Technical University of Denmark *Page 1 of 34 pages.*

Written examination: 18. December 2021

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

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

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

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

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. 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

For various reasons it can be interesting to analyze birth data - for example if one is interested in whether the number of births depends on the season. Birth data from Denmark is available from Statistics Denmark. Actually, by just visually inspecting the available birth data, it is easy to see that there are more births during summer than during winter. However, it is not clear if there is a difference between spring and fall.

To investigate this difference, the number of births in spring and fall each year in the period 2007 to 2020 was obtained. The differences are listed below (a positive difference means more births in the fall):

2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
826	435	-504	247	-211	357	-570	601	1459	770	830	-156	748	309

The sample mean is $\bar{x} = 367.2$ and sample standard deviation is $s = 571.5$.

Question I.1 (1)

What is correct calculation of the 95% confidence interval for the mean difference between spring and fall?

1 ☐ $367.2 \pm 2.14\sqrt{\frac{571.5}{14}}$

2* ☐ $367.2 \pm 2.16\sqrt{\frac{571.5^2}{14}}$

3 ☐ $571.5 \pm 2.14\sqrt{\frac{367.2}{14}}$

4 ☐ $571.5 \pm 1.96\sqrt{\frac{367.2^2}{14}}$

5 ☐ $367.2 \pm 1.96\sqrt{\frac{571.5}{14}}$

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

We use Method 3.9 to calculate the interval, first find the quantile in the t-distribution

```r
qt(0.975, df=14-1)

## [1] 2
```

$$367.2 \pm 2.160369\sqrt{\frac{571.5^2}{14}} = [37.2, 697.2]$$

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

Question I.2 (2)

A test of no difference in mean between spring and fall must be carried out, at significance level 5%. What conclusion can be drawn from this test based on the given information (both conclusion and argument must be correct)?

1* ☐ Since the p-value is less than 5% a significantly higher mean in the fall is detected.

2 ☐ Since the p-value is greater than 5% no difference in mean is detected.

3 ☐ Since the p-value is less than 5% a significantly lower mean in the fall is detected.

4 ☐ Since the p-value is greater than 5% a significant difference in mean is detected.

5 ☐ Since the p-value is less than 1% a significantly lower mean in the fall is detected.

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

We have to calculate the p-value for the test for $\mu = 0$. It's Method 3.23

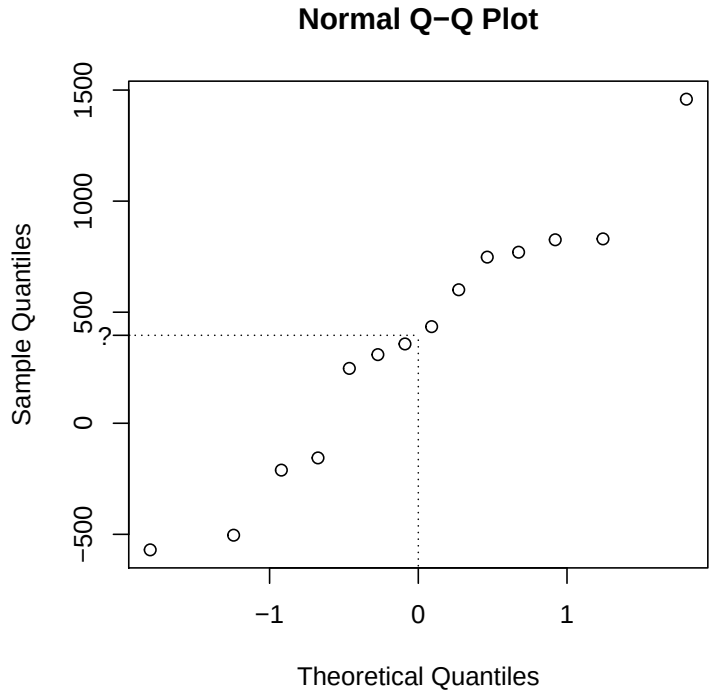
```r
tobs <- 367.2 / (571.5/sqrt(14))
tobs

## [1] 2

2 * (1 - pt(abs(tobs), df=14-1))

## [1] 0.03
```

It's below the significance level of 5% so the null hypothesis is rejected. The final step is to check the sample mean, which is higher than zero, hence we can conclude that there is a significant a higher number of births in the fall.

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

Question I.3 (3)

In the model validation the following normal q-q plot of the sample was generated for checking the assumption of normal distribution of the population.



