Technical University of Denmark *Page 1 of 35 pages.*

Written examination: 30. May 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	II.1	II.2	II.3	III.1	III.2	III.3	III.4	III.5	IV.1
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer	3	4	3	1	4	1	2	5	1	5

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

Exercise	VIII.1	VIII.2	VIII.3	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	3	2	5	4	2	4	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. 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 incidence of occupational disease in an industry is such that the workers have a 20% chance of suffering from it.

Question I.1 (1)

What is the probability that out of 6 randomly selected workers 4 or more will contract disease?

1 ☐ 0.000064

2 ☐ 0.01536

3* ☐ 0.01696

4 ☐ 0.90112

5 ☐ 0.9984

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

This a binomial experiment and we get the probability by:

```
1-pbinom(3, size=6, prob = 0.2)
```

```
## [1] 0.01696
```

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

Continue on page 3

Exercise II

In a study three new products were tested to compare the user experience of each. The products were named "A", "B" and "C". Prototypes of each product were randomly send to testers, and they reported back their experiences with the product in an interview. Their responses were rated according to how much they liked the product – and counted into one of three categories: "Low", "Medium", "High". The results were read into R by:

```
mat <- matrix(c(24, 21, 14,
                12, 15, 22,
                15, 26, 24), ncol = 3, byrow = TRUE)
colnames(mat) <- c("Low","Medium","High")
rownames(mat) <- c("A","B","C")
```

and presented in a table:

	Low	Medium	High
A	24	21	14
B	12	15	22
C	15	26	24

Researchers want to test if the three products are rated significantly different. Hence the following null hypothesis must be tested

$$H_0 : p_{i,1} = p_{i,2} = p_{i,3} \text{ for } i = 1, 2, 3$$

where $p_{i,j}$ denotes the proportion in row i and column j.

Question II.1 (2)

What is the expected value of counts in the "Medium" rating category for product "B" under the null hypothesis?

1 ☐ 0.087

2 ☐ 15.0

3 ☐ 16.3

4* ☐ 17.6

5 ☐ 19.1

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

We are given the code to read the data in R, so the easiest is:

```
# Write the table in R
mat

##   Low Medium High
## A  24     21   14
## B  12     15   22
## C  15     26   24

# The number of observations
(n <- sum(mat))

## [1] 173

# The expected counts under H0
fit$expected["B","Medium"]

## [1] 17.6

# or
margin.table(mat, 1)["B"] * margin.table(mat, 2)["Medium"] / n

##    B
## 17.6
```

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

Question II.2 (3)

What is the outcome of the test of the null hypothesis on a 5% significance level (both conclusion and argument must be correct)?

1 ☐ The p-value is below the significance level, hence the null hypothesis is accepted.

2 ☐ The p-value is below the significance level, hence the null hypothesis is rejected.

3* ☐ The p-value is above the significance level, hence the null hypothesis is accepted.

4 ☐ The p-value is above the significance level, hence the null hypothesis is rejected.

5 ☐ There has not been provided enough information to calculate the p-value and make the conclusion.

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

Again, since we have the code for reading data into R, it's easiest to:

```
fit <- chisq.test(mat, correct=FALSE)
fit

##
##  Pearson's Chi-squared test
##
## data:  mat
## X-squared = 8, df = 4, p-value = 0.09
```

The p-value is above the significance level, so the null hypothesis is accepted.

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

Question II.3 (4)

In this question we only consider observations for product "A":

Low	Medium	High
24	21	14

what is the 98% confidence interval for the proportion of "Low" ratings for product "A" (Note that, the result from relevant R function is slightly different from the correct answer, when rounded it's within ± 0.01)?

1* ☐ $[0.26, 0.56]$

2 ☐ $[0.32, 0.81]$

3 ☐ $[0.37, 0.76]$

4 ☐ $[0.49, 0.83]$

5 ☐ $[0.52, 0.81]$

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

We need to just consider part of the data, we easily do this not using R:

```r
(x <- mat["A","Low"])

## [1] 24

(n <- sum(mat["A", ]))

## [1] 59
```

and then we can calculate the interval, either using the formula:

```r
# Using the formula from the book
(pha <- x / n)

## [1] 0.407

pha + c(-1,1) * qnorm(0.99) * (sqrt(pha*(1-pha)/n))

## [1] 0.258 0.556
```

or using the function:

```r
# or using the function
prop.test(x, n, conf.level = 0.98, correct=FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  x out of n, null probability 0.5
## X-squared = 2, df = 1, p-value = 0.2
## alternative hypothesis: true p is not equal to 0.5
## 98 percent confidence interval:
##  0.272 0.557
## sample estimates:
##     p
## 0.407
```

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

Continue on page 7

Exercise III

In the process of curing of concrete, the temperature inside the concrete will increase rapidly for a period of time due to chemical processes. After the increase the temperature will decrease until the temperature of the surroundings is reached. The strength of the concrete can be assessed from the temperature profile (i.e. how the temperature increased and decreased).

