Technical University of Denmark *Page 1 of 36 pages.*

Written examination: 24. June 2022

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

Duration: 4 hours

Aids and facilities allowed: All

The questions were answered by

_____ _____ _____

(student number) (signature) (table number)

This exam consists of 30 questions of the "multiple choice" type, which are divided between 8 exercises. To answer the questions, you need to fill in the "multiple choice" form (6 separate pages) on Digital Exam with the numbers of the answers that you believe to be correct.

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

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

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

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

Exercise	VI.2	VI.3	VI.4	VI.5	VI.6	VII.1	VIII.1	VIII.2	VIII.3	VIII.4
Question	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Answer	5	1	3	4	4	1	5	1	2	5

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 birth weight of 50 newborn girls has been recorded in an unknown country, and the sample mean and standard deviation were found to be $\bar{x}_{\text{p}} = 3505.7$ g and $s_{\text{p}} = 467.9$ g.

Question I.1 (1)

What is the 95% confidence interval for the mean birth weight of girls, μ_{p}?

1 ☐ $[3328.3, 3683.0]$

2* ☐ $[3372.7, 3638.7]$

3 ☐ $[3371.4, 3640.0]$

4 ☐ $[3328.4, 3683.0]$

5 ☐ $[3499.6, 3511.8]$

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

Following equation (3-10):

$3505.7 - t_{0.975} \cdot 467.9/\sqrt{50} = 3505.7 - 2.009575 \cdot 467.9/\sqrt{50} = 3372.724$

$3505.7 + t_{0.975} \cdot 467.9/\sqrt{50} = 3505.7 + 2.009575 \cdot 467.9/\sqrt{50} = 3638.676$

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

Question I.2 (2)

The mean birth weight of Danish girls is known to be 3449 g. We want to test if the birth weight from the unknown country differs significantly from the birth weight of Danish girls, so we test the hypothesis $H_0 : \mu_{\text{p}} = 3449$ g using the observed data.

The corresponding test statistic is computed as $t_{\text{obs}} = 0.857$. Which of the following statements is correct, when we use a significance level of $\alpha = 0.05$ (both p-value and conclusion must be correct)?

1 ☐ p-value = 0.198, and the hyptothesis cannot be rejected.

2 ☐ p-value = 0.198, and the hyptothesis is rejected.

3* ☐ p-value = 0.396, and the hyptothesis cannot be rejected.

4 ☐ p-value = 1.604, and the hyptothesis is rejected.

5 ☐ p-value = 0.396, and the hyptothesis is rejected.

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

The p-value is computed by:

```
2*(1-pt(0.857,df = 49))
```

which gives p-value = 0.396. Since the p-value is greater than 0.05, the hypothesis cannot be rejected.

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

Continue on page 4

The birth weight of 50 newborn boys has also been recorded (in the same country). The sample mean and standard deviation were found to be $\bar{x}_{\mathrm{d}} = 3619.4\mathrm{g}$ and $s_{\mathrm{d}} = 409.0\mathrm{g}$. We want to test the hypothesis that girls and boys have the same mean birth weight against the alternative hypothesis that the means are different. We use a significance level of $\alpha = 0.05$ in the remainder of the questions.

Question I.3 (3)

Under the null-hypothesis of no difference between the mean birth weight of girls and boys, the (Welch) two-sample statistic, T, follows a t-distribution with ν degrees of freedom. What is ν in our case equal to?

1 ☐ 100

2 ☐ 98.24

3 ☐ 49

4 ☐ 98

5* ☐ 96.28

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

The result is found by using equation (3-50).

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

Continue on page 5

Question I.4 (4)

Assume that the number of degrees of freedom, ν, is stored in R as `v`. Which command results in the correct critical value in the t-distribution, mentioned in the previous question, to be used for the hypothesis test of equal means?

1* ☐ `qt(0.975,v)`

2 ☐ `1-pt(0.975,v)`

3 ☐ `pt(0.95,v)`

4 ☐ `1-qt(0.95,v)`

5 ☐ `qt(0.95,v)`

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

The critical value is always a quantile, and therefore the `qt()` function is used. Since $\alpha = 0.05$ and we make a two-sided test, we are looking for the 97.5% quantile.

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

Continue on page 6

Question I.5 (5)

The sampled birth weights of girls and boys are stored in `xp` and `xd`, respectively. Which of the commands below would generate the correct confidence interval for the difference in means?

1 ☐
```r
t.test(xp, xd, paired = TRUE)
```

2 ☐
```r
t.test(xp, xd, paired = TRUE, conf.level = 0.90)
```

3 ☐
```r
t.test(xp, xd, conf.level = 0.90)
```

4 ☐
```r
t.test(xp, xd, paired = TRUE, conf.level = 0.95)
```

5* ☐
```r
t.test(xp, xd)
```

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

Since $\alpha = 0.05$, we want the `conf.level` to be 0.95 and that is the default. Our observations er not paired and therefore `paired` should be `FALSE`, which is also the default.

