Written examination: 30. May 2021

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

Duration: 4 hours

Aids and facilities allowed: All

The questions were answered by

(student number)	(signature)	(table number)

This exam consists of 30 questions of the "multiple choice" type, which are divided between 11 exercises. To answer the questions, you need to fill in the "multiple choice" form (6 separate pages) on CampusNet with the numbers of the answers that you believe to be correct.

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

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

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

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										

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										

The exam paper contains 26 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

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

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.

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 ☐ $[0.52, 0.81]$

Continue on page 5

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 $-4.12 < 0$

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 ☐ $[-0.120, -0.117]$

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 ☐ $[0.00205, 0.00707]$

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

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

Continue on page 8

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

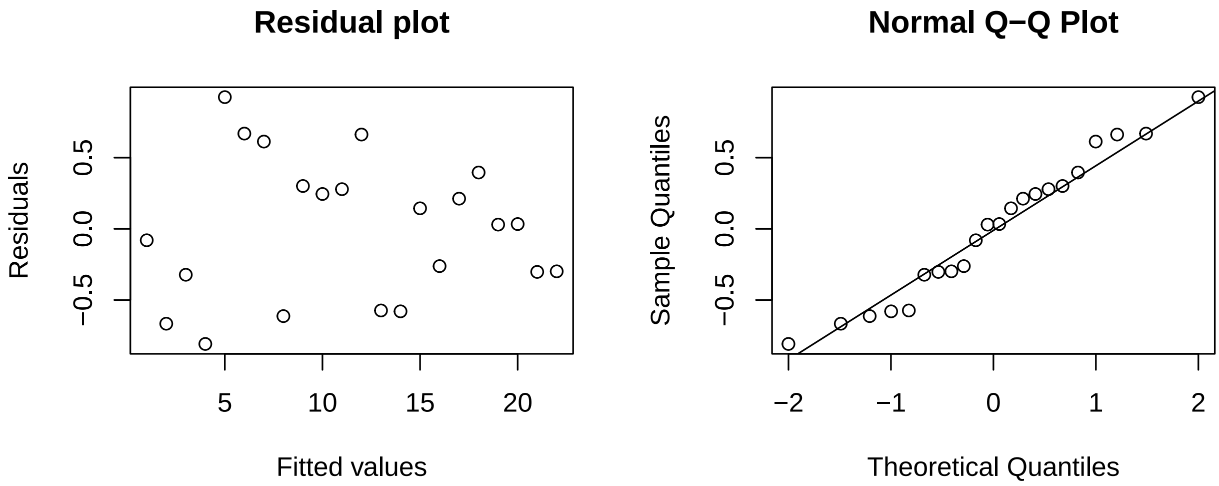
3 ☐ 0.017 °C

4 ☐ 2.7 %

5 ☐ 0.46 °C

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.

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 <u>not</u> 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 <u>not</u> fit well with the model.

5 ☐ None of the above statements are correct.