In the R-code below `x1` represents the change in temperature inside the concrete from the beginning of an hour to the end of the hour, on Day 3 after the cast began:

```
t.test(x1)

##
##  One Sample t-test
##
## data:  x1
## t = -4.1246, df = 21, p-value = 0.0004823
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.10939584 -0.03605871
## sample estimates:
##   mean of x
## -0.07272727
```

In the questions below it can be assumed that the observations are from a normal distribution with expectation μ_1 and variance σ_1^2, furthermore it may be assumed that the observations are independent.

Question III.1 (5)

Based on the R-output above and using significance level $\alpha = 0.05$, can it be concluded that the temperature is decreasing, corresponding to $\mu_1 < 0$ (both the conclusion and the argument must be correct)?

1 ☐ Yes, since $0.00048 < 0.05$

2 ☐ Yes, since $-0.073 < 0$

3 ☐ No, since $0.073 > 0.05$

4* ☐ Yes, since $-0.036 < 0$

5 ☐ No, since $-4.12 < 0$

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

Out of the five answer only one of them contains the correct answer and fulfilling argument. The p-value of $H_0 : \mu_1 \neq 0$ is 0.00048, so we reject that it is equal to zero, however it's not enough argument to conclude that μ_1 is below zero. The argument $-0.036 < 0$ is fulfilling, since -0.036 is the upper limit of the 95% confidence interval, hence when zero is not included in the interval, and the interval is below zero, then it's a sufficient argument to conclude that $\mu_1 \neq 0$.

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

In the R-code below `x2` represents the hourly temperature differences on Day 4 after the cast:

```
mean(x2)

## [1] -0.1181818

sd(x2)

## [1] 0.05884899

length(x2)

## [1] 22
```

It can be assumed that the observations are normal and independent, with mean μ_2 and variance σ_2^2.

Question III.2 (6)

μ_2 represents the expected value on the Day 4, what is the 95% confidence interval for μ_2?

1* ☐ $[-0.144, -0.092]$

2 ☐ $[-0.140, -0.0966]$

3 ☐ $[-0.135, -0.101]$

4 ☐ $[-0.124, -0.113]$

5 ☐ $[-0.120, -0.117]$

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

We take the values for `x2` given above and insert them in the one-sample confidence interval formula: Equation 3-10.

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

Question III.3 (7)

σ_2^2 represents the variance on the Day 4, what is the 95% confidence interval for the standard deviation σ_2?

1 ☐ $[0.0472, 0.0792]$

2* ☐ $[0.0453, 0.0841]$

3 ☐ $[0.0348, 0.120]$

4 ☐ $[0.00266, 0.00495]$

5 ☐ $[0.00205, 0.00707]$

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

We use the formula for the confidence interval for the standard deviation with the values given above for `x2`. It's the second formula in Method 3.19.

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

Question III.4 (8)

Assuming equal variance in the two groups, what is the usual test statistics for the test H_0 : $\mu_1 = \mu_2$ against the two-sided alternative?

1 ☐ 3.62

2 ☐ 1.81

3 ☐ 2.58

4 ☐ 1.78

5* ☐ 2.10

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

This is a bit tricky question, because calculating the answer takes a couple of steps. We want to use Method 3.53, so we need to first find pooled variance estimate. We have to find the mean and standard deviation of the two samples.

For `x1` we have to rearrange the confidence interval formula (Equation 3-10) and use the result from `t.test` given in the beginning of the Exercise:

```
s1 <- (0.10939584 - 0.03605871)/(2*qt(0.975,df=21))*sqrt(22)
m1 <- -0.07272727
```

For `x2` we have it:

```
s2 <- 0.05884899
m2 <- -0.1181818
```

Then we can insert that to get the answer:

```
s <- sqrt((21*s1^2 + 21*s2^2)/(44-2))
(m1-m2)/(s*sqrt(1/22+1/22))

## [1] 2.100419
```

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

Question III.5 (9)

Using the standard deviation from Day 4, and significance level $\alpha = 0.05$ how many observations would be needed to detect a mean of -0.05 (when using the null hypothesis that the slope is zero), if the required power is 0.9 (the correct answer is calculated using the formula in the book)?

1* ☐ 15

2 ☐ 8

3 ☐ 26

4 ☐ 5

5 ☐ 12

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

We use Method 3.65:

```
(s2 * (qnorm(0.9)+qnorm(0.975)) / (0.05))^2

## [1] 14.55574

# or with the R function (giving a slightly higher number because it uses the t-distr
power.t.test(delta=-0.05, sd=s2, sig.level=0.05, power=0.9, type="one.sample")

##
##      One-sample t test power calculation
##
```

```
##               n = 16.59608
##           delta = 0.05
##              sd = 0.05884899
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
```

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

Continue on page 13

Exercise IV

Yearly measurements of temperature (in °C) for a region of northern Italy were collected in the years 1984-2005. The data was read into R:

```
temperature <- c(8.43, 7.89, 8.28, 7.84, 9.62, 9.41, 9.40, 8.22, 9.18, 9.17,
                 9.25, 9.68, 8.49, 8.53, 9.30, 8.94, 9.46, 9.69, 9.37, 9.42,
                 9.13, 9.18)
year <- 1984:2005
```

and a linear regression was carried out on the data:

```
##
## Call:
## lm(formula = temperature ~ year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80872 -0.31761  0.03158  0.29517  0.92517
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -82.97094   33.74565  -2.459   0.0232 *
## year          0.04611    0.01692   2.725   0.0130 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5035 on 20 degrees of freedom
## Multiple R-squared:  0.2708,Adjusted R-squared:  0.2343
## F-statistic: 7.427 on 1 and 20 DF,  p-value: 0.01304
```

Question IV.1 (10)

What is the estimate for the expected temperature increase over 10 years?

