicate` command ensures that this procedure is repeated 10 times. An easier way to draw ten random numbers from an exponential distribution would be to use the command `rexp(10, 2)`.

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

Continue on page 19

Exercise VI

Jesus Rivas, a herpetologist, is currently doing research on green anacondas. These snakes, some of the largest in the world, can grow up to 25 feet in length. They have been known to swallow live goats and even people. Jesus Rivas and fellow researchers walk barefoot in shallow water in the Llanos grasslands shared by Venezuela and Colombia during the dry season. When they feel a snake with their feet, they grab it and hold it with the help of another person. After muzzling the snake with a sock and tape, they measure the length of the snake. 23 green anacondas were captured and their length was measured in feet. The sample data is stored in `length_ft`. You can see the corresponding histogram of the sample below.



Question VI.1 (15)

Which of the following is the correct 99% confidence interval for the median anaconda length assuming that parametric bootstrapping was used for estimation of the interval?

```
median_ft <- median(length_ft)
mean_ft <- mean(length_ft)
sd_ft <- sd(length_ft)
n <- length(length_ft)
k <- 10000

sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.025, 0.975))

##     2.5%    97.5%
## 12.02935 14.59873
```

```r
sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft))
sim_medians <- apply(sim_samples, 2, mean)
quantile(sim_medians, c(0.005, 0.995))
```

```
##     0.5%    99.5%
## 11.94304 14.68206
```

```r
sim_samples <- replicate(k, rchisq(n, mean_ft))
sim_medians <- apply(sim_samples, 2, mean)
quantile(sim_medians, c(0.005, 0.995))
```

```
##     0.5%    99.5%
## 10.75972 16.24469
```

```r
sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.005, 0.995))
```

```
##     0.5%    99.5%
## 11.64546 15.04535
```

```r
sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft^2))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.005, 0.995))
```

```
##      0.5%     99.5%
##  9.121213 17.500782
```

1 ☐ [12.03, 14.60]

2 ☐ [11.94, 14.68]

3 ☐ [10.76, 16.24]

4* ☐ [11.65, 15.05]

5 ☐ [9.12, 17.50]

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

Parametric bootstrapping requires knowledge regarding the population's distribution. As it can be seen from the histogram above the snake length follows approx. a normal distribution. As we are interested to simulate the 99% confidence interval for the median snake length only the fourth answer can be correct.

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

Question VI.2 (16)

Which of the following is the correct 99% confidence interval for the median anaconda length assuming that non-parametric bootstrapping was used for estimation of the interval?

```
median_ft <- median(length_ft)
mean_ft <- mean(length_ft)
sd_ft <- sd(length_ft)
n <- length(length_ft)
k <- 10000

sim_samples <- replicate(k, sample(length_ft, n, replace = TRUE))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.005, 0.995))

##     0.5%    99.5%
## 11.93076 15.22501

sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.005, 0.995))

##     0.5%    99.5%
## 11.61621 14.97613

sim_samples <- replicate(k, sample(length_ft, n, replace = TRUE))
sim_medians <- apply(sim_samples, 2, mean)
quantile(sim_medians, c(0.01, 0.99))

##       1%      99%
## 12.08800 14.46791

sim_samples <- replicate(k, sample(length_ft, n, replace = TRUE))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.01, 0.99))

##       1%      99%
## 12.48738 15.03513

sim_samples <- replicate(k, sample(length_ft, n, replace = TRUE))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.025, 0.975))

##     2.5%    97.5%
## 12.82957 14.46058
```

1* ☐ [11.93, 15.23]

2 ☐ [11.59, 15.05]

3 ☐ [12.13, 14.50]

4 ☐ [12.49, 15.04]

5 ☐ [12.83, 14.46]

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

In case of non-parametric bootstrapping we sample with replacement from our sample data. We are still interested in a 99% confidence interval for the median, hence only answer 1 can be correct.

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

Continue on page 24

Exercise VII

Question VII.1 (17)

You have been collecting amber with a friend and you found in total 20 pieces. You agreed to share it by randomly drawing 10 pieces each. Three of the pieces are very attractive. What is the probability that you will get all three attractive pieces?

1 ☐ 0.0877%

2 ☐ 0.877%

3* ☐ 10.5%

4 ☐ 13.0%

5 ☐ 24.0%

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

This is hyper geometric, since it is drawing without replacement, so

```
dhyper(3, 3, 17, 10)

## [1] 0.1052632
```

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

Question VII.2 (18)

Let X represent the weight in grams of a new piece of amber that you find at your favourite location. From experience you know that when you find a piece of amber there, then its weight follows a log-normal distribution, such that $X \sim LN(1, 0.7^2)$.

What is the mean weight μ_X of amber pieces at your favourite location according to this model?

1 ☐ 2.01 g

2 ☐ 2.72 g

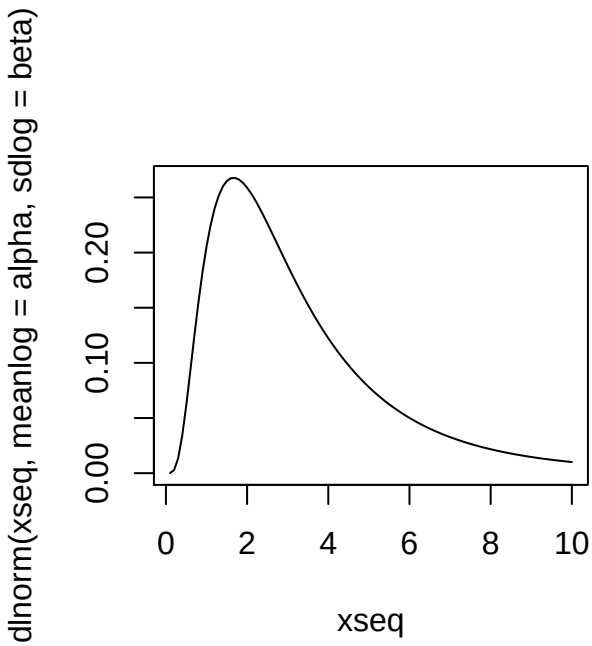
3* ☐ 3.47 g

4 ☐ 5.93 g

5 ☐ 9.21 g

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

```r
alpha <- 1
beta <- 0.7
##
xseq <- seq(0.1, 10, by=0.1)
plot(xseq, dlnorm(xseq, meanlog=alpha, sdlog=beta), type = "l")
##
exp(alpha+beta^2/2)

## [1] 3.472935
```



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

Question VII.3 (19)

Based on the information given in the last question: If you find 20 pieces at your favourite location, what is the probability that at least 3 of them weigh more than 10 grams?