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

Continue on page 7

Exercise II

A Danish company wants to investigate whether the employees' professional training on a virtual reality (VR) platform affect their task quality score. 200 employees participated. The following count data provides an overview of task quality (poor, medium and good) versus VR training engagement level (below average, average, and above average).

Task Quality Score \VR engagement	Below Average	Average	Above Average	Row Total
Poor	11	27	15	53
Medium	14	40	30	84
Good	5	23	35	63
Column Total	30	90	80	200

The null-hypothesis of independence between task quality score and VR training engagement score is to be tested by χ^2-test.

Employees with "medium" and "good" task quality score are considered as "Efficient Employees".

Question II.1 (6)

What is the expected number of individuals with below average VR training engagement score and poor task quality score under H_0 (i.e assuming H_0 is true)?

1* ☐ 7.95

2 ☐ 21.83

3 ☐ 25.2

4 ☐ 19.43

5 ☐ 9.45

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

The expected number under the null hypothesis for each cell is found as

$$\text{"column total"} \cdot \frac{\text{"row total"}}{\text{"total"}},$$

for table cell (1,1), which is the number of individuals with below average VR training engagement score and poor task quality score. So, the answer is

$$e_{11} = 30 \cdot \frac{53}{200} = 7.95.$$

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

Question II.2 (7)

What is the 95% confidence interval for the proportion of "Efficient Employees" based on the data given above?

1* ☐ $[0.674, 0.796]$

2 ☐ $[0.621, 0.749]$

3 ☐ $[0.532, 0.668]$

4 ☐ $[0.426, 0.578]$

5 ☐ $[0.706, 0.824]$

---------- 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 <- 200
p <- (84+63)/n
p + c(-1, 1) * sqrt(p * (1 - p) / n) * qnorm(0.975)

## [1] 0.6738355 0.7961645
```

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

Question II.3 (8)

What is the 95% confidence interval for the difference in the proportion of good task quality scorers with VR training engagement score "above average" and "average" ($p_{\text{AbAvg, good}} - p_{\text{Avg, good}}$)?

1 ☐ $[0.019, 0.361]$

2 ☐ $[0.043, 0.212]$

3 ☐ $[0.011, 0.313]$

4* ☐ $[0.041, 0.323]$

5 ☐ $[0.044, 0.091]$

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

The CI is calculated by

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

The result can be found in R by

```
p1 <- 35/80
p2 <- 23/90
p1-p2 + c(-1,1) * sqrt(p1*(1-p1)/80+p2*(1-p2)/90)*qnorm(0.975)

## [1] 0.04074504 0.32314385
```

which is answer no 4.

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

As a help for the next question the following R-code, where `training` is the table of counts, has been executed (some numbers have been replaced by letters)

```
chisq.test(training, correct = FALSE)
##
## Pearson's Chi-squared test
##
##  data:  training
## X-squared = 10.985, df = A, p-value = B
```

Question II.4 (9)

Considering the χ^2-test statistic, what is the p-value and the correct conclusion using significance level $\alpha = 0.05$ (all parts of the answer must be correct)?

1* ☐ There is a significant dependence between VR training engagement and task quality, as p-value $= 0.027 < 0.05 = \alpha$

2 ☐ There is no evidence of significant dependence between VR training engagement and task quality, as p-value $= 0.027 < 0.05 = \alpha$

3 ☐ The Null-hypothesis cannot be rejected since p-value $= 0.50 > \alpha = 0.05$

4 ☐ There is a significant dependence between VR training engagement and task quality, as p-value $= 0.037 < 0.05 = \alpha$

5 ☐ There is no evidence of significant dependence between VR training engagement and task quality, as p-value $= 0.037 < 0.05 = \alpha$

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

The test is conducted noting that the degrees of fredom (A) is 4 and hence the p-values can be calculated by

```
(B <- 1-pchisq(10.985,df=4))

## [1] 0.02673311
```

or this can be done be setting up the table and using chisq.test

```
training <- matrix(c(11,27,15,14,40,30,5,23,35), ncol =3, byrow =TRUE)
chi <- chisq.test(training, correct = FALSE)
chi

##
##  Pearson's Chi-squared test
##
## data:  training
## X-squared = 10.985, df = 4, p-value = 0.02673
```

The correct option is 1) as it shows the correct p-value with correct rejection rule. The rejection rule is: Reject H_0 if p-value $< \alpha$. The hypothesis is that there is no dependence. By rejecting the hypothesis, we can state that "There is significant dependence between VR training engagement score and task quality score.

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

Continue on page 11

Exercise III

In an experiment regarding poisoning of rats, survival time (days) for 24 rats were measured. Each rat received poison and was afterwards treated with one of four treatments, A, B, C, D. Let Y denote the logarithm of the survival time (`logt`) which is used for the analysis.

A one-way ANOVA model was fitted to the data:

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