On the y-axis a value is marked with "?". It's the value which, as indicated with the dotted lines, is in the middle of the two first points laying on either side of zero on the x-axis.

What is this value called?

1 ☐ The first quartile of the sample.

2 ☐ The third quartile of the sample.

3 ☐ The sample mean.

4* ☐ The median of the sample.

5 ☐ The Inter Quartile Range (IQR) of the sample.

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

The q-q plot is generated by sorting the observations and plotting them versus the i/(n+1) quantiles of the distribution, here the normal. Since the (standard) normal distribution is symmetric around zero, then the two first points on either side of zero are the two middle points in the sample (given equal number of observations). So it's the median, which is in the middle of those two points.

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

Continue on page 6

Exercise II

A sample was obtained, where 0 indicates a non-success and 1 indicates a success. The sample consists of 14 observations of value 0 and 18 observations of value 1.

Question II.1 (4)

What is the sample standard deviation?

1 ☐ $s = 0.254$

2 ☐ $s = 0.496$

3* ☐ $s = 0.504$

4 ☐ $s = 15.6$

5 ☐ $s = 16.1$

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

We can use the formula for sample standard deviation in Def. 1.11 and calculate it:

```
x <- c(rep(0,14), rep(1,18))
sqrt(1/31 * sum((x - mean(x))^2))

## [1] 0.5040161
```

or simply use:

```
sd(x)

## [1] 0.5040161
```

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

Question II.2 (5)

The sample was from a binomial experiment. The parameter p is the probability of a success.

What is the estimate of p?

1 ☐ 0.3164

2 ☐ 0.1914

3* ☐ 0.5625

4 ☐ 0.4375

5 ☐ 18

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

We use Method 7.3 for calculating the best estimate of p, so

```
x <- 18
n <- 18+14
x/n

## [1] 0.5625
```

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

Question II.3 (6)

Given $p = 0.5$, what is the probability of observing 18 or more successes in a new sample of the same size?

1* ☐ 0.298

2 ☐ 0.811

3 ☐ 0.702

4 ☐ 0.189

5 ☐ 0.110

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

```
1 - pbinom(17, 32, 0.5)

## [1] 0.2983074
```

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

Continue on page 8

Exercise III

Question III.1 (7)

We wish to simulate 35 samples from a normal distribution with mean = -2 and variance = 4. Which of the following code snippets does not generate this?

1 ☐ `rnorm(n = 35, mean = -2, sd = 2)`

2 ☐ `-2 + rnorm(n = 35, mean = 0, sd = 2)`

3* ☐ `rnorm(n = 35, mean = -2, sd = 1) * 4`

4 ☐ `-2 + rnorm(n = 35, mean = 0, sd = 1) * (-2)`

5 ☐ `-2 + rnorm(n = 35, mean = 0, sd = 4) / 2`

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

The third snippet generates random samples from $Y \sim N(8, 16)$. $Y = 4 \cdot X$, where $X \sim N(-2, 1)$, hence $Y \sim N(4 \cdot (-2), 4^2 \cdot 1) = N(8, 16)$.

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

Continue on page 9

Exercise IV

In an experiment, researchers measured the diameter of the top of the rootstock (in mm) and the fruit production (in mg) of 20 plants of the same biennial species.

Data are stored in R as `root`, and a linear regression $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ is carried out.

Question IV.1 (8)

First, the researchers calculate the sample correlation to be $r = 0.953$. Which of the following statements is correct (both conclusion and argument must be correct)?

1* ☐ There is a strong positive relationship between rootstock size and fruit production. A test of the hypothesis $\beta_1 = 0$ can confirm if this relationship is significant or not.

2 ☐ There is not a relationship between rootstock size and fruit production since r is between ± 1.96, where 1.96 is the 97.5% quantile in a standard normal distribution, $N(0, 1)$.

3 ☐ There is a strong relationship between rootstock size and fruit production, but we cannot tell if the relation is positive or negative.

4 ☐ There is a negative relationship between rootstock size and fruit production. Therefore, both x and y are normally distributed.

5 ☐ The sample correlation is a not an indicator of relationship between rootstock size and fruit production.

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

There is a strong positive relationship between rootstock size and fruit production as $r > 0$ and close to 1.