1 ☐ 0.31%

2* ☐ 2.36%

3 ☐ 3.14%

4 ☐ 4.24%

5 ☐ 12.31%

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

Now this is drawing with replacement, since every time we find a new piece it's from an "infinite" sized population. So first we find the probability that a new piece is more than 10 grams i.e.

$$P(X > 10) = 1 - P(X < 10)$$

(X is weight), so in R:

```
(p <- 1 - plnorm(10, meanlog=alpha, sdlog=beta))
```

```
## [1] 0.03138368
```

and this is the success probability in the binomial drawing. The probability of finding 3 or more pieces is then

$$P(Y \geq 3) = 1 - P(Y \leq 2)$$

```
1 - pbinom(2, 20, p)
```

```
## [1] 0.02363236
```

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

Continue on page 27

Exercise VIII

Let the random variable X_i represent the i'th observation in a sample of n observations from a population which is uniformly distributed between α and β. The observations are sampled randomly and thus independently of each other. So $X_i \sim U(\alpha, \beta)$ and i.i.d.

Question VIII.1 (20)

The sample mean is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

What is the distribution of $\bar{X}$ as n goes to infinity?

1 ☐ $N(0, 1^2)$

2 ☐ $U(\alpha, \beta)$

3 ☐ t-distribution with $n-1$ degrees of freedom

4* ☐ $N(\frac{\alpha+\beta}{2}, \frac{(\beta-\alpha)^2}{12n})$

5 ☐ $U(\alpha^n, \beta^n)$

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

From the CLT Theorem 3.14 we know that the sample mean, i.e. a sum of i.i.d. random variables, is normal distributed, with same mean and variance divided by the number of variables n.

The mean and variance of the uniform distribution is found in Theorem 2.36, which inserted gives the answer.

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

Question VIII.2 (21)

Define $Y_i = 2 + \frac{1}{10}X_i$, which of the following statements is correct?

1 ☐ $\mathrm{E}(Y_i) = \frac{1}{10}\mathrm{E}(X_i)$

2 ☐ $\mathrm{E}(Y_i) = \frac{1}{100}\mathrm{E}(X_i)$

3 ☐ $\mathrm{V}(Y_i) = \frac{1}{10}\mathrm{V}(X_i)$

4* ☐ $\mathrm{V}(Y_i) = \frac{1}{100}\,\mathrm{V}(X_i)$

5 ☐ $Y_i \sim U(\alpha, \beta)$

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

We use the identities for linear variables in Theorem 2.54.

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

Continue on page 29

Exercise IX

In power systems the balancing power is the generation or load which can quickly be increased or decreased to stabilize the voltage on the grid. The balancing power is often traded on a market, as on the Dutch aFRR market, where bids are settled for 15 minute intervals. If you participate on such a marked, it is important to know how much energy is activated.

First the activated up-regulation volume is analyzed, that is how much energy in total was activated for increased generation per day. The average daily values in MWh for three winter months are read into the vector `xwinter` and the following analysis is carried out

```
t.test(xwinter)

##
##  One Sample t-test
##
## data:  xwinter
## t = 14, df = 89, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   9.346 12.341
## sample estimates:
## mean of x
##     10.84
```

Question IX.1 (22)

Let μ_{winter} be the mean up-regulation volume on winter days. Assuming a significance level $\alpha = 0.05$, what should be the conclusion on the following null hypothesis (both conclusion and argument must be correct)?

$$H_0 : \mu_{\text{winter}} = 10$$

1 ☐ The null hypothesis is rejected, since the p-value is below $2 \cdot 10^{-16}$ which is below 5%

2 ☐ The null hypothesis is accepted, since the p-value is below $2 \cdot 10^{-16}$ which is below 5%

3 ☐ The null hypothesis is rejected, since the p-value is below $2 \cdot 10^{-16}$ which is above 5%

4 ☐ The null hypothesis is accepted, since the p-value is below $2 \cdot 10^{-16}$ which is above 5%

5* ☐ The null hypothesis is accepted, since 10 is inside the 95% confidence interval

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

It's clear that the mean under the null hypothesis (μ_0) is inside the condidence interval, in which case we know that the null hypothesis will not be rejected, i.e. it must be accepted.

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

Question IX.2 (23)

What is the 99% confidence interval for μ_{winter}?

1 ☐ $[7.77, 13.91]$

2 ☐ $[8.01, 12.10]$

3 ☐ $[8.28, 13.41]$

4* ☐ $[8.86, 12.82]$

5 ☐ $[9.35, 12.34]$

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

Half the width of the confidence interval is

$$t_{0.975}\frac{s}{\sqrt{n}} = (12.341 - 9.346)/2 = 1.4975$$

so by looking up $t_{0.975}$ in R

```
qt(0.975, df=89)
```

```
## [1] 1.987
```

we find the standard error to

$$\frac{s}{\sqrt{n}} = \frac{1.4975}{t_{0.975}} = 0.7537$$

so we can find the 99% confidence interval by

$$\bar{x} \pm t_{0.995}\frac{s}{\sqrt{n}}$$

```
10.84 + c(-1,1) * qt(0.995, df=89) * 0.7537
```

```
## [1]  8.856 12.824
```

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

Question IX.3 (24)

What is the number of observations in `xwinter`?

1 ☐ 88

2 ☐ 89

3* ☐ 90

4 ☐ 91

5 ☐ 92

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

In a one-sample t-test we know that the degrees of freedom is $n - 1$, and since df is 89, then n is 90.

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

Question IX.4 (25)

In order to find out if there is a difference between winter and summer, the daily averages of up-regulation volume for the summer months in the same year are loaded into `xsummer`.

Based on the given data in the exercise, which of the following tests is best suited for concluding if there is a significant difference between the daily mean of up-regulation volume in winter and in summer?

1* ☐ A two-sample t-test

2 ☐ A paired two-sample t-test

3 ☐ A two-way ANOVA test

4 ☐ A test for the slope coefficient in a linear regression model

5 ☐ A χ^2-test

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

We have two samples, one from winter and one from summer, so it's a two sample test. They cannot be paired, since they are not on same dates and don't share other features that we are informed about.

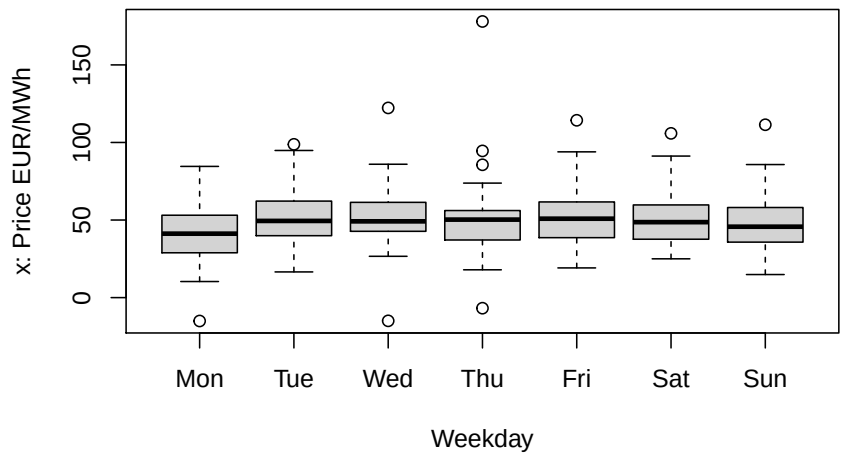
---------- FACIT-END ----------

Continue on page 33

Exercise X

This exercise is about the Dutch aFRR balancing power market as mentioned in the previous exercise. For providers of flexible power it is important to investigate the prices at which the balancing power is sold and bought on the market. A year of daily average price of down-regulation power is read into `x`. 364 observations (days) were included in the data.