```
logt <- c(-1.02, -1.24, -0.92, -1.47, -0.08, -0.49, -0.71, 0.22, -0.82, -1.05,
          -1.17, -0.92, -0.58,  0.02, -0.34, -0.97, -1.51, -1.56, -1.2, -0.99,
          -1.47, -1.39, -1.2,  -1.02)
treatments <-
  as.factor(c("A", "A", "A", "A", "B", "B", "B", "B", "C", "C", "C", "C",
              "D", "D", "D", "D", "A", "A", "B", "B", "C", "C", "D", "D"))
```

We are additionally informed that $SS(Tr) = 2.286$ and $SSE = 3.241$, and the values of the group means:

```
tapply(logt, treatments, mean)

##          A          B          C          D
## -1.2866667 -0.5416667 -1.1366667 -0.6816667
```

Question III.1 (10)

What is the estimate of the effect of treatment B, $\hat{\alpha}_B$?

1 ☐ -0.542

2* ☐ 0.370

3 ☐ 0.542

4 ☐ 2.22

5 ☐ 2.33

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

$$\hat{\alpha}_B = \bar{y}_B - \bar{\bar{y}} = -0.542 - (-0.912) = 0.370$$

where $\bar{\bar{y}}$ is calculated by

```
logt <- c(-1.02, -1.24, -0.92, -1.47, -0.08, -0.49, -0.71, 0.22, -0.82, -1.05,
          -1.17, -0.92, -0.58,  0.02, -0.34, -0.97, -1.51, -1.56, -1.2, -0.99,
          -1.47, -1.39, -1.2,  -1.02)
mean(logt)

## [1] -0.9116667
```

or by

```
mean(c(-1.28667,-0.54167,-1.13667,-0.68167))

## [1] -0.91167
```

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

Continue on page 13

Question III.2 (11)

What is the value of the usual test statistic (F), for testing difference in treatments?

1 ☐ 0.0121

2 ☐ 0.7051

3* ☐ 4.702

4 ☐ 14.11

5 ☐ 16.93

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

Either load the data into R and read the value from the ANOVA table, or use the formula (8-19):

$$F = \frac{SS(Tr)/(k-1)}{SSE/(n-k)} = \frac{2.286/3}{3.241/20} = 4.702$$

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

The researchers are particularly interested in comparing treatments B and D, as their knowledge in chemistry predicts that these treatments should be roughly equally good.

Question III.3 (12)

What is the conclusion on a 5% significance level regarding the post hoc difference in means between treatments B and D ($\alpha_{\mathrm{D}} - \alpha_{\mathrm{B}}$) (both argument and conclusion must be correct)?

1 ☐ The 95% confidence interval for the difference in means is $[-0.902, 0.622]$. Hence the treatments are not significantly different.

2 ☐ The 95% confidence interval for the difference in means is $[-0.659, 0.338]$. Hence the treatments are not significantly different.

3 ☐ The 95% confidence interval for the difference in means is $[-0.659, 0.338]$. Hence the treatments are significantly different.

4* ☐ The 95% confidence interval for the difference in means is $[-0.625, 0.345]$. Hence the treatments are not significantly different.

5 ☐ The 95% confidence interval for the difference in means is $[-0.206, -0.020]$. Hence the treatments are significantly different.

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

Use method 8.9. No Bonferroni adjustment as we are doing one pre-specified comparison:

$\bar{y}_{\mathrm{D}} - \bar{y}_{\mathrm{B}} \pm t_{0.975}\sqrt{MSE \cdot (1/n_{\mathrm{D}} + 1/n_{\mathrm{B}})} = -0.682 - (-0.542) \pm 2.086 \cdot \sqrt{0.16207/3} = [-0.625, 0.345]$

Since 0 is in the confidence interval, the difference in means is not significant.

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

Continue on page 15

Exercise IV

In an office building, the duration of a room being available (i.e. empty) was measured during a period of approximately 8 months, there is one observation every time the room change from occupied to empty (this imply that there might be more than one observation pr. day). The duration of availability is measured as available hours during normal office hours, and the measurements were stored in the vector `time`. A summary of the measured duration of availability is given below.