Continue on page 11

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

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))
```

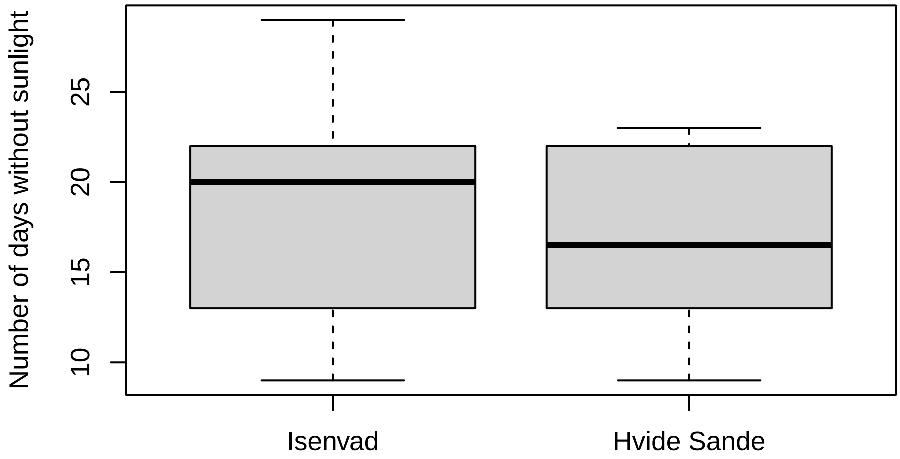
3 ☐
```
simsamples <- replicate(10000, sample(x, replace = FALSE))
quantile(apply(simsamples, 2, median), c(0.025, 0.975))
```

4 ☐ `t.test(x, conf.level=0.9)`

5 ☐ `t.test(x, conf.level=0.95)`

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

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.

Continue on page 14

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 ☐ $n = 279$

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.

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.

Continue on page 16

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

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

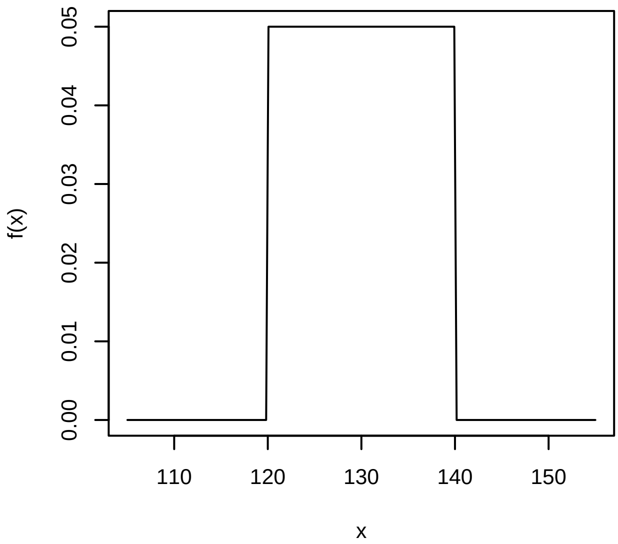
Continue on page 17

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

Question VIII.2 (22)

What is the standard deviation of X?

1 ☐ 1.67

2 ☐ 3.33

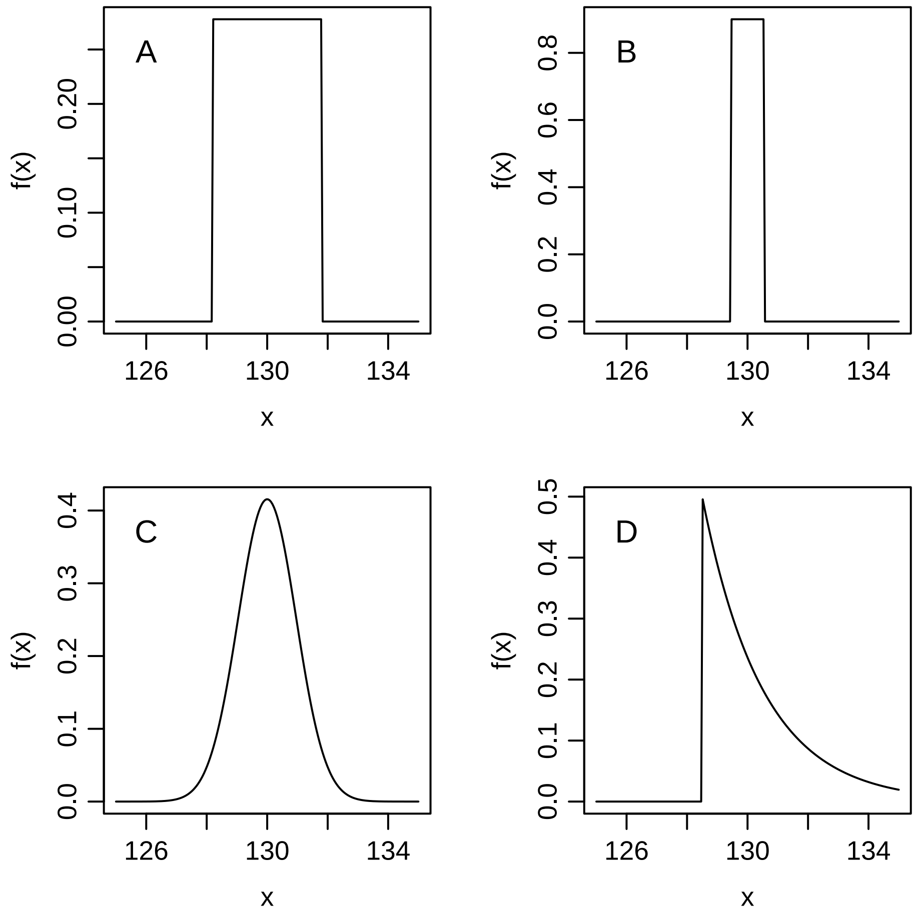
3 ☐ 4.47

4 ☐ 5.77

5 ☐ 33.33

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 $\bar{X}$.

Continue on page 20

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.

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.

Continue on page 22

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