To see if there are differences between the days of the week, box-plots are generated for each day (note that the prices are given per energy unit, this detail doesn't matter in this exercise):



A one-way ANOVA was carried out. The result are given below:

```
anova(lm(x ~ weekday))

## Analysis of Variance Table
##
## Response: x
##            Df Sum Sq Mean Sq F value  Pr(>F)
## weekday     6   4934  822.42  2.0969 0.05296 .
## Residuals 357 140016  392.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question X.1 (26)

Given a significance level of 5%, what is the critical value for the F-test of equal weekday means?

1 ☐ 1.549

2 ☐ 1.791

3 ☐ 1.943

4* ☐ 2.124

5 ☐ 2.444

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

The test statistic follows an F-distribution under the null hypothesis, see Theorem 8.6 and the critical value is the $1-\alpha$ quantile in the F-distribution with $k-1$ and $n-k$ degrees of freedom, and $k = 7$ different weekdays, so

```
qf(0.95, 6, 357)

## [1] 2.123994
```

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

Question X.2 (27)

Assuming that all model assumptions are fulfilled, what is the estimate of the variance of the daily average down-regulation price on Fridays using this model (both value and explanation must be correct)?

1* ☐ $\hat{\sigma}^2 = 392.2$, since the variance estimate is pooled and thus it is the same for all weekdays

2 ☐ $\hat{\sigma}^2 = \frac{140016}{4934} = 28.38$, since the variance estimate is pooled and thus it is the same for all weekdays

3 ☐ $\hat{\sigma}^2 = \frac{140016}{7} = 20002$, since the variance estimate must be split on the different weekdays, thus adjusted by the degrees of freedom for `weekdays`

4 ☐ $\hat{\sigma}^2 = \frac{140016}{6} = 23336$, since the variance estimate must be split on the different weekdays, thus adjusted by the degrees of freedom for `weekdays`

5 ☐ This cannot be calculated with the given information

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

Since one of the model assumptions for the model in the ANOVA, is that the variance is homogeneous, meaning that it's the same for all groups, then the variance is pooled. We can read it off directly from the ANOVA table printed in the result.

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

Question X.3 (28)

What is the proportion of variance explained by the model?

1 ☐ 0.57%

2* ☐ 3.4%

3 ☐ 18.4%

4 ☐ 32.3%

5 ☐ 96.6%

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

It's the proportion of variance explained by the "treatment" (here `weekday`) of the total variance SST, hence

```
4934 / (140016 + 4934)

## [1] 0.03403932
```

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

Continue on page 36

Exercise XI

The following sample is available:

```
  x <- c(1.11, 0.94, -2.43, -0.90,  0.29, -1.41,  0.38,  0.99, -0.50)
```

Question XI.1 (29)

What is the median of the sample?

1 ☐ -0.10

2 ☐ -0.17

3 ☐ 0.34

4 ☐ 0.38

5* ☐ 0.29

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

We can do it very easily, since $n = 9$, hence uneven, it's the middle value of the sorted sample:

```
sort(x)

## [1] -2.43 -1.41 -0.90 -0.50  0.29  0.38  0.94  0.99  1.11
```

or just use the in built function:

```
  median(x)

## [1] 0.29

  quantile(x, type=2)

##    0%   25%   50%   75%  100%
## -2.43 -0.90  0.29  0.94  1.11
```

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

Question XI.2 (30)

Let the i'th observation of the sample be represented by X_i and assume $X_i \sim N(\mu, \sigma^2)$ where X_i is i.i.d.

What is the distribution of the sample mean $\bar{X}$?

1 ☐ The t-distribution with 8 degrees of freedom

2 ☐ The t-distribution with 9 degrees of freedom

3* ☐ The normal distribution $N(\mu, \sigma^2/n)$

4 ☐ The χ^2-distribution with 8 degrees of freedom

5 ☐ The χ^2-distribution with 9 degrees of freedom

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

We have Theorem 3.3, so that's how it works!

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

Continue on page 38

The exam is over! Enjoy your Christmas holidays!

Technical University of Denmark *Page 1 of 36 pages.*

Written examination: 17. May 2020

Course name and number: **Introduction to Statistics (02323)**

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 11 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	4	3	3	5	2	5	4	2	5

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

Exercise	VIII.3	VIII.4	VIII.5	VIII.6	IX.1	IX.2	X.1	X.2	X.3	XI.1
Question	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Answer	1	5	3	4	3	4	4	1	2	4

The exam paper contains 36 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 characteristics of electrical components are not exactly as specified, e.g. if you buy a resistor then the resistance through it is not exactly as specified. In the production of electric circuits, it is of great interest not to get too much variation in the quality of the overall circuit. An example is the resistance through two parallel connected resistors, which is calculated by

$$R = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2}}$$

where R_1 is the resistance through one of the resistors and R_2 through the other resistor. The resistance is measured in ohm.

Assume that $R_1 \sim N(4, 0.2)$ and $R_2 \sim N(2, 0.2)$.

Question I.1 (1)

You buy 100 R_1 resistors - which can be assumed to be independent of each other. What is the probability that none of these has a resistance below 3 ohms?

1 ☐ 1.27%

2 ☐ 2.78%

3 ☐ 13.9%

4* ☐ 27.9%

5 ☐ 42.4%

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

First calculate the probability that a single of them is not below 3:

```
pnorm(3, 4, sqrt(0.2))
```

```
## [1] 0.01267366
```

and now we have 100 independent draws (we draw them from an "infinite" population of resistors), hence we use the binomial distribution calculating zero "successes"

```r
dbinom(0, 100, pnorm(3, 4, sqrt(0.2)))
```

```
## [1] 0.2793009
```

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

Question I.2 (2)

Calculate an estimate of the standard deviation of the total resistance R (the answer is rounded to two significant digits, tip: if you use simulation then remember to make sufficient repetitions to get a stable result)?

1 ☐ 0.026

2 ☐ 0.094

3 ☐ 0.16

4* ☐ 0.21

5 ☐ 0.44

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

Its a non-linear function, so use simulation