```
summary(time)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2500  0.9375  2.5000  2.7017  3.7500 11.7500
```

Question IV.1 (13)

What is the Inter Quartile Range (IQR) for the presented data?

1 ☐ 0.20

2 ☐ 2.70

3 ☐ 2.50

4* ☐ 2.81

5 ☐ 11.5

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

IQR is the difference between the first and the third quartile, both are given directly in the summary from R and the IQR is

```
3.75-0.9375

## [1] 2.8125
```

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

Assume that a random variable X follows an exponential distribution with expected value equal to the observed average of the time of availability.

Question IV.2 (14)

What is the median of X?

1 ☐ 2.50

2 ☐ 0.169

3* ☐ 1.87

4 ☐ 2.70

5 ☐ 0.741

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

The median is the value $\tilde{x}$ such that

$$0.5 = P(X \leq \tilde{x}) = F(\tilde{x}) \tag{3}$$

and hence

$$\tilde{x} = F^{-1}(0.5) \tag{4}$$

this can be calculated i R by

```
qexp(0.5,1/2.702)

## [1] 1.872884
```

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

Continue on page 17

Question IV.3 (15)

Still assuming the exponential distribution with mean equal to the observed average time of availability, which of the following pieces R-code calculate a 95% parametric bootstrap confidence interval for the expected value of the time of availability (in all cases `n = length(time)` and `k`=10^4)?

1 ☐
```
m <- mean(time)
X <- matrix(rexp(n * k, m), ncol=k)
quantile(apply(X, 2, mean), prob = c(0.025, 0.975))
```

2* ☐
```
m <- mean(time)
X <- matrix(rexp(n * k, 1/m), ncol=k)
quantile(apply(X, 2, mean), prob = c(0.025, 0.975))
```

3 ☐
```
X <- replicate(k, sample(time, replace = TRUE))
quantile(apply(X, 2, mean), prob = c(0.025, 0.975))
```

4 ☐
```
X <- replicate(n, sample(time, replace = TRUE, size = k))
quantile(apply(X, 2, mean), prob = c(0.025, 0.975))
```

5 ☐
```
m <- mean(time)
X <- matrix(rnorm(n * k, m, sd(time)), n)
quantile(apply(X, 2, mean), prob = c(0.025, 0.975))
```

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

We should use the exponential distribution hence it will be either 1 or 2, answer 1 use the wrong rate (`m`), while 2 use the correct rate (`1/m`), and also correctly calculate the mean value and find the correct quantiles.

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

Continue on page 18

It is of interest to examine if the coefficient of variation is equal 1, for that purpose the following R-code have been evaluated (including the results):

```
k <- 1e4
n <- length(time)

X <- replicate(k, sample(time, replace = TRUE))
quantile(apply(X, 2, sd) / apply(X, 2, mean),
         prob = c(0.025, 0.05, 0.95, 0.975))

##      2.5%        5%       95%     97.5%
## 0.7586455 0.7727351 0.9228017 0.9382224

quantile(apply(X, 2, var) / apply(X, 2, mean),
         prob = c(0.025, 0.05, 0.95, 0.975))

##     2.5%       5%      95%    97.5%
## 1.508020 1.570758 2.324900 2.400079

X2 <- replicate(k, rexp(n, m))
quantile( apply(X2, 2, sd) / apply(X2, 2, mean) ,
         prob = c(0.025, 0.05, 0.95, 0.975))

##      2.5%        5%       95%     97.5%
## 0.8767490 0.8933601 1.1094034 1.1368452

quantile(apply(X2, 2, var) / apply(X2, 2, mean),
         prob = c(0.025, 0.05, 0.95, 0.975))

##      2.5%        5%       95%     97.5%
## 0.2740744 0.2873999 0.4669260 0.4916159
```

Question IV.4 (16)

Based on the R-code above what can we conclude using significance level $\alpha = 0.05$, and not using any distribution assumption (both conclusion and argument should be correct)?

1 ☐ It cannot be rejected that the coefficient of variation is equal to 1, since $1 > 0.94$

2 ☐ The coefficient of variation is less than 0.7 since a 95% confidence interval is $[0.27, 0.49]$

3 ☐ It cannot be rejected that the coefficient of variation is equal to 1, since $1 \notin [1.51, 2.4]$

4 ☐ It cannot be rejected that the coefficient of variation is equal to 1, since $1 \in [0.88, 1.14]$

5* ☐ The cefficient of variation is not equal to 1, since $1 \notin [0.76, 0.94]$

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

Not using any distribution assumption, imply that we should use non-parametric bootstrap. This imply that we should use either the first or the second stated result. In the first result the calculated numbers are

$$\frac{s}{\bar{x}} \tag{5}$$

where $\bar{x}$ and s is the observed mean and standard deviation from the simulation, which is the coefficient of variation and hence we should use the first result. This imply that the 95% confidence interval is $[0.76; 0.94]$ and hence the we can conclude (using the 5% significance level), that the coefficient of variation is not equal 1 (which is answer no 5).

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

Continue on page 20

A similar set of masurements was taken from another room, it is desired to compare the mean time of availability between the two rooms. The summary for the data from the second room is given below.