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

Next, the following R-code is run. Some output has been masked with an `x`:

```
summary(lm(fruit ~ diameter, data = root))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -125.28      14.56       x        x       x
## diameter       23.25       1.74       x        x       x
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 7.761 on 18 degrees of freedom
```

Furthermore, the symbols related to significance of parameters (`*` or `.`) have been removed.

Question IV.2 (9)

The average rootstock diameter was $\bar{x} = 8.31$ mm. Compute the average fruit production, $\bar{y}$.

1 ☐ 29.0 mg

2 ☐ 31.6 mg

3* ☐ 67.9 mg

4 ☐ 193.2 mg

5 ☐ 318.5 mg

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

It follows from (5-10) that $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, so the average fruit production may be computed as

$$\bar{y} = -125.28 + 23.25 \cdot 8.31 = 67.9\,\text{mg}$$

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

Question IV.3 (10)

Calculate the 95% confidence interval for β_1. You may also use that $S_{\text{xx}} = 19.90$.

1* ☐ [19.59, 26.91]

2 ☐ [19.68, 26.82]

3 ☐ [19.84, 26.66]

4 ☐ [20.23, 26.27]

5 ☐ [22.43, 24.07]

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

The confidence interval is found using Method 5.15. $\hat{\beta}_1 = 23.25$. $t_{0.975} = 2.10$; here is used 18 degrees of freedom. $\hat{\sigma}_{\beta_1} = 1.74$, which can be read from the R output, or found by formula 5-44, $\hat{\sigma}_{\beta_1} = \hat{\sigma}/\sqrt{(SSD_x)} = 7.761/\sqrt{19.90}$.

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

Question IV.4 (11)

In this experiment, rootstocks were measured by their diameter. If we assume the rootstocks to be circular, area and diameter are related by the following formula:

$$\text{area} = \tfrac{\pi}{4} \cdot \text{diameter}^2$$

where $\pi \approx 3.1416$. A rootstock was measured to have a diameter of 9.60 mm. The standard deviation on the measurement is $\sigma = 0.05$ mm. Using the error propagation rule, approximate the standard deviation on the measurement of the area of this rootstock:

1 ☐ 0.05 mm^2

2* ☐ 0.754 mm^2

3 ☐ 0.960 mm^2

4 ☐ 1.61 mm^2

5 ☐ 3.72 mm^2

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

The error propagation rule says that $\sigma_{\text{area}} \approx \sigma_{\text{diam}} \cdot f'(x_0)$. $f'(x) = 2 \cdot \frac{\pi}{4} \cdot x = \frac{\pi}{2} \cdot x$. Inserting the right numbers, we get:

$$\sigma_{\text{area}} \approx 0.05 \cdot 2 \cdot \pi/4 \cdot 9.60 = 0.754$$

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

The researchers subsequently expand their experiment to study the effect of grazing herbivores, which is measured on a scale from 0 to 1. Data are stored in `root2`, and the following R-code is run:

```r
summary(lm(fruit ~ diameter + grazing, data = root2))
```

```
##
## Call:
## lm(formula = fruit ~ diameter + grazing, data = root2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1964  -2.8268   0.3196   3.9146  17.3264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -91.774      7.118  -12.89 2.95e-15 ***
## diameter      23.568      1.149   20.51  < 2e-16 ***
## grazing      -36.114      3.358  -10.75 6.10e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.749 on 37 degrees of freedom
## Multiple R-squared:  0.9291,Adjusted R-squared:  0.9252
## F-statistic: 242.3 on 2 and 37 DF,  p-value: < 2.2e-16
```

Question IV.5 (12)

Look at the R output above. Which of the following statements is correct, given a significance level of $\alpha = 1\%$?

1 ☐ Rootstock size appears to have a significant effect on fruit production, while grazing does not.

2 ☐ Rootstock size is not significant, because the p-value is greater than 0.01.

3* ☐ Both rootstock size and grazing are significant, because the p-values are less than 0.01.

4 ☐ Neither rootstock size nor grazing appear to be significant, because the p-values are less than 0.05.

5 ☐ None of the variables involved are normally distributed, because all p-values are lower than 0.01.

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

Both rootstock size and grazing are highly significant.

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

Continue on page 13

Exercise V

This exercise contains two independent questions.

Question V.1 (13)

In the following R code two samples of different size are simulated:

```
x1 <- rnorm(10, mean=4, sd=5)
x2 <- rnorm(20, mean=4, sd=5)
t.test(x1, x2)
```

If this code is run, what is the probability that the obtained p-value in the result of the `t.test` function is below some level $\alpha \in [0, 1]$?