1 ☐ 0.046 °C

2 ☐ -8.2 °C

3 ☐ 0.017 °C

4 ☐ 2.7 %

5* ☐ 0.46 °C

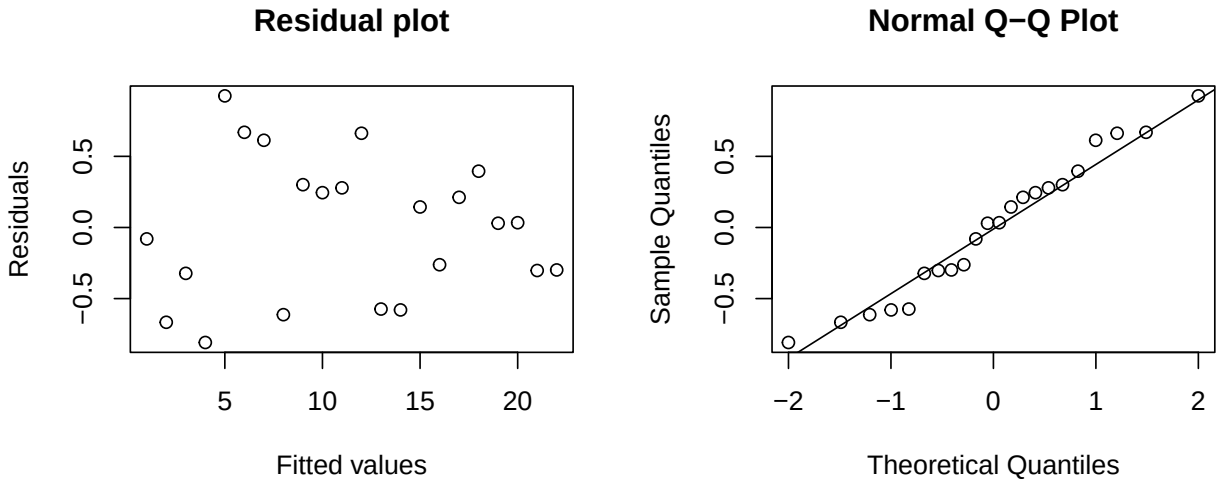
---------- FACIT-BEGIN ----------

The annual increase is 0.046. Multiply by 10 to get 0.46 °C.

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

Question IV.2 (11)

The plots below show a residual plot and a normal q-q plot of the residuals:



Which of the following statements is the correct interpretation of the two plots?

1 ☐ We see no linear tendency in the residual plot. This is evidence for the null hypothesis of no significant effect of time on temperature.

2 ☐ The residual plot looks (reasonably) fine, but the q-q plot looks questionable. This indicates a problem with the normality assumption.

3 ☐ The residual plot looks (reasonably) fine, but the q-q plot looks questionable. This indicates a problem with the linear dependence assumption.

4 ☐ Both plots look questionable. This indicates problems with the linear dependence assumption and the normailty assumption.

5* ☐ Both the residual plot and the q-q plot look (reasonably) fine and this confirms the validity of the model.

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

Both q-q plot and residual plot look fine, see Section 5.7.

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

Question IV.3 (12)

Suppose the temperature for the same region in 2017 was 10.91 °C. Which of the following statements is a correct (both argument and conclusion must be correct)?

1 ☐ The 95% confidence interval is $[8.21, 9.86]$. The observation fits reasonably well with the model.

2 ☐ The 95% confidence interval is $[8.21, 9.86]$. The observation does not fit well with the model.

3* ☐ The 95% prediction interval is $[8.70, 11.37]$. The observation fits reasonably well with the model.

4 ☐ The 95% prediction interval is $[8.70, 11.37]$. The observation does not fit well with the model.

5 ☐ None of the above statements are correct.

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

Either use the prediction interval formula in Equation 5-60 or use the `predict` function in R, e.g.

```
predict(lm(temperature ~ year), newdata = data.frame(year = 2017), interval = "predictio

##        fit      lwr      upr
## 1 10.03201 8.696461 11.36756
```

Since the observation is inside the prediction interval, it is reasonable to say that the observation fits well with the model.

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

Continue on page 16

Exercise V

In order to understand why it can be hard to live in Denmark during the dark times in the winter, an analysis was carried out. From the Danish Meteorological Institute the sunlight arriving during December and January, at the Isenvad station in mid-Jutland, was obtained for 10 years. From the data the longest period with no registered sunlight was calculated for each winter:

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Period length in days	3.7	1.9	4.8	11.7	2.8	4.7	2.7	4.9	6.7	3.8

Question V.1 (13)

What is the median of the sample?