```
summary(time2)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.250   1.500   2.500   2.551   3.250  14.250
```

Assuming independece between the rooms, it was decided to test if there is a significant difference in the expected times of availability between the rooms using a test that does not use any distribution assumptions.

Question IV.5 (17)

Which of the following pieces of R-code can be used to test the hypothesis that there is no difference between the mean time of availability for the two rooms.

1 ☐
```
t.test(time, time2, paired = TRUE)
```

2 ☐
```
prop.test(sim1,sim2)
```

3 ☐
```
sim <- replicate(k, sample(time - time2, replace = TRUE))
quantile(apply(sim1, 2, mean), prob = c(0.025, 0.975))
```

4* ☐
```
sim1 <- replicate(k, sample(time, replace = TRUE))
sim2 <- replicate(k, sample(time2, replace = TRUE))
quantile(apply(sim1,2,mean) - apply(sim2,2,mean), prob = c(0.025,0.975))
```

5 ☐
```
t.test(time, time2)
```

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

Again we should not use any distribution assumption and hence we can exclude the answers based on the `t.test` (i.e. answer 1, and 5). `prop.test` is used to compare proportions and we have continuous data so we can exclude 2. As we assume independence between the rooms the correct answer 4 that calculate a two sample test of difference in mean between the rooms.

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

Continue on page 21

Exercise V

The production manager of a paper factory has analysed paper produced by the machinery on different days. He wishes to test if the quality varies according to machinery and day of production.

The analysis resulted in the following two-way ANOVA table (some values have been omitted):

```
## Analysis of Variance Table
##
## Response: quality
##           Df  Sum Sq  Mean Sq F value  Pr(>F)
## day        6 0.24902 0.041504  2.2731    X
## machine    3 0.21025 0.070085  3.8384    X
## Residuals 18 0.32866 0.018259
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Question V.1 (18)

How many observations were included in the study?

1 ☐ 6

2 ☐ 9

3 ☐ 18

4 ☐ 27

5* ☐ 28

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

The degrees of freedom are number of categories - 1. Hence we have 7 days and 4 machines. One observation per combination: total number of observations in the study is 28.

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

Continue on page 22

Question V.2 (19)

What is the conclusion about effects of day of production on a 5% significance level (both argument and conclusion must be correct):

1 ☐ There is a significant effect since the p-value is 0.028

2* ☐ There is not a significant effect since the p-value is 0.083.

3 ☐ There is not a significant effect since the p-value is 0.157

4 ☐ There is not a significant effect since $SSE = 0.329$ is within ± 1.96, where 1.96 is the 97.5% quantile in a standard normal distribution, $N(0, 1)$.

5 ☐ We do not have sufficient information to conclude whether day of production is significant or not.

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

We should use the provided F statistic for `day`. The distribution under the null hypothesis is $F(6, 18)$, so the answer is given by:

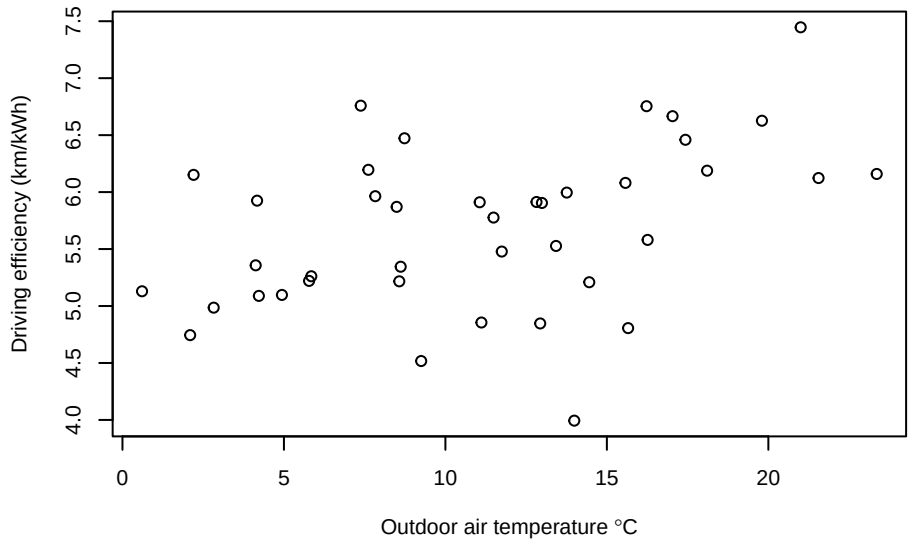
```
1 - pf(2.2731, df1 = 6, df2 = 18)

## [1] 0.08268843
```

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

Continue on page 23

Exercise VI

The owner of an electrical car wanted to find out what effect the ambient temperature has on the driving range. So she collected the driving efficiency (trip length per unit of energy), as well as the outdoor air temperature, on every trip she made during a period. The data can be seen in the scatter plot below:



A simple linear regression model with the driving efficiency as model output, and outdoor air temperature as model input, was fitted. The results were:

```
##
## Call:
## lm(formula = Efficiency ~ Toutdoor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.84694 -0.27181  0.01402  0.43993  1.26562
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.10493    0.22551  22.638   <2e-16 ***
## Toutdoor     0.05259    0.01799   2.924   0.0058 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6573 on 38 degrees of freedom
## Multiple R-squared:  0.1836,Adjusted R-squared:  0.1622
## F-statistic: 8.548 on 1 and 38 DF,  p-value: 0.0058
```

Question VI.1 (20)

Which of the following statements is correct (both the conclusion and the argument must be correct)?