```r
  k <- 1000000
  R1 <- rnorm(k, mean=4, sd=sqrt(0.2))
  R2 <- rnorm(k, mean=2, sd=sqrt(0.2))
  R <- 1 / (1/R1 + 1/R2)
  sd(R)
```

```
## [1] 0.2093243
```

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

Continue on page 4

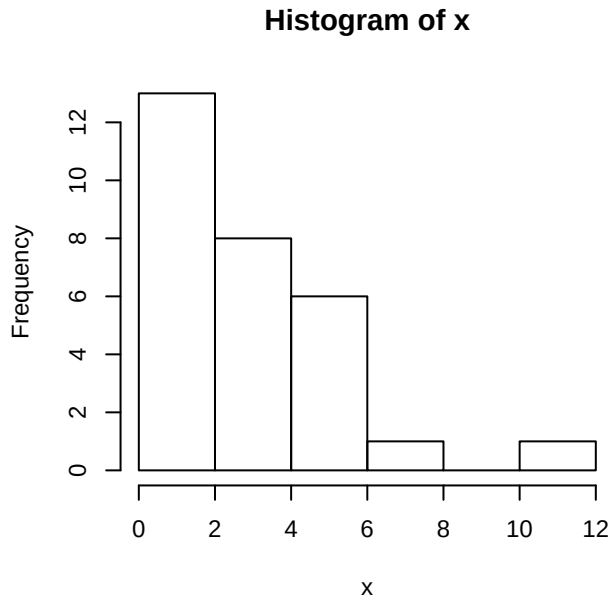
Exercise II

In a computer system an optimization routine is used and the execution time for this routine is under investigation. The execution time is measured in hours and loaded into R with the following code:

```
x <- c(1.6, 2, 3.4, 4, 2.1, 0.6, 0.4, 0.4, 6, 0.4, 4.9, 2, 2, 4.6, 0.5,
       3.4, 7.2, 10.5, 3.2, 1.3, 5.7, 1.9, 2.6, 2.5, 4.4, 1.8, 3.9, 6, 0.9)
```

Question II.1 (3)

It is desirable to assess which distribution the outcomes in the sample could originate from. Therefore the following histogram has been generated of the observations in `x`:



On the basis of the information given, evaluate which of the following distributions is most likely to have generated the observations in the sample?

1 ☐ A normal distribution

2 ☐ A Poisson distribution

3* ☐ An exponentiel distribution

4 ☐ A t-distribution

5 ☐ A binomial-distribution

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

Lets go through the answers:

- With most values in the interval closes to zero, but no values below, then it is very unlikely that it is from a normal distribution.
- It's not Poisson, since it is not integer values.
- It looks very must like an exponential distribution.
- Like the normal, it's very unlikely that it is a t-distribution.
- It's not Binomial, since it is not integer values.

Hence the only likely answer is the exponential distribution.

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

Question II.2 (4)

Based on the sample, what is the estimate of the mean and standard deviation of the computation times?

1 ☐ $\hat{\mu} = 2.53$ and $\hat{\sigma} = 1.66$

2 ☐ $\hat{\mu} = 3.36$ and $\hat{\sigma} = 0.48$

3* ☐ $\hat{\mu} = 3.11$ and $\hat{\sigma} = 2.37$

4 ☐ $\hat{\mu} = 1.98$ and $\hat{\sigma} = 5.63$

5 ☐ $\hat{\mu} = 3.96$ and $\hat{\sigma} = 2.81$

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

Copy from the pdf to read the sample into R:

```
x <- c(1.6, 2, 3.4, 4, 2.1, 0.6, 0.4, 0.4, 6, 0.4, 4.9, 2, 2, 4.6, 0.5,
3.4, 7.2, 10.5, 3.2, 1.3, 5.7, 1.9, 2.6, 2.5, 4.4, 1.8, 3.9, 6, 0.9)
```

and then

```
mean(x)

## [1] 3.11

sd(x)

## [1] 2.37
```

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

Question II.3 (5)

You want to give a guarantee that the execution time is below a certain level, and therefore a confidence interval of the 90% quantile should be calculated. A function is defined in R to calculate it by:

```
q90 <- function(x){ quantile(x, prob=0.9, type=2) }
```

Which of the following R codes calculates a 95% percent non-parametric bootstrap confidence interval for the 90% quantile of the distribution of computation times?

1 ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.05, 0.95))
```

2 ☐
```
simsamples <- replicate(10000, sample(x, replace = FALSE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.025, 0.975))
```

3 ☐
```
simsamples <- replicate(10000, sample(x, replace = FALSE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.05, 0.95))
```

4 ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.1, 0.90))
```

5* ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.025, 0.975))
```

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

The differences in the answers are:

- replace: must be TRUE
- quantiles calculated: they must be 2.5% and 97.5% to have the right significance level with 95% percent in between

So

```
simsamples <- replicate(10000, sample(x, replace = TRUE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.025, 0.975))

##  2.5% 97.5%
##   4.6  10.5
```

See Section 4.3.

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

Continue on page 8

Exercise III

An NGO has 15 callers employed to recruit new members. Let X represent the number of members a single caller recruits during one working day. The number of new members each caller recruits in a day can be assumed to be independent of each other. From experience it is known that a good model for X is a binomial distribution, where the probability of getting a new member in a call is 7%. It is assumed that each caller can do 120 calls in one day.

Question III.1 (6)

What is the probability that a caller on a single day recruits more than 5 new members?

1 ☐ 0.12

2* ☐ 0.85

3 ☐ 0.45

4 ☐ 0.17

5 ☐ 0.96

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

It's a binomial setup, so discrete, and the probability is

$$P(X > 5) = 1 - P(X \leq 5)$$

which is found in R by

```
n <- 120
p <- 0.07
1 - pbinom(5, n, p)

## [1] 0.8522782
```

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

Question III.2 (7)

If Y is the total number of new members the 15 callers can recruit in a day, what is the mean and variance of Y?

1 ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 10.8$

2 ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 41.9$

3 ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 43.5$

4 ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 102.4$

5* ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 117.2$

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

The mean and variance is given for the binomial distribution in Theorem 2.21. Hence $\mu_X = np = 8.4$ and $\sigma_X^2 = np(1-p) = 7.812$.

We have the total recruitment of the 15 callers by the sum

$$Y = \sum_{1}^{15} X$$

Applying the rules in Theorem 2.56, we get $\mathrm{E}(Y) = \sum_1^{15} 8.4 = 126$ and $\mathrm{V}(Y) = \sum_1^{15} 7.812 = 117.18$.

The trick is not to make

$$Y = 15X$$

which it would be only the recruitment of a single caller times 15, which would give a wrong variance $\mathrm{V}(Y) = \mathrm{V}(15X) = 15^2 \cdot 7.812 = 1757.7$! Well, it's not among the answers, to that made it less of a pitfall.