1 ☐ $\alpha \cdot \frac{2+10}{10+20}$

2 ☐ $\alpha \cdot \sqrt{\frac{1}{10} + \frac{1}{20}}$

3 ☐ $\alpha \cdot (\frac{1}{10} + \frac{1}{20})$

4 ☐ $1 - \alpha$

5* ☐ α

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

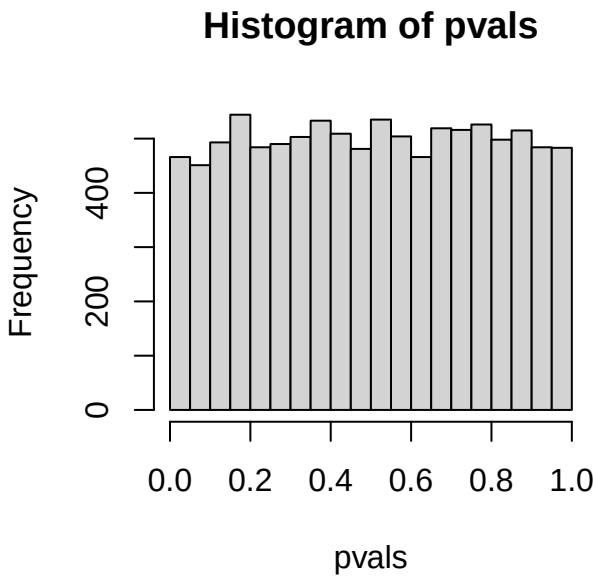
Since there is no difference in means between two populations from which the samples are simulated, then the p-value of the t-test of no difference in mean, hence the null hypothesis is true, will actually be uniform distributed.

So if α is the significance level, we know that obtaining a p-value below α will make us wrongly reject the null hypothesis, i.e. make a Type I error.

We can check it by

```
k <- 10000
pvals <- numeric(k)
for(i in 1:k){
  x1 <- rnorm(10, mean=4, sd=2)
  x2 <- rnorm(20, mean=4, sd=10)
  pvals[i] <- t.test(x1, x2)$p.value
}
hist(pvals)
```

```r
# E.g. alpha = 5%
sum(pvals < 0.05) / k

## [1] 0.0466
```



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

Question V.2 (14)

A plan has been made for a new experiment for testing difference in mean between two populations. The power of the test must be calculated at significance level 5%. The minimum difference in mean to be detected is set to 1, and the sample size is 30 in each sample.

The standard deviation of the populations are assumed to be equal. To get a value for the design, the pooled estimate from a previous similar experiment is used. The previous sample from the first population had sample standard devation $s_1 = 1.8$ and sample size $n_1 = 20$, and the previous sample from the second population had sample standard devation $s_2 = 1.4$ and sample size $n_2 = 30$.

With this planned experiment what is the power of the test?

1 ☐ 0.921

2 ☐ 0.339

3 ☐ 0.998

4 ☐ 0.227

5* ☐ 0.679

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

We first need to calculate the pooled estimate of the standard deviation using Method 3.52

```
var1 <- 1.8^2
n1 <- 20
var2 <- 1.4^2
n2 <- 30
varp <- ((n1-1)*var1+(n2-1)*var2) / (n1+n2-2)
```

and this can be used for calculating the power of the test:

```
power.t.test(n=30, delta=1, sd=sqrt(varp), sig.level=0.05)
```

```
##
##      Two-sample t test power calculation
##
##               n = 30
##           delta = 1
##              sd = 1.570563
##       sig.level = 0.05
##           power = 0.6790425
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

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

Continue on page 17

Exercise VI

Researchers want to compare rainfall in April between two regions "A" and "B" of a particular country. They have 20 observations from each region, all which can be assumed to be independent. Data are measured in mm.

Assume data for region A is stored in `rainA` and data for region B is stored in `rainB`.

Question VI.1 (15)

The folowing code was run:

```
sum(rainA)

## [1] 610.2105

quantile(rainA, probs = c(0.025, 0.975))