1 ☐ 3.8

2* ☐ 4.25

3 ☐ 4.7

4 ☐ 4.77

5 ☐ 4.875

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

Sort the observations and then, since there is an equal number, take the average of the two observations in the middle. If the values were written into R:

```
median(x)

## [1] 4.25
```

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

Question V.2 (14)

One wants to estimate a 90% confidence interval for the mean of the longest period without sunshine in Isenvad during the years. The sample is stored in the vector `x`. Which of the following code snippets calculates the confidence interval without any assumption of distribution?

1 ☐
```
simsamples <- replicate(10000, sample(x, replace = FALSE))
quantile(apply(simsamples, 2, mean), c(0.025, 0.975))
```

2* ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
quantile(apply(simsamples, 2, mean), c(0.05, 0.95))
```

3 ☐
```
simsamples <- replicate(10000, sample(x, replace = FALSE))
quantile(apply(simsamples, 2, median), c(0.025, 0.975))
```

4 ☐
```
t.test(x, conf.level=0.9)
```

5 ☐
```
t.test(x, conf.level=0.95)
```

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

It cannot be with the `t.test` function, since that would require an assumption of normal distribution (although if we have $n > 30$ observations the CLT will make t.test result ok without normal dist.).

So it has to be with bootstrapping, and we can rule out the answer with `median` and the answer with `replace = FALSE`. So only one answer is left:

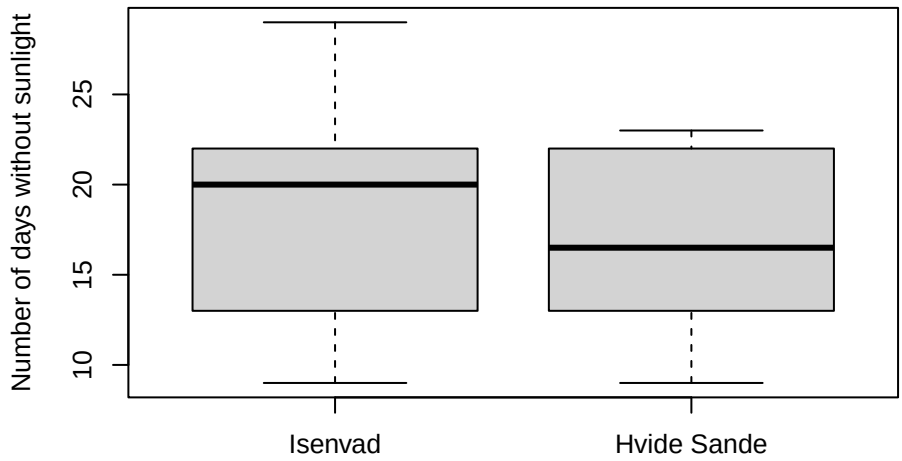
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
quantile(apply(simsamples, 2, mean), c(0.05, 0.95))

##   5%  95%
## 3.53 6.25
```

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

Question V.3 (15)

Furthermore, an analysis was carried out where the sunlight in Isenvad in the middle of Jutland, was compared to the sunlight, in Hvide Sande at the west coast of Jutland. The number of days with no registered sunlight during December and January at each location, for each of the same 10 years was calculated. The observations are summarized in the following boxplot:



The observations are sorted according to the year and for Isenvad stored in `x` and for Hvide Sande in `y`:

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
x	9	13	18	29	19	22	13	22	26	21
y	9	12	13	23	17	22	15	17	22	16

The following R code was executed:

```
mean(x)

## [1] 19.2

mean(y)

## [1] 16.6

k <- 10000
simxsamples <- replicate(k, sample(x, replace = TRUE))
simysamples <- replicate(k, sample(y, replace = TRUE))
sim1 <- apply(simxsamples, 2, mean) - apply(simysamples, 2, mean)

simsamples <- replicate(10000, sample(x-y, replace = TRUE))
sim2 <- apply(simsamples, 2, mean)

quantile(sim1, c(0.005, 0.995))

##  0.5% 99.5%
##  -3.5   8.7
```

```r
quantile(sim2, c(0.005, 0.995))
```

```
##  0.5% 99.5%
##   0.4   4.6
```

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

1 ☐ Two parametric bootstrapping 99% confidence intervals were calculated.

2 ☐ At a 5% significance level it cannot be concluded that there is a significant difference between the number of days with no sunlight for the two locations.

3* ☐ At a 5% significance level it can be concluded that there is a significant difference between the number of days with no sunlight for the two locations.

4 ☐ The sample mean for Isenvad is lower than for Hvide Sande.

5 ☐ None of the above statements are correct.

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

We have to use the result of `sim2`, since it's the samples can be paired on the year, so we do the calculations as one-sample on the differences between each year. Since the 99% confidence interval does not contain zero, then the 95% confidence interval (which is always narrower) will also not contain zero.