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

Continue on page 10

Exercise IV

In Denmark, people often discuss whether it has been a good or a bad summer. The table below shows the average temperatures for the months May to September in the years 2014-2018:

	2014	2015	2016	2017	2018	Average
May	11.7	9.7	12.9	12.0	15.0	12.26
June	14.9	12.7	16.0	14.7	16.5	14.96
July	19.5	15.5	16.4	15.5	19.2	17.22
August	16.0	17.4	16.1	16.0	17.5	16.60
September	14.6	13.2	16.2	13.3	14.1	14.28
Average	15.34	13.70	15.52	14.30	16.46	15.01

To investigate if there is a difference between the years, the following R code has been run, where `year` is indicator for the years 2014-2018 and `temp` is the average temperature:

```
anova(fit <- lm(temp ~ year))

## Analysis of Variance Table
##
## Response: temp
##           Df Sum Sq Mean Sq F value Pr(>F)
## year       4   23.4    5.85    1.15   0.36
## Residuals 20  101.7    5.08
```

Question IV.1 (8)

At the significance level $\alpha = 0.05$ what is the conclusion from the R code above (by difference is meant significant difference in mean. Remember all parts of the answer must be correct)?

1 ☐ There is no difference in temperature between the years, since $5.85 > 5.08$.

2 ☐ There is a difference in temperature between the years, since $0.36 > 0.05$.

3 ☐ There is a difference in temperature between the years, since $23.4 < 101.7$.

4* ☐ There is no difference in temperature between the years, since $0.36 > 0.05$.

5 ☐ There is a difference in temperature between the years, since $0.36 < 5$.

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

We read the p-value for the test for a significant effect of year under `Pr(>F)` and compare it to the significance level. See Theorem 8.6.

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

Question IV.2 (9)

The model used in the test above can be written as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \text{ and i.i.d.}$$

What is the estimate of σ^2?

1 ☐ $\hat{\sigma}^2 = (23.4 + 101.7)/(4 + 20)$

2* ☐ $\hat{\sigma}^2 = 101.7/20$

3 ☐ $\hat{\sigma}^2 = \sqrt{101.7}$

4 ☐ $\hat{\sigma}^2 = (5.85 + 5.08)/(4 + 20)$

5 ☐ $\hat{\sigma}^2 = \sqrt{5.08}$

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

It's the value listed under `Mean Sq`. It is $MSE = \frac{SSE}{n-k}$, see the table on the page after Theorem 8.6.

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

Question IV.3 (10)

What is

$$\sum_{\text{all } j,i} (y_{ij} - \bar{y})^2$$

where y_{ij} represents the observed temperatures and $\bar{y}$ is the average of all the temperatures?

1 ☐ 106.78

2 ☐ 10.93

3 ☐ 29.25

4 ☐ 1.15

5* ☐ 125.1

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

You have to recognize this formula as the total variance: SST. It's given in Theorem 8.2 and it is calculated by summing all the variances in the table:

```
23.4+101.7
```

```
## [1] 125
```

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

Continue on page 13

Exercise V

A sample of $n = 50$ observations was taken from a population. The population distribution is not known, but it is known that the mean μ is equal to 0. Let the i'th observation be represented by X_i and the sample mean by $\bar{X}$. The sample standard deviation is denoted by s.

Question V.1 (11)

You now want to get an idea of the distribution of the population, which of the following actions would be the most obvious to carry out?

1* ☐ Make a density histogram of the sample.

2 ☐ Make a scatter plot of the population.

3 ☐ Make a qq-normal plot of the sample.

4 ☐ Make a pie chart of the population.

5 ☐ Make a t-test.

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

Going through the answers:

- Yes, that gives a view on the distribution of the sample
- We can't really make a scatter plot with only one variable (i.e. we need two paired samples)
- A qq-normal plot gives a good idea about if the sample is normal distributed, but not easy to see if it's another distribution
- It's not really useful for a single sample (as presented in the book)
- Not meaningful to see the distribution

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

Question V.2 (12)

The following transformation has been defined

$$Y = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Which of the following distributions approximates the distribution of Y?

1 ☐ An exponential distribution.

2 ☐ A t-distribution with 2 degrees of freedom.

3 ☐ A χ^2-distribution.

4 ☐ An F-distribution.

5* ☐ A standard normal distribution.

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

According to the Central Limit Theorem (CLT) 3.14 the sample mean is approximated well with a standard normal distribution if the sample size is above 30 observations. Note in the Theorem the population standard deviation σ is used, but when we replace with the sample standard deviation s, then it becomes a t-distribution (see Theorem 3.5), but when we are over 30 observations, then the t and the normal is very close, so then we can use either.

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

Continue on page 15

Exercise VI

In a research project on energy consumption in schools, it was investigated how students and teachers set the radiator thermostats in classrooms. Thermostat settings were recorded during a period with cold weather at a number of schools in randomly selected classrooms. It was predetermined that thermostats are well set if the setting is between 2 and 3 on all radiators in a room, as it otherwise indicates under or oversized radiators. Besides, it's not desirable for thermostats to be set differently in a room, as this results in inferior comfort and poorer return water cooling.

The following observations were made during the period:

	School 1	School 2	School 3	School 4
Not-well set	18	11	22	9
Well set	38	36	15	12

Hence, at School 2 there were 11 out of 47 rooms in which the thermostats were not well set.

Question VI.1 (13)

What is the 95% confidence interval for the proportion of thermostats that were not set well at School 1 (note that the result from the R functions and the books formula may be slightly different, but both results are always closest to the correct answer)?

1 ☐ $[0.07, 0.17]$

2* ☐ $[0.20, 0.44]$

3 ☐ $[0.26, 0.53]$

4 ☐ $[0.05, 0.22]$

5 ☐ $[0.18, 0.51]$

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

We use the CI formula in Method 7.3:

```
p <- 18/(18+38)

## With the formula of the book
p - qnorm(0.975) * sqrt(p*(1-p)/(18+38))

## [1] 0.1991

p + qnorm(0.975) * sqrt(p*(1-p)/(18+38))

## [1] 0.4437
```

Note that this gives a slightly different result than the R function, see Remark 7.8. It's still closest to the correct answer, since it gives:

```
## With the R function
prop.test(x = 18, n = 18+38, correct=FALSE)$conf.int

## [1] 0.2140 0.4518
## attr(,"conf.level")
## [1] 0.95
```

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

Question VI.2 (14)

It was planned to compare schools to investigate if there were differences in practice of thermostat setting at the schools. Under the null hypothesis of no difference, then what is the expected number of not-well set thermostats at School 3?

1* ☐ $e_{13} = 37 \cdot \frac{60}{161} = 13.8$

2 ☐ $e_{13} = 15 \cdot \frac{22}{37} = 8.9$

3 ☐ $e_{13} = 60 \cdot \frac{124}{161} = 46.2$

4 ☐ $e_{13} = 22 \cdot \frac{22}{37} = 13.1$

5 ☐ $e_{13} = 15 \cdot \frac{124}{161} = 11.6$

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

We find the formula for expected counts under the null hypothesis that all proportions are equal in Section 7.4.