##      2.5%     97.5%
##  8.278595 66.521274
```

Which of the following statements can be concluded about the sample data from region A:

1* ☐ The mean of the sample is 30.5.

2 ☐ The median of the sample is 37.4.

3 ☐ The variance of the sample is 14.9.

4 ☐ The standard deviation of the sample is 14.9.

5 ☐ None of the above.

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

The sum of the sample is 610.2. Therefore the mean is $\frac{610.2}{20} = 30.5$.

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

Continue on page 18

Question VI.2 (16)

The researchers decide to compare the medians of the two samples.

Which of the following code snippets correctly computes a 95% confidence interval for difference in medians using non-parametric bootstrap?

1* ☐
```
sim_median_diff <- replicate(1000,
                             median(sample(rainA, 20, replace = TRUE)) -
                             median(sample(rainB, 20, replace = TRUE)))
quantile(sim_median_diff, c(0.025, 0.975))
```

2 ☐
```
sim_median_diff <- replicate(1000,
                   median(sample(rainA - rainB, 20, replace = TRUE)))
quantile(sim_median_diff, c(0.025, 0.975))
```

3 ☐
```
t.test(rainA, rainB, paired = FALSE, conf.level = 0.95)$conf.int
```

4 ☐
```
t.test(rainA, rainB, paired = TRUE, conf.level = 0.95)$conf.int
```

5 ☐
```
t.test(rainA, rainB, paired = TRUE, conf.level = 0.975)$conf.int
```

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

Snippets 3-5 use a t-test and hence do not use non-parametric bootstrap. The setup in this experiment is two samples, unpaired, which the first snippet does (the second snippet assumes paired two-sample).

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

Question VI.3 (17)

As a result of the previous question, the researchers got the confidence interval [-16.3, 4.09].

Which of the following statements can be concluded?

1 ☐ The median rainfall in region A is significantly lower than the median rainfall in region B on a 5% significance level.

2 ☐ The median rainfall in region B is significantly lower than the median rainfall in region A on a 5% significance level.

3* ☐ There is not a significant difference between the median rainfall in regions A and B on a 5% significance level.

4 ☐ There is a linear relationship between rainfall in region A and region B

5 ☐ None of the above.

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

Since 0 is contained in the 95% confidence interval, there is not a significant difference between the medians of the two samples.

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

Continue on page 20

Exercise VII

Developers of forest management techniques want to know how to increase biodiversity. They carried out experiments where different techniques of management were applied to randomly selected locations in a forest. At each location one of four different foresting techniques was applied.

After five years the biodiversity was calculated at each location with the Shannon biodiversity index, which measures the richness of species.

The observed values for the four techniques named A to D are:

A	B	C	D
0.9	1.8	1.5	0.9
1.5	1.8	1.7	1.6
1.4	2.1	1.8	1.2
1.1	1.7	1.0	1.6
2.0	2.3	1.9	1.5

An ANOVA analysis was carried out to determine if there is any significant difference in mean between any of the techniques. It can be assumed that the necessary assumptions for using ANOVA are met. The ANOVA result is:

```
anova(lm(y ~ technique))

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## technique  3 1.0855 0.36183    X       X      X
## Residuals 16 1.8400 0.11500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that some of the values have been replaced with an `X`.

Question VII.1 (18)

What is the total variance (SST)?

1 ☐ 0.36183

2 ☐ 1.0855

3 ☐ 1.1783

4 ☐ 1.8400

5* ☐ 2.9255

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

We add together the the sum of squared errors (SSE) and the treatment sum of squares ($SS(Tr)$)

$$SST = 1.0855 + 1.8400 = 2.9255$$

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

Question VII.2 (19)

Calculate the F test statistic for the ANOVA. What is the conclusion of the test for difference in means on a significance level of $\alpha = 5\%$ (both conclusion and argument must be correct)?

1* ☐ The null hypothesis is not rejected since $F = 3.146 < 3.239$.

2 ☐ The null hypothesis is rejected since $F = 3.146 < 4.077$.

3 ☐ The null hypothesis is not rejected since $F = 3.239 < 4.077$.

4 ☐ The null hypothesis is not rejected since $F = 0.0542 < 4.077$.

5 ☐ The null hypothesis is rejected since $F = 0.0542 > 0.05$.

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

The test statistic follows an F-distribution under the null hypothesis and using Method 8.6 we get the formula to calculate it with the available values

$$F = \frac{1.0855/3}{1.8400/16} = \frac{0.36183}{0.11500} = 3.1464$$

The critical value in the F-distribution is:

```
qf(0.95, df1=3, df2=16)

## [1] 3.238872
```

which we find in the answers and since the observed value is lower, then the null hypothesis is not rejected.