1* ☐ At significance level of $\alpha = 0.01$ a significant correlation between the driving efficiency and the outdoor temperature could be detected, since $0.0058 < 0.01$.

2 ☐ At significance level of $\alpha = 0.01$ a significant correlation between the driving efficiency and the outdoor temperature could not be detected, since $0.05259 > 0.01$.

3 ☐ At significance level of $\alpha = 0.05$ a significant correlation between the driving efficiency and the outdoor temperature could be detected, since $0.01799 < 0.05$.

4 ☐ At significance level of $\alpha = 0.05$ a significant correlation between the driving efficiency and the outdoor temperature could not be detected, since $0.01799 < 0.05$.

5 ☐ At significance level of $\alpha = 0.05$ a significant correlation between the driving efficiency and the outdoor temperature could not be detected, since $0.6573 > 0.05$.

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

All the answers are about the correlation between the two variables, and we know that a test for significant correlation is equivalent with a test for significant slope in the simple regression model, i.e. $H_0 : \beta_1 \neq 0$.

So, we need to find the right p-value for this null hypothesis for making the conclusion. It's the `Pr(>|t|)` printed for the `Toutdoor` variable, which is only present in one answer. We check that answer and see that it is correct in terms of significance level and conclusion: Yes, it is significant since the p-value is lower than the significance level.

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

Question VI.2 (21)

The battery size was 54 kWh. How long is the predicted mean driving range at a temperature level of 5 °C according to the model and the estimated parameters?

1 ☐ 250 km

2 ☐ 260 km

3 ☐ 270 km

4 ☐ 280 km

5* ☐ 290 km

Continue on page 25

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

We must calculate the predicted value, so we insert the estimated parameters in the (simple linear) model, with $x = 5$ and get

```
5.10493 + 0.05259 * 5
```

```
## [1] 5.36788
```

which is 'kilometers per kWh', hence we scale it with the battery size to get

```
(5.10493 + 0.05259 * 5) * 54
```

```
## [1] 289.8655
```

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

Question VI.3 (22)

The $i = 5$ data point had the observation of temperature at 2.096 °C and at a driving efficiency of 4.744 km/kWh. What is the residual (i.e. the realized error) for this data point?

1* ☐ -0.471

2 ☐ 0.226

3 ☐ 0.657

4 ☐ 0.843

5 ☐ 1.634

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

The residual is the observed error, which is the difference between the observed and predicted model output value at the point

$$y_i - (\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i)$$

So it's found by

```
4.744 - (5.10493 + 0.05259 * 2.096)
```

```
## [1] -0.4711586
```

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

Question VI.4 (23)

The car owner wanted to investigate the effect of humidity on the driving range and therefore got hold of observations of air humidity from a nearby weather station and matched them with her observations.

She wanted to fit a multiple linear regression model with both the temperature and the humidity as inputs, but before she did some considerations. Which of the following statements about fitting a multiple linear regression model is not correct?

1 ☐ It's most often a good idea to investigate scatter plots of all possible pairs of the variables (a pairs plot in R).

2 ☐ It's important to carry out a model selection.

3* ☐ The level of correlation between the inputs cannot impact the results.

4 ☐ The number of observations impacts the results.

5 ☐ It's important to carry out a model validation with the selected model.

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

We check each answer and find that the one about correlated inputs not having impact on results is wrong. If the level of correlation of the inputs is high, then it impacts the results heavily, it's called collinearity, see Section 6.3.

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

Continue on page 28

Question VI.5 (24)

She fitted a multiple linear regression model with both the outdoor air temperature and the humidity. The obtained result was:

```
##
## Call:
## lm(formula = Efficiency ~ Toutdoor + Humidity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76776 -0.34382 -0.01327  0.38670  1.34920
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.96398    0.24226  20.490  < 2e-16 ***
## Toutdoor     0.06454    0.01952   3.306  0.00211 **
## Humidity    -0.16622    0.11379  -1.461  0.15250
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6477 on 37 degrees of freedom
## Multiple R-squared:  0.2282,Adjusted R-squared:  0.1864
## F-statistic: 5.469 on 2 and 37 DF,  p-value: 0.008302
```

What is the conclusion of a backward selection step on a 5% significance level for the fitted multiple linear regression model from this result (both the conclusion and the argument must be correct)?

1 ☐ None of the two inputs should be removed from the model, since $0.2282 > 0.05$.