So we sum the row of Not-well set:

```
18+11+22+9

## [1] 60
```

and the total number is

```
60 + 38 + 36 + 15 + 12
```

```
## [1] 161
```

so we have 37 counts at School 3,

```
37 * 60/161
```

```
## [1] 13.79
```

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

Question VI.3 (15)

You want to investigate whether there was a difference in the proportion of not-well set thermostats at the four schools. Use a significance level of 1%. What will be the conclusion (both conclusion and argument must be correct)?

1 ☐ A difference between the schools cannot be detected, since the relevant test statistic is below the critical level of 15.4.

2 ☐ A difference between the schools can be detected, since the relevant test statistic is below the critical level of 11.3.

3 ☐ A difference between the schools cannot be detected, since the relevant test statistic is above the critical level of 0.0057.

4 ☐ A difference between the schools cannot be detected, since the relevant test statistic is below the critical level of 0.0057.

5* ☐ A difference between the schools can be detected, since the relevant test statistic is above the critical level of 11.3.

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

If we don't want to do all the calculations, then we can put in R (we could have done that already) by:

```
M <- as.table(rbind(c(18,11,22,9),c(38,36,15,12)))
M
```

```
##    A  B  C  D
## A 18 11 22  9
## B 38 36 15 12
```

```r
## Chi^2 test
(Xsq <- chisq.test(M))

##
##  Pearson's Chi-squared test
##
## data:  M
## X-squared = 13, df = 3, p-value = 0.006

## Observed statistic
Xsq$statistic

## X-squared
##     12.57

## p-value
1 - pchisq(Xsq$statistic, df=ncol(M)-1)

## X-squared
##  0.005671

## Critical value
qchisq(0.99, df=length(x)-1)

## [1] 11.34
```

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

Continue on page 19

Exercise VII

The following 3 questions deal with various statistical problems that may arise when treating water.

Question VII.1 (16)

When controlling drinking water the quality is measured by regular water analysis. There is, of course, legislation on quality among other things requirements for concentrations of different substances. One requirement is that the conductivity of the water at the consumer's tap must not be more than 2500 μS/cm at 20 °C. If the conductivity is above this level, the concentration of salts is too high and the water is referred to as aggressive.

The conductivity of water at randomly selected consumers' taps has been measured. Let a measurement be represented by the stochastic variable X_i, which can be assumed to be normally distributed. Twenty independent measurements have been collected to determine the conductivity and the observed values are stored in the vector `x` in R. You now want to test if the water on average is aggressive, and the following null hypothesis about the mean value, μ, of the conductivity in the drinking water at the consumer's tap, is formulated

$$H_0 : \mu = 2500$$

with the alternative hypothesis

$$H_1 : \mu \neq 2500$$

The following from R is given:

```
t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 4.8527, df = 19, p-value = 0.0001106
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   704.1022 1772.1090
## sample estimates:
## mean of x
##  1238.106
```

Based on the result above and with a 5% significance level, which of the following statements is correct (both conclusion and reasoning must be correct)?

1 ☐ We reject H_0 as the relevant confidence interval does not contain 0.

2 ☐ We accept H_0, since the sample average is within the relevant confidence interval.

3 ☐ We accept H_0, since the test statistic is greater than 1.96.

4 ☐ We reject H_0, since the p-value is 0.0001106.

5* ☐ We reject H_0, since 2500 is not within the relevant confidence interval.

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

From the R-output we see that the 95% confidence interval is 704-1772, i.e. does not contain 2500. Thus we reject the null-hypothesis, since the 2500 is outside the acceptance region (=the confidence interval). Answer 4 is not correct because the p-value of the `t.test` result above was calculated for a different Null hypothesis, i.e. $H_0 : \mu = 0$.

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

Question VII.2 (17)

At waterworks the water is purified by one of two methods, A or B, and the remaining concentration of a substance is measured. Measurements of remaining concentration are given in the stochastic variables $Y_{A,i}$ and $Y_{B,i}$ for methods A and B, respectively. It is of interest to investigate which of the two methods best purifies the water. $Y_{A,i}$ and $Y_{B,i}$ can be assumed to be normally distributed and their variances, σ_A^2 and σ_B^2, can be assumed to be equal. For each of the two methods a sample of 20 observations is taken. We want to test the null hypothesis

$$H_0 : \quad \mu_A = \mu_B$$

where the alternative hypothesis is

$$H_1 : \quad \mu_A \neq \mu_B$$

Which of the following procedures is a correct approach?

1 ☐ A paired t-test with 19 degrees of freedom.

2 ☐ A paired t-test with 18 degrees of freedom.

3* ☐ A two-sample t-test with 38 degrees of freedom.

4 ☐ A two-sample t-test with 39 degrees of freedom.

5 ☐ A F-test testing variance homogeneity.

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

We use Welch's t-tests where the number of degrees of freedom is given by (3.50):

$$\nu = (\sigma_A^2/20 + \sigma_B^2/20)^2)/((\sigma_A^2/20)^2/(20-1) + (\sigma_B^2/20)^2/(20-1))$$

since $\sigma_A = \sigma_B$ we have

$$\nu = (2\sigma_A^2/20)^2/(2(\sigma_A^2/20)^2/(20-1)) = 2(20-1) = 38$$

Alternatively, we could have used the pooled t-test because the variances of both samples are similar (3.53.

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

Question VII.3 (18)

A new study is planned on the concentration of a substance in a drinking water drilling. We want to achieve a power of 90% to detect a mean value difference of 2 units from a given value. From experience it is known that the standard deviation is 3.5 units and you want to perform the test at a 5% significance level (remember that results from R functions may differ slightly from the result obtained with the book's formulas). How many observations must be taken to fulfil these requirements?

1 ☐ At least 15 observations.

2* ☐ At least 35 observations.

3 ☐ At least 48 observations.

4 ☐ At least 67 observations.

5 ☐ At least 102 observations.

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

We can either use the formula in Method 3.65 or the function in R:

```
## The formula
zb <- qnorm(0.9)
za <- qnorm(1-0.05/2)
delta <- 2
sigma <- 3.5

(sigma * (zb+za)/delta)^2
```

```
## [1] 32.17898

## The R function
power.t.test(power=0.9, delta=2, sd=3.5, sig.level=0.05, type="one.sample")

##
##      One-sample t test power calculation
##
##               n = 34.15781
##           delta = 2
##              sd = 3.5
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
```

The results of the formula and from the R output differ slightly due to rounding. However, both indicate that only answer 2 can be correct.