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

Question VII.3 (20)

It was pre-planned to calculate the 95% confidence interval for the difference in mean between two specific techniques. Which of the following R codes calculates the width of this interval?

1 ☐ `2 * qt(0.95, df=16) * sqrt(0.115 * 1/25)`

2 ☐ `qt(0.95, df=16) * sqrt(1.84 * 1/5)`

3* ☐ `2 * qt(0.975, df=16) * sqrt(0.115 * 2/5)`

4 ☐ `2 * qt(0.975, df=20) * sqrt(0.115 * 2/5)`

5 ☐ `qt(0.975, df=20) * sqrt(1.84 * 1/5)`

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

We find the formula for the confidence interval in Method 8.9. We find the values and get

```
n <- 20
k <- 4
MSE <- 0.115
ni <- 5
nj <- 5
2 * qt(0.975, df=n-k) * sqrt(MSE * (1/ni + 1/nj))

## [1] 0.9093381
```

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

Continue on page 23

Exercise VIII

Question VIII.1 (21)

As part of their bachelor project, two students have performed experiments and collected data, and now wish to examine the relationship between two variables "X" and "Y".

However, they cannot agree on how to correctly check the assumptions for linear regression. Only one of the statements below is correct. Which one?

1 ☐ A QQ-plot of the "Y" values would reveal if the normality assumption is met.

2 ☐ A QQ-plot of the "X" values would reveal if the normality assumption is met.

3 ☐ A boxplot of the "X" and "Y" values would reveal if the variance homogenity assumption is met.

4* ☐ A residual plot with fitted values on one axis and residuals on the other axis would reveal if the linearity assumption is met.

5 ☐ Parametric bootstrap would reveal if the assumptions of linear regression is met.

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

If the linearity assumption is met, one should not be able to see a pattern of residuals vs. fitted values. In linear regression, QQ plots are used for the residuals, not X or Y.

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

Continue on page 24

Exercise IX

Data from a randomized block Design experiment were read into R:

```
y <- c(3.5, 3.0, 5.4, 7.2,
       7.7, 9.0, 7.0, 6.0,
       0.4, 1.1, 1.0, 1.8)

treatm <- as.factor(c(1, 1, 1, 1,
                      2, 2, 2, 2,
                      3, 3, 3, 3))

block <- as.factor(c(1, 2, 3, 4,
                     1, 2, 3, 4,
                     1, 2, 3, 4))
```

A model for the data from such an experiment is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \text{ where } \varepsilon_{ij} \sim N(0, \sigma^2) \text{ and independent}$$

where $i = 1, \ldots, k$ index the treatment ($k = 3$) and $j = 1, \ldots, l$ index the block ($l = 4$).

Following the book, we estimate μ by the overall mean, $\hat{\mu} = \bar{\bar{y}}$.

Question IX.1 (22)

Given the assumptions behind the model is fulfilled what is the estimate of the effect from Block 1?

1 ☐ $\hat{\beta}_1 = 4.425$

2 ☐ $\hat{\beta}_1 = 4.775$

3 ☐ $\hat{\beta}_1 = 3.867$

4* ☐ $\hat{\beta}_1 = -0.558$

5 ☐ $\hat{\beta}_1 = 2.988$

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

We use Equation (8-37) as done in the example following right after the equation in the book. So:

```r
(mu <- mean(y))

## [1] 4.425

(alpha <- tapply(y, treatm, mean) - mu)

##     1     2     3
##  0.35  3.00 -3.35

(beta <- tapply(y, block, mean) - mu)

##           1           2           3           4
## -0.55833333 -0.05833333  0.04166667  0.57500000

mean(c(3.5, 7.7, 0.4)) - mu

## [1] -0.5583333

mean(y[block==1]) - mu

## [1] -0.5583333
```

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

Question IX.2 (23)

What is the estimate of σ?