2 ☐ The outdoor temperature should be removed from the model, since $0.00211 < 0.05$.

3 ☐ The outdoor temperature should be removed from the model, since $0.06454 > 0.05$.

4* ☐ The humidity should be removed from the model, since $0.15250 > 0.05$.

5 ☐ The humidity should be removed from the model, since $0.16622 > 0.05$.

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

In a backward selection step the least insignificant input is removed from the model. In this case we see that the humidity is the only insignificant input (i.e. it has p-value above the significance level for the test of being different from zero), and hence is must be removed.

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

Continue on page 29

The model can be formulated as

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

with the columns of $\boldsymbol{X}$ being; a vector of ones, measured ourdoor temperature, and measured humidity.

Question VI.6 (25)

What are the diagonal elements of the matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$?

1 ☐ $(0.374, 0.0301, 0.176)$

2 ☐ $(0.0906, 0.000588, 0.0200)$

3 ☐ $(0.612, 0.174, 0.419)$

4* ☐ $(0.140, 0.000908, 0.0309)$

5 ☐ $(0.577, 0.0465, 0.271)$

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

The standard errors can be calculated by

$$SE_i = \hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}}$$

hence the diagonal elements can be calculated by

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii} = \frac{SE_i^2}{\hat{\sigma}^2}$$

In R this is calculated by

```
c(0.24226,0.01952,0.11379)^2/0.6477^2
```

```
## [1] 0.1398993291 0.0009082634 0.0308645821
```

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

Continue on page 31

Exercise VII

A family is playing an old game called Mouse. In this game 10 pieces of candy are put on a plate. One family member is the player and looks away, while the others point to 2 pieces of candy, which are then called the "mice". The player now selects one piece at a time. If the selected one is a mouse, the turn is over and the player keeps all the pieces picked up before selecting the "mouse". It can be assumed that the player selects the pieces completely at random.

Question VII.1 (26)

What is the probability that the player gets all 8 possible pieces?

1* ☐ 2.2%

2 ☐ 3.6%

3 ☐ 5.8%

4 ☐ 6.4%

5 ☐ 9.2%

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

This is an experiment which follows the hypergeometric distribution, hence it is "count the number of successes in n binary draws without replacement). The answer can be found in the following different ways:

```
dhyper(8, 8, 2, 8)

## [1] 0.02222222

dhyper(0, 2, 8, 8)

## [1] 0.02222222

8/10 * 7/9 * 6/8 * 5/7 * 4/6 * 3/5 * 2/4 * 1/3

## [1] 0.02222222

2/10 * 1/9

## [1] 0.02222222
```

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

Continue on page 32

Exercise VIII

Let $X_i \sim LN(\mu, \sigma^2)$, $i = \{1, ..., n\}$ (i.e. $\log(X_i) \sim N(\mu, \sigma^2)$) be independent random variables.

Question VIII.1 (27)

What is the probability $P(X_1 X_2 > k)$?

1 ☐ $1 - F(k)$, where F is the distribution function for a log-normal distribution with mean μ and variance σ^2

2 ☐ $1 - F\left(\frac{k-2\mu}{2\sigma^2}\right)$, where F is the distribution function for a standard normal random variable.

3 ☐ $1 - F(k)$, where F is the distribution function for a log-normal distribution with mean 2μ and variance σ

4 ☐ $1 - F(k)$, where F is the distribution function for a log-normal distribution with mean 2μ and standard deviation 2σ

5* ☐ $1 - F\left(\frac{\log k - 2\mu}{\sqrt{2}\sigma}\right)$, where F is the distribution function for a standard normal random variable.

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

First note that $Y = \log(X_1) + \log(X_2) \sim N(2\mu, 2\sigma^2)$, and therefore $X_1 X_2 \sim LN(2\mu, 2\sigma^2)$, hence 1 and 3 are wrong. 4 is also wrong since the mean of the log-normal distribution is not equal μ (standard deviation is also wrong). The remaning 2 answers rely on standardization and

$$Z = \frac{Y - 2\mu}{\sqrt{2}\sigma} \sim N(0, 1)$$

and also

$$\begin{aligned} P(X_1 X_2 > k) =& P(\log(X_1) + \log(X_2) > \log(k)) \\ =& P\left(\frac{Y - 2\mu}{\sqrt{2}\sigma} > \frac{\log(k) - 2\mu}{\sqrt{2}\sigma}\right) \\ =& P\left(Z > \frac{\log(k) - 2\mu}{\sqrt{2}\sigma}\right) \\ =& 1 - F\left(\frac{\log(k) - 2\mu}{\sqrt{2}\sigma}\right) \end{aligned}$$

where F is the distribution function for a standard normal distribution, hence answer no 5 is the correct answer.

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

Question VIII.2 (28)

The quantity

$$Q = \left(\prod_{i=1}^{n} X_i \right)^{1/n} \tag{6}$$

is also called the geometric mean. What is the mean and variance of Q?