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

Continue on page 20

Exercise VI

A family is looking for a summer house. They really love summer and sunshine, and will mostly use the summer house in July. Therefore they downloaded data of sunlight hours observed at Hvide Sande located at the west coast of Jutland, and similarly at Hammer Odde located at Bornholm. They have taken the difference in sunlight hours between the two locations for each day during the last 10 years in July.

Let the i'th observed difference in hours be represented by x_i, such that $x_i > 0$ implies that there was more sunlight at Bornholm compared to the west coast of Jutland.

They decide to restrict their search for a summer house to the location with more sunlight hours, if a statistical test can reveal statistical evidence of a difference between the locations at a significance level of 5%. The values are saved in the vector `x` in R and following result was obtained:

```
t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 4.722, df = 278, p-value = 3.708e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.150853 5.226247
```

Question VI.1 (16)

How many observations were included in the analysed sample?

1 ☐ $n = 9$

2 ☐ $n = 10$

3 ☐ $n = 11$

4 ☐ $n = 278$

5* ☐ $n = 279$

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

We read off the degrees of freedom in the result to $df = 278$ and we know that for the one-sample t-test carried out $df = n - 1$, so $n = df + 1$=279.

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

Question VI.2 (17)

What does the family conclude based on the result (both argument and location decision must be correct)?

1* ☐ There is strong evidence against $H_0 : \mu_X = 0$, leading them to only search for a summer house at Bornholm.

2 ☐ There is strong evidence against $H_0 : \mu_X = 0$, leading them to only search for a summer house at the west coast of Jutland.

3 ☐ There is weak evidence against $H_0 : \mu_X = 0$, leading them to only search for a summer house at Bornholm.

4 ☐ There is weak evidence against $H_0 : \mu_X = 0$, leading them to only search for a summer house at the west coast of Jutland.

5 ☐ There is litte or no evidence against $H_0 : \mu_X = 0$, leading to them search for a summer house at both locations.

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

Using Table 3.1 and the p-value in the result the family conclude that there are strong evidence against $H_0 : \mu_X = 0$ and since the estimated mean is above zero, then it leads them to conclude that the there is more sunlight at Bornholm – thus they will search for a summerhouse only there.

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

Question VI.3 (18)

Which of the following statements regarding the sample mean of `x` is correct?

1 ☐ The sample mean is 2.613.

2 ☐ The sample mean is 3.075.

3* ☐ The sample mean is 3.689.

4 ☐ The sample mean is 4.722.

5 ☐ Not enough information has been given to calculate the sample mean.

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

It's simply the middle of the confidence interval:

```
2.150853 + (5.226247 - 2.150853) * 0.5
```

```
## [1] 3.68855
```

```
mean(x)
```

```
## [1] 3.68855
```

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

Continue on page 23

Exercise VII

Customers at a bank arrive at random and independently; the probability of an arrival in any 1-minute period is the same as the probability of an arrival in any other 1-minute period. Answer the following questions, assuming a mean arrival rate of three customers per minute.

Question VII.1 (19)

What is the probability of exactly three customers arriving in a randomly selected 1-minute period?

1* ☐ 0.2240

2 ☐ 0.4232

3 ☐ 0.5768

4 ☐ 0.6472

5 ☐ 0.7760

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

We must use the Poisson distribution and the probability for exactly three costumers arriving is given by the pdf:

```
dpois(3,3)

## [1] 0.2240418
```

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

Question VII.2 (20)

What is the probability of at least three arrivals in a randomly selected 1-minute period?

1 ☐ 0.2240

2 ☐ 0.4232

3* ☐ 0.5768

4 ☐ 0.6472

5 ☐ 0.7760

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

We must have at least three or above, so $P(X \geq 3) = 1 - P(X < 2)$, so:

```
1- ppois(2, lambda=3)
```

```
## [1] 0.5768099
```

?ppois

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

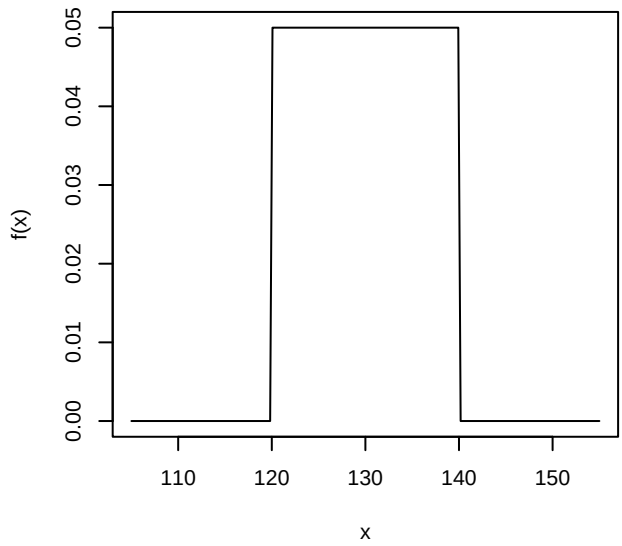
Continue on page 25

Exercise VIII

Consider the random variable X that represents the flight time (in minutes) of an airplane traveling from Chicago to New York. The probability density function for X is

$$f(x) = \begin{cases} 1/20 & 120 \leq x \leq 140 \\ 0 & \text{otherwise} \end{cases}$$

which is plotted below:



Question VIII.1 (21)

What is the probability of a flight time between 120 and 140 min?