1 ☐ $\hat{\sigma} = 0.65$

2* ☐ $\hat{\sigma} = 1.57$

3 ☐ $\hat{\sigma} = 14.9$

4 ☐ $\hat{\sigma} = 8.93$

5 ☐ $\hat{\sigma} = 3.85$

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

The easiest method is probably to input the ANOVA model and find the estimate in the result:

SSTr MSTr

SSE MSE $MSE = \sigma^2$

```
anova(lm(y ~ treatm + block))

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## treatm     2 81.380  40.690 16.4293 0.003681 **
## block      3  1.943   0.648  0.2614 0.850900
## Residuals  6 14.860   2.477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

and remember to take the square root to get the standard deviation:

```
sqrt(2.477)

## [1] 1.573849
```

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

Continue on page 27

Exercise X

The table below shows the number of people killed in traffic in Denmark in some categories (i.e. not all traffic deaths are included), for the years 2016-2019.

Year	2016	2017	2018	2019	Total
Ordinary car	96	99	62	87	344
Motorcycle	26	11	21	27	85
Bike	31	27	28	31	117
Pedestrian	36	20	30	30	116
Total	189	157	141	175	662

For the remainder of the exercise, we define a "soft traffic user" as being either a bicyclist or a pedestrian.

Question X.1 (24)

What is the usual 95% confidence interval for the proportion of "soft traffic users" killed in traffic based on the data given above (i.e. using the total over the 4 years)?

1* ☐ $[0.316, 0.388]$

2 ☐ $[0.496, 0.590]$

3 ☐ $[0.321, 0.382]$

4 ☐ $[0.326, 0.378]$

5 ☐ $[0.504, 0.583]$

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

This answer is given by the formula

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

which can be calculated in R by

```
n <- 662
p <- (117+116)/n
p + c(-1, 1) * sqrt(p * (1 - p) / n) * qnorm(0.975)

## [1] 0.3155833 0.3883442
```

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

Question X.2 (25)

As part of the analysis it is desired to test if there is a statistically significant difference in the proportions of soft traffic users being killed in the years 2016 and 2019 ($H_0 : p_{2016} - p_{2019} = 0$). What is the conclusion using the usual test and significance level $\alpha = 5\%$ (both conclusion and argument must be correct)?

1 ☐ The test statistic is $Z = 0.118$. There is not a significant difference, since $Z < 0.95$

2 ☐ The test statistic is $Z = 0.237$. There is a significant difference, since $Z > 0.05$.

3 ☐ The test statistic is $Z = 0.237$. There is not a significant difference, since $Z < 1.96$

4 ☐ The test statistic is $Z = 0.237$. There is not a significant difference, since $Z > 0.05$.

5* ☐ The test statistic is $Z = 0.118$. There is not a significant difference, since $Z < 1.96$.

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

The test statistics is calculated by

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{2}$$

where $\hat{p}_1$ and $\hat{p}_2$ is the observed proportion in the two groups, and $\hat{p}$ is the proportion under the null-hypothesis.

```
 ## 2: CI p soft 16 - p soft 19
p1 <- (31+36)/189
p2 <- c(31+30)/175
p <- (31+36+31+30)/(189+175)
(z <- (p1-p2)/sqrt(p*(1-p)*(1/189+1/175)))

## [1] 0.118303
```

The observed statistics should be compared with the critical value $z_{1-\alpha/2} = 1.96$, which is answer no. 5.

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

Question X.3 (26)

What is the usual 95% confidence interval for the difference in the proportion of motor cyclists being killed in 2016 and 2017 ($p_{2016} - p_{2017}$)?

1 ☐ $[0.070, 0.138]$

2 ☐ $[0.013, 0.122]$

3 ☐ $[0.022, 0.113]$

4* ☐ $[0.004, 0.131]$

5 ☐ $[0.044, 0.091]$

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

The CI is calculated by

$$\hat{p}_1 - \hat{p}_2 \pm \sqrt{\hat{p}_1(1-\hat{p}_1/n_1 + \sqrt{\hat{p}_2(1-\hat{p}_2}/n_2} \tag{3}$$

The result can be found in R by

```
p1 <- 26/189
p2 <- 11/157
p1-p2 + c(-1,1) * sqrt(p1*(1-p1)/189+p2*(1-p2)/157)*qnorm(0.975)

## [1] 0.004212527 0.130792360
```

which is answer no 4.

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

As an aid for the next questions, the following result from R is given (some number are replaced by letters), where `dat` is the appropriate part the table above (i.e. excluding the totals)

```
> chisq.test(dat)

 chisq.test(dat)