1* ☐ $E[Q] = e^{\mu + \frac{1}{2}\frac{\sigma^2}{n}}$ and $V[Q] = \left(e^{\sigma^2/n} - 1\right) e^{2\mu + \frac{\sigma^2}{n}}$

2 ☐ $E[Q] = \mu$ and $V[Q] = \sigma^2/n$

3 ☐ $E[Q] = e^{\mu + \frac{1}{2}\sigma^2}$ and $V[Q] = \left(e^{\sigma^2} - 1\right) e^{2\mu + \sigma^2}$

4 ☐ $E[Q] = e^{\mu/n}$ and $V[Q] = e^{\sigma^2/n}$

5 ☐ $E[Q] = (\mu)^{1/n}$ and $V[Q] = (\sigma^2/n)^{1/n}$

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

Taking the logarithm we get

$$\log(Q) = \frac{1}{n} \sum_{i=1}^{n} \log(X_i) \sim N(\mu, \sigma^2/n)$$

and hence $Q \sim LN(\mu, \sigma^2/n)$, and therefore

$$\begin{aligned} E[Q] &= e^{\mu + \frac{1}{2}\frac{\sigma^2}{n}} \\ V[Q] &= \left(e^{\sigma^2/n} - 1\right) e^{2\mu + \frac{\sigma^2}{n}} \end{aligned}$$

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

Continue on page 35

Question VIII.3 (29)

Let $X_i \sim N(\mu_1, \sigma_1^2)$ and $Y_j \sim N(\mu_2, \sigma_2^2)$ ($i = \{1, ..., n_1\}$ and $j = \{1, ..., n_2\}$) be independent random variables. Furthermore, let

$$Z_1 = \sum_{i=1}^{n_1} (X_i - \mu_1)^2$$
$$Z_2 = \sum_{j=1}^{n_2} (Y_j - \mu_2)^2$$

for which value of σ_1^2 does it hold that $E[Z_1/Z_2] = 1$?

1 ☐ $\sigma_1^2 = \frac{\sigma_2^2 n_1 (n_2 - 2)}{n_2^2}$

2* ☐ $\sigma_1^2 = \frac{\sigma_2^2 (n_2 - 2)}{n_1}$

3 ☐ $\sigma_1^2 = \frac{\sigma_2^2 \mu_1 (n_1 - 1)(n_2 - 3)}{\mu_2 (n_2 - 1)^2}$

4 ☐ $\sigma_1^2 = \frac{\sigma_2^2 n_2}{n_1}$

5 ☐ $\sigma_1^2 = \frac{\sigma_2^2 \mu_1 (n_2 - 3)}{\mu_2 (n_2 - 1)}$

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

First note that $Z_1/\sigma_1^2 \sim \chi^2(n_1)$ and $Z_2/\sigma_2^2 \sim \chi^2(n_2)$ and hence

$$\frac{Z_1/(\sigma_1^2 n_1)}{Z_2/(\sigma_2^2 n_2)} \sim F(n_1, n_2)$$

and we can write

$$\begin{aligned} E\left[\frac{Z_1}{Z_2}\right] &= \frac{n_1 \sigma_1^2}{n_2 \sigma_2^2} E\left[\frac{Z_1/(\sigma_1^2 n_1)}{Z_2/(\sigma_2^2 n_2)}\right] \\ &= \frac{n_1 \sigma_1^2}{n_2 \sigma_2^2} \frac{n_2}{n_2 - 2} \\ &= \frac{n_1 \sigma_1^2}{\sigma_2^2} \frac{1}{n_2 - 2} \end{aligned}$$

and hence $E[Z_1/Z_2] = 1$ for

$$\sigma_1^2 = \frac{\sigma_2^2 (n_2 - 2)}{n_1}$$

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

Question VIII.4 (30)

Let $X_i \sim N(0,1)$ $(i = \{1, ..., n\})$ be iid random variables, and let S^2 and $\bar{X}$ be defined in the usual way, what is the distribution of $Q = n\bar{X}^2 + (n-1)S^2$?

1 ☐ $Q \sim F(1, n)$

2 ☐ $Q \sim N(0, n-1)$

3 ☐ $Q \sim \chi^2(n-1)$

4 ☐ $Q \sim F(1, n-1)$

5* ☐ $Q \sim \chi^2(n)$

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

$\bar{X} \sim N(0, 1/n)$ and hence $\sqrt{n}\bar{X} \sim N(0,1)$ and therefore $n\bar{X}^2 \sim \chi^2(1)$ and further $(n-1)S^2 \sim \chi^2(n-1)$. Since $\bar{X}$ and S^2 are indepedent it follows that $n\bar{X}^2 + (n-1)S^2 \sim \chi^2(n)$.

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

The exam is finished. Enjoy the summer!