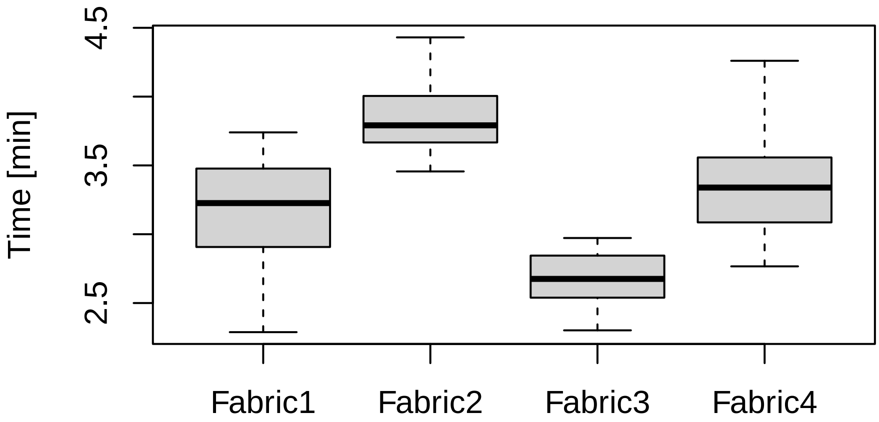
3 ☐ For Residuals: Df = 19 and Mean Sq = 0.20

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

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

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.

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)`

Continue on page 25

Exercise XI

The following sample is given:

i	x_i
1	8.41
2	8.33
3	6.79
4	6.76
5	4.92
6	8.79
7	9.03
8	7.75

Question XI.1 (29)

What is the estimated variance $\hat{\sigma}_x^2$?

1 ☐ $\hat{\sigma}_x^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_x)^2}{n} = 1.65$

2 ☐ $\hat{\sigma}_x^2 = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_x)^2}{n-1}} = 1.37$

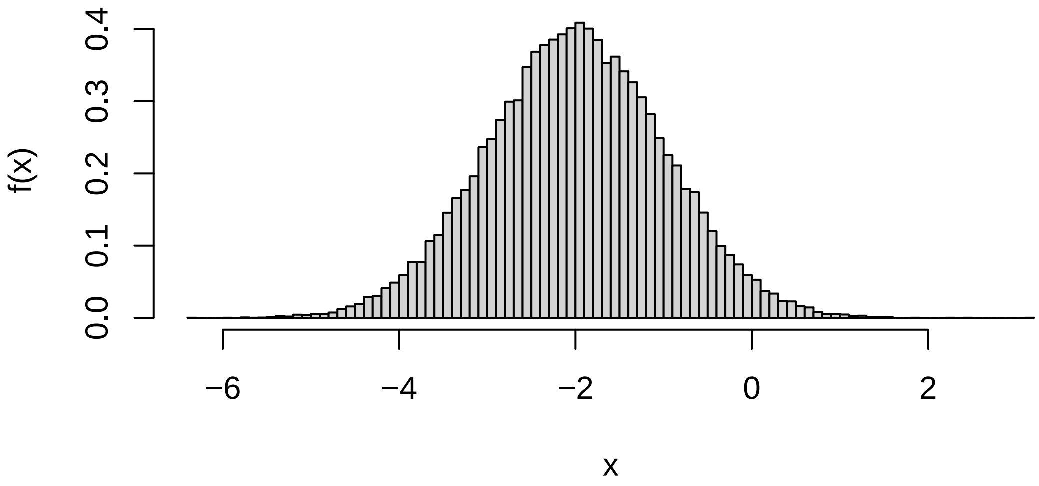
3 ☐ $\hat{\sigma}_x^2 = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_x)^2}{n}} = 1.29$

4 ☐ $\hat{\sigma}_x^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_x)^2}{n-1} = 1.89$

5 ☐ The sample size is too small to calculate the variance estimate.

Question XI.2 (30)

Consider the histogram of another sample:



Which of the following distributions would most likely fit population from which the sample was taken?

1 ☐ The exponential distribution

2 ☐ The log-normal distribution

3 ☐ The uniform distribution

4 ☐ The standard normal distribution

5 ☐ The normal distribution

The exam is finished. Enjoy the summer!