1 ☐ 0.2

2 ☐ 0.5

3 ☐ 0.8

4 ☐ 0.9

5* ☐ 1.0

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

It's the uniform distribution from 120 to 140, so the entire interval and the probability is therefore 1.

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

Question VIII.2 (22)

What is the standard deviation of X?

1 ☐ 1.67

2 ☐ 3.33

3 ☐ 4.47

4* ☐ 5.77

5 ☐ 33.33

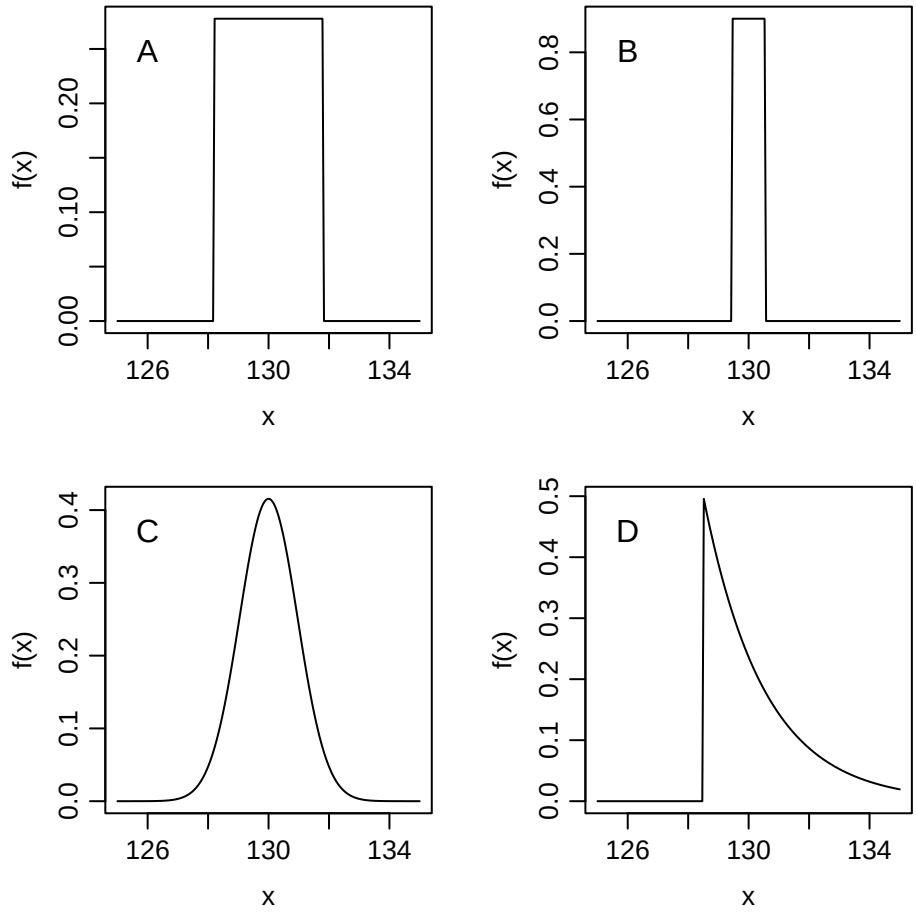
---------- FACIT-BEGIN ----------

Use the formula for the variance of a uniform distribution in Method 2.36.

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

Question VIII.3 (23)

If a random sample of $n = 36$ observations was taken of the flight times, which of the following plots would then represent a good approximation of the probability density function (pdf) of the sample mean $\bar{X}$?



1 ☐ Plot A

2 ☐ Plot B

3* ☐ Plot C

4 ☐ Plot D

5 ☐ None of the plots can be a good approximation to the pdf of the sample mean $\bar{X}$.

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

We know that the sample mean is approximated well with a normal distribution with mean μ_X and variance σ_X^2/n, which follows from the CLT in Theorem 3.14.

Since it's a uniform distribution we use Theorem 2.36 to obtain the mean of X

$$\mu_X = 130$$

and the standard deviation of X

$$\sigma_X = \sqrt{\sigma_X} = \sqrt{33.33} = 5.77$$

So we find the plot of a normal distribution with mean 130 and standard deviation

$$\sigma_{\bar{X}} = \sigma_X/\sqrt{n} = 5.77/6 = 0.96$$

(remember that roughly three times the standard deviations from the mean, the normal pdf is down at zero).