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

Continue on page 23

Exercise VIII

On 30 randomly selected summer days, corresponding values of the temperature at noon, x measured in degrees Celsius, and the number of ice creams sold in an ice cream chain, Y, have been recorded. The following model has been fitted in R:

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

The result of this is shown below:

```
summary(fit1)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -161.24  -81.60  -46.14  103.83  249.85
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2382.001    116.620  -20.43   <2e-16 ***
## x             230.703      5.083   45.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 124.7 on 28 degrees of freedom
## Multiple R-squared:  0.9866,Adjusted R-squared:  0.9861
## F-statistic:  2060 on 1 and 28 DF,  p-value: < 2.2e-16
```

Question VIII.1 (19)

Based on the above R output, what is the estimate of the variance of the errors $\hat{\sigma}^2$?

1 ☐ 116.6

2 ☐ 116.6^2

3 ☐ 124.7

4* ☐ 124.7^2

5 ☐ $(230.7/28)^2$

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

$\hat{\sigma}$ is seen directly from "Residual standard error". Thus the variance is obtained by squaring this value.

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

Question VIII.2 (20)

Based on the above R output, what is the prediction of the mean value of ice creams sold, $\hat{y}_{\text{new}}$, at $x_{\text{new}} = 25°$ C?

1 ☐ 231 ice creams.

2 ☐ 2382 ice creams.

3* ☐ 3386 ice creams.

4 ☐ 5768 ice creams.

5 ☐ 11535 ice creams.

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

```
-2382 + 230.7 * 25
```

```
## [1] 3385.5
```

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

Question VIII.3 (21)

Based on the above R output, what is the critical value for the test

$$H_0 : \beta_1 = 0$$

using a 1% significance level?

1* ☐ 2.76

2 ☐ 2.05

3 ☐ 1.96

4 ☐ 1.70

5 ☐ 2.56

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

We know that the standardized parameters follow a t-distribution under H_0 from Theorem 5.12. Hence, as in Example 5.13, we find the critical values of the test by:

```
qt(0.995, df=28)

## [1] 2.763262
```

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

Question VIII.4 (22)

Which of the following statements about a prediction interval for $Y_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} + \varepsilon_{\text{new}}$ is not correct?

1 ☐ The prediction interval is wider than a corresponding confidence interval.

2 ☐ The width of the prediction interval depends on the sample size.

3 ☐ The prediction interval is symmetrical around the predicted value.

4 ☐ The width of the prediction interval depends on the value of x_{new}.

5* ☐ If the sample size becomes large enough, then the width of the prediction interval becomes 0.

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

The prediction interval is an interval for a new value, thus the width will not be smaller than $2 \cdot t_{1-\alpha/2}\sigma$, where σ represents the standard error of the predicted value $\hat{y}_{new}$ (5.18).

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

Question VIII.5 (23)

It is suspected that the linear model is not a correct model, so now you instead fit the model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$. The model is fitted by (note that `x2` is `x` squared):

```r
x2 <- x^2
fit2 <- lm(y ~ x + x2)
summary(fit2)

##
## Call:
## lm(formula = y ~ x + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -193.67  -59.32  -25.73   64.96  263.87
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -159.3055   470.1439  -0.339    0.737
## x             24.9850    42.9277   0.582    0.565
## x2             4.5715     0.9502   4.811 5.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 93.15 on 27 degrees of freedom
## Multiple R-squared:  0.9928,Adjusted R-squared:  0.9922
## F-statistic:  1856 on 2 and 27 DF,  p-value: < 2.2e-16
```

Based on this R output and with a 1% significance level, what can now be concluded about the relationship between ice cream sales and temperature (both conclusion and argument must both be correct)?

1 ☐ The relationship differs statistically significantly from a linear relationship, since $\hat{\beta}_2$ is positive.

2 ☐ The relationship does not differ statistically significantly from a linear relationship, since the p-value for $\hat{\beta}_1$ is above 1%.

3* ☐ The relationship differs statistically significantly from a linear relationship, since the p-value for $\hat{\beta}_2$ is below 1%.

4 ☐ We cannot reject that the relationship is linear, since $\hat{\beta}_2 < \hat{\beta}_1$.

5 ☐ We cannot reject that the relationship is linear, since $R^2 \approx 1$.

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

From the output we see that the p-value corresponding to the second order term is below 1%.

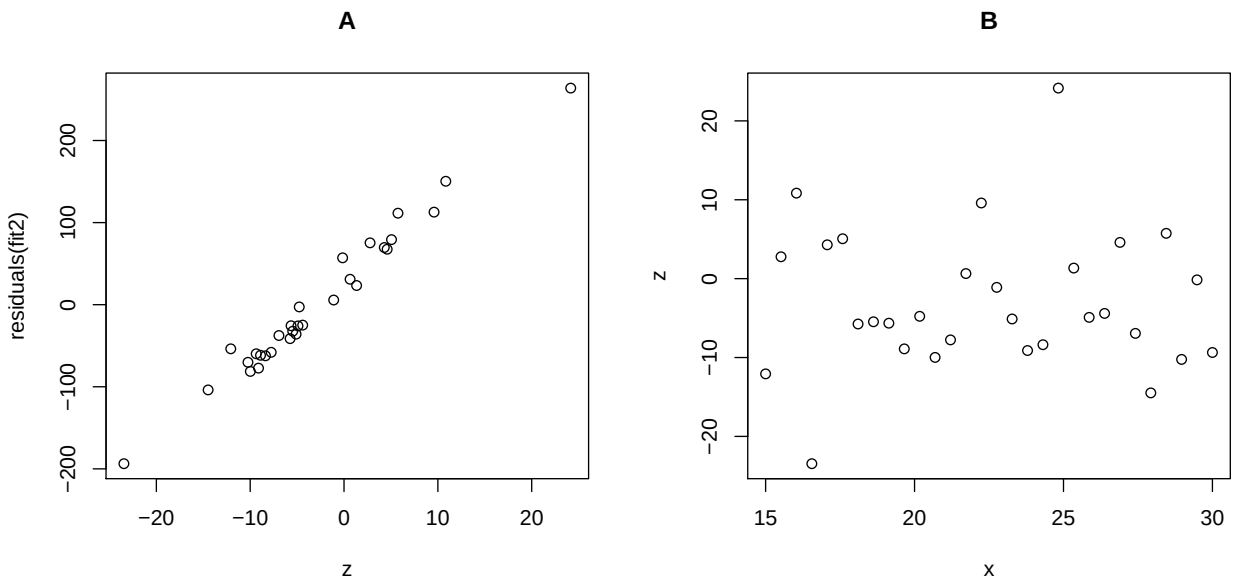
---------- FACIT-END ----------

Question VIII.6 (24)

In this exercise the result from the fit of the model from the previous question is investigated.

Below are shown two plots (A and B), where:

- z is a new variable, which represent measurements of sunshine on a day (the unit doesn't matter).
- x is the temperature on a day.
- `residuals(fit2)` is the residuals from the fit.