Pearson's Chi-squared test

data:  dat
X-squared = 15.356, df = A, p-value = B
```

Question X.4 (27)

On significance level $\alpha = 5\%$, what is the conclusion about the entire distribution over the 4 years (both conclusion and argument should be correct)?

1 ☐ There is a significant change over the years, since $15.36 > 12.59$, where 12.59 is the critical value from the χ^2-test.

2* ☐ It cannot be rejected that the distribution is unchanged, since the p-value is $0.08 > 0.05$.

3 ☐ It cannot be rejected that the distribution is unchanged, since the p-value is $0.50 > 0.05$

4 ☐ There is a significant change over the years, since $15.36 > 9$, where 9 is the critical value from the χ^2-test.

5 ☐ It cannot be rejected that the distribution is unchanged, since $15.35 < 26.27$, where 26.27 is the critical value from the χ^2-test.

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

The numbers A, B and the critical value is

```
A <- 9
(B <- 1-pchisq(15.356,df=A))

## [1] 0.08161004

(cv <- qchisq(0.95,df=A))

## [1] 16.91898
```

Hence 1) use a wrong critical value, 2) use the correct p-value and the right significance level and the conclusion is also correct, 3) use the wrong p-value, 4) and 5) use the wrong critical value. Hence the only correct answer is answer no. 2.

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

Question X.5 (28)

What is the contribution to the test statistic from the cell "ordinary car" in 2016?

1* ☐ 0.050

2 ☐ 0.508

3 ☐ 0.279

4 ☐ 0.520

5 ☐ 0.285

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

The test statistic is calculated by

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{4}$$

with $e_{ij} = \frac{n_i n_j}{n}$, hence the contribution from cell (1,1) is

```
e <- 344*189/662
(e-96)^2/e

## [1] 0.04979708
```

i.e. answer no 1.

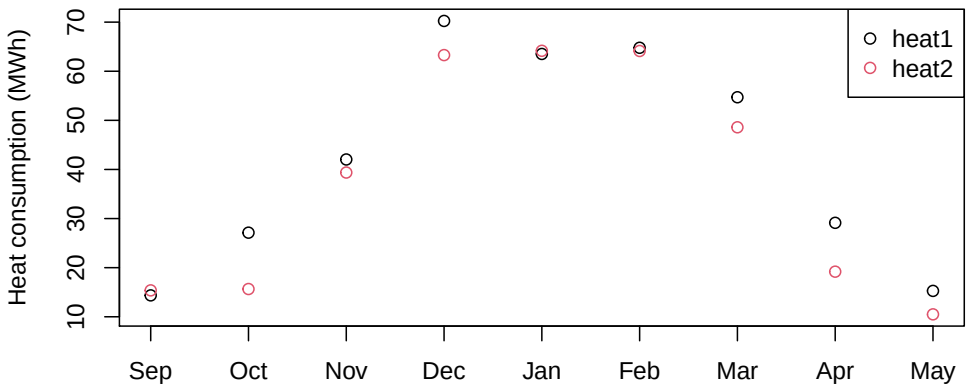
---------- FACIT-END ----------

Continue on page 32

Exercise XI

Current policy dictates that heat consumption in buildings must be decreased in the forthcoming years. An experiment was carried out in two identical apartment buildings. The occupants in one of the two buildings were given advices on how to save energy during a heating season (Sep-May).

The weekly heat consumption covering the period has been read into two vectors in R, one for each building: `heat1` and `heat2`. They are plotted below.



Question XI.1 (29)

A test for difference in mean heat consumption between the buildings is carried out. Which of the following R codes calculates the correct result of such a test?

1 ☐ `summary(lm(heat1 ~ heat2))`

2 ☐ `prop.test(heat1 > heat2, length(heat1))`

3 ☐ `t.test(heat1, heat2)`

4* ☐ `t.test(heat1-heat2)`

5 ☐ `1 - pt(mean(heat1)-mean(heat2), df=length(heat1))`

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

The samples are paired in time, so we have to test the difference.

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

Question XI.2 (30)

The p-value of the test was 1.6%. According to the book, how should one communicate this result?

1 ☐ We have little or no evidence that there is a significant difference in mean heat consumption.

2 ☐ We have little or no evidence that there is no significant difference in mean heat consumption.

3* ☐ We have some evidence that there is a significant difference in mean heat consumption.

4 ☐ We have some evidence that there is no significant difference in mean heat consumption.

5 ☐ We have strong evidence that there is no significant difference in mean heat consumption.

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

According to Table 3.1 we should use the wording "some evidence" against H_0, which is that there no difference. So we have some evidence that there is a difference.

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

Continue on page 34

SÆTTET ER SLUT. God jul!