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

Continue on page 29

Exercise IX

We observed two variables, y and x_1, and carried out a linear regression:

```
summary(lm(y ~ x1))

##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44978 -0.20443 -0.12711  0.00835  1.11002
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.5178     0.1456  -10.43 6.21e-06 ***
## x1            0.4161     0.1547    2.69   0.0275 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4563 on 8 degrees of freedom
## Multiple R-squared:  0.4749,Adjusted R-squared:  0.4092
## F-statistic: 7.234 on 1 and 8 DF,  p-value: 0.02751
```

Question IX.1 (24)

Which of the following statements is correct regarding the output line `(intercept)` in the lm result?

1 ☐ The Std. Error expresses the uncertainty on the estimate of the regression slope.

2* ☐ The Std. Error expresses the uncertainty on the expected value of an observation, where $x_1 = 0$.

3 ☐ The t-value can be used to assess if there is a significant association between x_1 and y.

4 ☐ The t-value is a measure of model validity. A small t-value indicates a valid model.

5 ☐ Neither the Std. Error nor the t-value are related to the uncertainty of the model.

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

The expected value of a new observation is the same as the confidence interval for the line, as stated in text above Method 5.18: "confidence intervals, where we predict the mean value of future outcomes".

Checking the formula for the standard error of the intercept β_0 in Equation 5-43, with the formula for the confidence interval for the line in Equation 5-59, we can see that with $x_{\text{new}} = 0$ it's the same expression for the standard error.

One can also exclude the other answers and see that the only reasonable answer is option 2, so even without getting the detail about the expected value, it's possible to find the correct answer.

Option 1 and 3 are correct for `x1`, not for the intercept.

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

Question IX.2 (25)

Suppose we included an additional variable x_2 to our model, i.e. $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$.

Which of the following statements is correct?

1 ☐ The explained variance r^2 (`Multiple R-squared` in the result) will decrease compared to the model with only x_1.

2 ☐ If we were to perform backward selection with significance level $\alpha = 0.05$, we would remove x_2 if the corresponding p-value in the result is 0.0275.

3 ☐ At most one of x_1 and x_2 will be significant on a 5% level.

4 ☐ At least one of x_1 and x_2 will be significant on a 5% level.

5* ☐ Each statement above is either not correct, or we have insufficient information conclude if it is.

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

The explained variance always increases if more variables are included. This rules out option 1. If x_1 are x_2 are (sufficiently) collinear, it may happen that none of them are significant. This rules out option 4. Potentially, there is an effect of x_2 in addition to that of x_1. This rules out option 3. The p-values usually change when more variables are included. This rules out option 2.

In conclusion, option 5 is the correct answer.

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

Continue on page 31

Exercise X

As part of a multilab study, three fabrics were tested for flammability at the National Bureau of Standards. The following burn times in minutes were recorded after a paper tab was ignited on the hem of a dress made of each fabric:

Fabric1	Fabric2	Fabric3
3.11	3.43	2.56
3.09	4.03	3.14
2.67	3.54	3.11
2.66	3.24	1.69
2.16	3.77	1.91
3.22	3.86	2.62
3.28	3.39	3.25

A one-way analysis of variance (ANOVA) was carried out. The resulting ANOVA table can be seen below (some elements has been substituted with question marks):

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Fabric	?	3.71	?	8.91	0.0020
Residuals	?	3.75	?		

Question X.1 (26)

Which of the following statements is correct?

1 ☐ For Residuals: Df = 21 and Mean Sq = 0.18

2 ☐ For Residuals: Df = 20 and Mean Sq = 0.19

3 ☐ For Residuals: Df = 19 and Mean Sq = 0.20

4* ☐ For Residuals: Df = 18 and Mean Sq = 0.21

5 ☐ For Fabric: Df = 3 and Mean Sq = 1.237

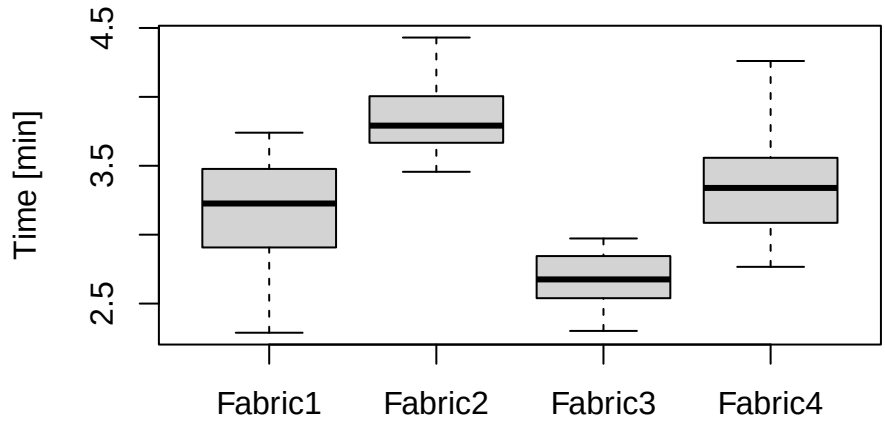
---------- FACIT-BEGIN ----------

The number of observations was $n = 21$. Therefore, the total number of degrees of freedom was $n - 1 = 20$. The number of degrees of freedom for Fabric is given by $k - 1 = 2$. The number of degrees of freedom for the residuals is given by $(n - 1) - (k - 1) = 18$. The residual variance can be calculated by dividing the Sum Sq with Df. Given the information provided above only answer 4 can be correct.