Which of the following statements is correct?

1 ☐ Figure B is used to examine whether the assumption of variance homogeneity is met.

2 ☐ Figure A is used to investigate whether the normal distribution assumption is met.

3 ☐ Figure B indicates a strong relationship between z og x.

4* ☐ Figure A can be used to investigate whether z should be included in the model.

5 ☐ Figure B can be used to check whether the relationship between x and Y is modelled correctly.

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

In the residual plot (A) we see a clear pattern. Because there is a clear relationship between z and the residuals, it indicates that z should be included in the model (5.28).

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

Continue on page 29

Exercise IX

The number of shooting stars per hour, X, is given by $X \sim Po(3)$, i.e. Poisson distributed with mean 3 shooting stars per hour.

Question IX.1 (25)

If one counts shooting stars for four hours, how many shooting stars might one expect to see (remember that the expectation value is equal to the mean)?

1 ☐ 8

2 ☐ 9

3* ☐ 12

4 ☐ 16

5 ☐ 24

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

We scale λ by 4 and obtain: $\lambda_4 = 4 \cdot 3 = 12$

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

Question IX.2 (26)

Suppose one has just observed a shooting star. What is then the probability of waiting more than 10 minutes for the next shooting star?

1 ☐ $P(X = 0)$, $X \sim Po(3)$

2 ☐ $P(X > 0)$, $X \sim Po(3)$

3 ☐ $P(Y > 10)$, $Y \sim Exp(3)$

4* ☐ $P(Y > \frac{10}{60})$, $Y \sim Exp(3)$

5 ☐ $P(Y > \frac{10}{6})$, $Y \sim Exp(3)$

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

Here we use the connection between Poisson process and the exponential distribution, i.e. the waiting times between poisson events are exponentially distributed. If Y is the waiting time

between two events, we need to find the probability $P(Y > 10 \text{ min}) = P(Y > 10/60 \text{ hours})$, where $Y \sim Exp(3)$.

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

Continue on page 31

Exercise X

In a production of steel pipes one is interested in the diameter of the pipes. Therefore, a sample of 30 tubes is taken and the sample variance is calculated to be $s^2 = 531$ mm^2.

Question X.1 (27)

What is the 99% confidence interval for the standard deviation of the diameter of the tubes?

1 ☐ [18.4, 31.0]

2 ☐ [18.1, 30.6]

3 ☐ [310, 1079]

4* ☐ [17.2, 34.3]

5 ☐ [294, 1175]

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

We use Method 3.19 to calculate the confidence interval. First the $1 - \alpha/2$ and $\alpha/2$ quantiles of the χ^2-distribution with 29 degrees of freedom are found in R:

```
qchisq(0.995, df=29)

## [1] 52.3

qchisq(0.005, df=29)

## [1] 13.1
```

$$\left[\sqrt{\frac{29 \cdot 531}{52.3}}, \sqrt{\frac{29 \cdot 531}{13.1}}\right] = [17.2, 34.3]$$

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

Question X.2 (28)

The diameter of the steel pipes is usually measured by one of two different measurement methods. It is now suspected that the two methods don't measure identically. 11 pipes are therefore randomly selected. Each pipe is now measured by both methods. The observation made with measurement Method 1 on pipe i is at position i in the vector `x` and corresponding measurement with Method 2 at position i in the vector `y`. The measurements by both methods can be assumed to be normally distributed.

The following analyses are now carried out in R:

```
t.test(x-y)

##
##  One Sample t-test
##
## data:  x - y
## t = -2.541, df = 10, p-value = 0.0293
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2.409246 -0.158027
## sample estimates:
## mean of x
##  -1.28364

t.test(x,y)

##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -0.1353, df = 20, p-value = 0.894
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.0683  18.5010
## sample estimates:
## mean of x mean of y
##   96.9018   98.1855
```

You want to test the null hypothesis that the two measurement methods have the same mean

$$H_0 : \mu_X = \mu_Y$$

against the alternative hypothesis that they are different

$$H_1 : \mu_X \neq \mu_Y$$

At a 5% significance level, which of the following possibilities is correct (both conclusion and argument must be correct)?

1* ☐ We reject H_0, since $p < 0.05$.

2 ☐ We cannot reject H_0, since $p = 0.89$.

3 ☐ We reject H_0, since the alternative hypothesis is different from 0.

4 ☐ We cannot reject H_0, since the test statistic, t_{obs}, is -0.14.

5 ☐ We reject H_0, since the difference between the sample averages is greater than 1.

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

It is a paired setup, and the measurements for each pipe is located at the same position in both vectors. Thus we use the output from the first call.

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

Question X.3 (29)

The company that produces the steel pipes is getting a new and faster machine for producing steel pipes. Regardless of the answer in the previous question, only measurement Method 1 is used below. The desired diameter of the steel pipes is 100 mm and the machine will be properly calibrated. You will plan a new experiment where you want to test the null hypothesis

$$H_0 : \mu_X = 100$$

against the alternative hypothesis

$$H_1 : \mu_X \neq 100$$

We use $\hat{\sigma}_x^2 = 502.23$, a significance level of 5% and the following R-code has been executed, since a sample size of $n = 40$ is wanted:

```
power.t.test(n=40, sd=sqrt(502.23), power=0.9, type="one.sample")

##
##      One-sample t test power calculation
##
##               n = 40
##           delta = 11.8
##              sd = 22.4
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
```

Based on the R code above, what can now be concluded before the experiment is performed (both argument and conclusion must be correct)?

1 ☐ If the true mean is 88.2 or lower, we have more than 90% chance of rejecting H_1.

2* ☐ If the true mean is 88.2 or lower, we have more than 90% chance of rejecting H_0.

3 ☐ If the true mean is 88.2 or lower, we will reject H_0.

4 ☐ If the true mean differs with less than 11.8 we accept H_0.

5 ☐ If the true mean differs with less than 11.8 we reject H_0.

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

Delta is 11.8. To obtain a power of at least 90% we need to be at least as far away from the hypotesized mean as given by delta. This corresponds to detecting a true mean of 88.2 or below or 111.8 and above. Therefore only answer 2 can be correct.

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

Continue on page 35

Exercise XI

In the production of a particular type of plate, each plate has a 20% probability of having a flaw. A random sample of 10 plates is now taken.

Question XI.1 (30)

What is the probability that at most 3 plates in the sample have a flaw?

1 ☐ 0.95

2 ☐ 0.32

3 ☐ 0.68

4* ☐ 0.88

5 ☐ 0.60

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

$X \sim B(10, 0.2)$. The Probability in question is $P(X \leq 3)$. It can be found using the following R code:

```
pbinom(3,size=10,prob=0.2)

## [1] 0.8791261
```

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

The exam is finished. Have a great summer!