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

Question X.2 (27)

The flammability study was repeated for four other types of fabric. The results are shown in the boxplot:



A one-way ANOVA was carried out. The result is presented in the ANOVA table below (the values are as usual rounded):

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Fabric	3	13.82	4.61	43.20	0.0000
Residuals	76	8.10	0.11		

What would be the estimate for the variance in burn time, if all data were considered as one sample from a single population?

1 ☐ $\hat{\sigma}^2 = \frac{13.82}{3} + \frac{8.10}{76}$

2* ☐ $\hat{\sigma}^2 = \frac{13.82+8.10}{79}$

3 ☐ $\hat{\sigma}^2 = 4.61 + 0.11$

4 ☐ $\hat{\sigma}^2 = \frac{4.61}{3} + \frac{0.11}{76}$

5 ☐ We are not given sufficient information to calculate the variance estimate.

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

ANOVA decomposes the total Sum of Squares (SST), such that $SST = SS(Tr) + SSE$. We can extract $SS(Tr)$ and SSE from the ANOVA table above to find SST. The requested variance estimate can be calculated using the formula in Equation 8-10 and get $\hat{\sigma}^2 = \frac{SST}{n-1}$ with $df(total) = n - 1 = df(Fabric) + df(Residuals) = 76 + 3 = 79$.

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

Question X.3 (28)

Look at the ANOVA table from the previous question. We want to test the following hypothesis:

$$H_0 : \mu_{\text{Fabric1}} = \mu_{\text{Fabric2}} = \mu_{\text{Fabric3}} = \mu_{\text{Fabric4}} = \mu$$

Assuming a significance level $\alpha = 0.05$, which R command results in the correct critical value in the F-distribution to be used for the hypothesis test?

1 ☐ `pf(0.05, 3, 76)`

2 ☐ `pf(0.95, 3, 76)`

3 ☐ `pf(0.975, 3, 79)`

4* ☐ `qf(0.95, 3, 76)`

5 ☐ `qf(0.975, 3, 76)`

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

It must be qf to get a quantile from the F-distribution, and have $df_1 = k - 1$ and $df_2 = n - k$, where k is the number of groups.

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

Continue on page 34

Exercise XI

One is interested in examining the impact of four different teaching methods (A-D) with respect to student performance. A randomized block design was utilized, meaning that three students underwent all four teaching methods including corresponding examination in randomized order. The following data was gathered, where the best possible exam performance is 100 (percent):

	Student1	Student2	Student3
A	84	89	91
B	85	87	91
C	85	88	89
D	86	90	96

A two-way ANOVA with significance level $\alpha = 0.05$ was carried out to investigate if teaching method had a significant impact on student performance. The ANOVA table can be seen below. Some elements have been replaced by question marks:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Student	2	91.17	45.583	21.312	0.002
Method	3	20.92	?	?	?
Residuals	6	12.83	2.139		

Question XI.1 (29)

Which of the following statements is correct?

1 ☐ The p-value for Method is 0.56 and there is no significant effect of teaching method.

2 ☐ The p-value for Method is 0.18 and there is a significant effect of teaching method.

3 ☐ The p-value for Method is 0.18 and there is no significant effect of teaching method.

4* ☐ The p-value for Method is 0.10 and there is no significant effect of teaching method.

5 ☐ The p-value for Method is 0.10 and there is a significant effect of teaching method.

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

We can calculate $MS(Method) = \frac{SS(Method)}{Df(Method)}$ and $F(Method) = \frac{MS(Method)}{MS(Residuals)}$. We are interested in finding the p-value which represents the probability of obtaining an F-test statistic, which is at least as extreme as the observed (under the assumption that the Null hypothesis is true). We can find the p-value using the following R command:

`1 - pf(20.92/(3*2.14), 3, 6)` ≈ 0.10

The p-value is greater than 0.05, hence the there is no significant impact of teaching method.

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

Question XI.2 (30)

The ANOVA above indicates that there is a significant difference between student performance. We are now planning post-hoc tests for pairwise comparison of the student performance means. In order not to increase the risk of making a Type-I error we would like to correct our significance level α using Bonferroni correction.

Which of the following statements is correct?

1 ☐ We are performing 12 post-hoc test, hence we must divide α by 12.

2 ☐ We are performing 12 post-hoc test, hence we must divide α by 11.

3 ☐ We are performing 4 post-hoc test, hence we must divide α by 4.

4 ☐ We are performing 3 post-hoc test, hence we must divide α by 2.

5* ☐ We are performing 3 post-hoc test, hence we must divide α by 3.

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

We are aiming at comparing the 3 student performance means. Hence, we must carry our 3 post-hoc hypothesis tests. Bonferroni correction makes sure that the family wise Type-I risk does not inflate by dividing α by the number of hypothesis tests carried out. Therefore only answer 5 is correct.

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

The exam is finished. Enjoy the summer!