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

Written examination: 17. May 2020

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 10 exercises. To answer the questions, you need to fill in the "multiple choice" form (6 separate pages) on CampusNet with the numbers of the answers that you believe to be correct.

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

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

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

Exercise	IV.4	IV.5	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	1	2	1	5	5	3	2	4	3

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

The exam paper contains 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 characteristics of electrical components are not exactly as specified, e.g. if you buy a resistor then the resistance through it is not exactly as specified. In the production of electric circuits, it is of great interest not to get too much variation in the quality of the overall circuit. An example is the resistance through two parallel connected resistors, which is calculated by

$$R = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2}}$$

where R_1 is the resistance through one of the resistors and R_2 through the other resistor. The resistance is measured in ohm.

Assume that $R_1 \sim N(4, 0.2)$ and $R_2 \sim N(2, 0.2)$.

Question I.1 (1)

You buy 100 R_1 resistors - which can be assumed to be independent of each other. What is the probability that none of these has a resistance below 3 ohms?

1 ☐ 1.27%

2 ☐ 2.78%

3 ☐ 13.9%

4* ☐ 27.9%

5 ☐ 42.4%

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

First calculate the probability that a single of them is not below 3:

```
pnorm(3, 4, sqrt(0.2))
```

```
## [1] 0.01267366
```

and now we have 100 independent draws (we draw them from an "infinite" population of resistors), hence we use the binomial distribution calculating zero "successes"

```r
dbinom(0, 100, pnorm(3, 4, sqrt(0.2)))
```

```
## [1] 0.2793009
```

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

Question I.2 (2)

Calculate an estimate of the standard deviation of the total resistance R (the answer is rounded to two significant digits, tip: if you use simulation then remember to make sufficient repetitions to get a stable result)?

1 ☐ 0.026

2 ☐ 0.094

3 ☐ 0.16

4* ☐ 0.21

5 ☐ 0.44

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

Its a non-linear function, so use simulation

```r
  k <- 1000000
  R1 <- rnorm(k, mean=4, sd=sqrt(0.2))
  R2 <- rnorm(k, mean=2, sd=sqrt(0.2))
  R <- 1 / (1/R1 + 1/R2)
  sd(R)
```

```
## [1] 0.2093243
```

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

Continue on page 4

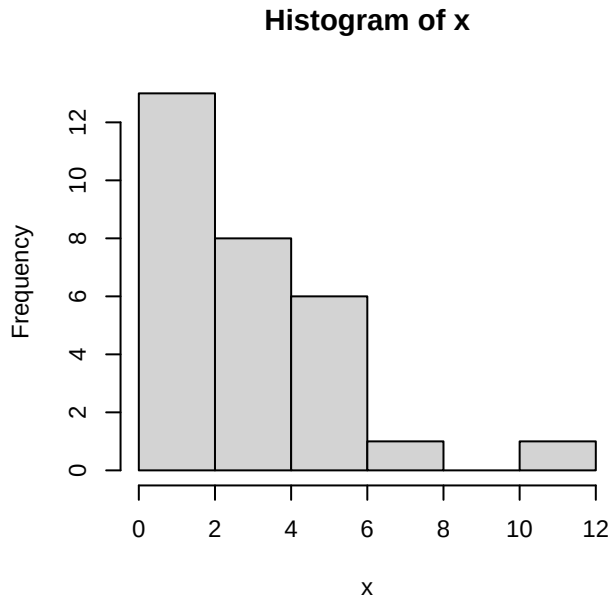
Exercise II

In a computer system an optimization routine is used and the execution time for this routine is under investigation. The execution time is measured in hours and loaded into R with the following code:

```
x <- c(1.6, 2, 3.4, 4, 2.1, 0.6, 0.4, 0.4, 6, 0.4, 4.9, 2, 2, 4.6, 0.5,
       3.4, 7.2, 10.5, 3.2, 1.3, 5.7, 1.9, 2.6, 2.5, 4.4, 1.8, 3.9, 6, 0.9)
```

Question II.1 (3)

It is desirable to assess which distribution the outcomes in the sample could originate from. Therefore the following histogram has been generated of the observations in `x`:



On the basis of the information given, evaluate which of the following distributions is most likely to have generated the observations in the sample?

1 ☐ A normal distribution

2 ☐ A Poisson distribution

3* ☐ An exponentiel distribution

4 ☐ A t-distribution

5 ☐ A binomial-distribution

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

Lets go through the answers:

- With most values in the interval closes to zero, but no values below, then it is very unlikely that it is from a normal distribution.
- It's not Poisson, since it is not integer values.
- It looks very must like an exponential distribution.
- Like the normal, it's very unlikely that it is a t-distribution.
- It's not Binomial, since it is not integer values.

Hence the only likely answer is the exponential distribution.

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

Question II.2 (4)

Based on the sample, what is the estimate of the mean and standard deviation of the computation times?

1 ☐ $\hat{\mu} = 2.53$ and $\hat{\sigma} = 1.66$

2 ☐ $\hat{\mu} = 3.36$ and $\hat{\sigma} = 0.48$

3* ☐ $\hat{\mu} = 3.11$ and $\hat{\sigma} = 2.37$

4 ☐ $\hat{\mu} = 1.98$ and $\hat{\sigma} = 5.63$

5 ☐ $\hat{\mu} = 3.96$ and $\hat{\sigma} = 2.81$

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

Copy from the pdf to read the sample into R:

```
x <- c(1.6, 2, 3.4, 4, 2.1, 0.6, 0.4, 0.4, 6, 0.4, 4.9, 2, 2, 4.6, 0.5,
3.4, 7.2, 10.5, 3.2, 1.3, 5.7, 1.9, 2.6, 2.5, 4.4, 1.8, 3.9, 6, 0.9)
```

and then

```
mean(x)

## [1] 3.11

sd(x)

## [1] 2.37
```

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

Question II.3 (5)

You want to give a guarantee that the execution time is below a certain level, and therefore a confidence interval of the 90% quantile should be calculated. A function is defined in R to calculate it by:

```
q90 <- function(x){ quantile(x, prob=0.9, type=2) }
```

Which of the following R codes calculates a 95% percent non-parametric bootstrap confidence interval for the 90% quantile of the distribution of computation times?

1 ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.05, 0.95))
```

2 ☐
```
simsamples <- replicate(10000, sample(x, replace = FALSE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.025, 0.975))
```

3 ☐
```
simsamples <- replicate(10000, sample(x, replace = FALSE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.05, 0.95))
```

4 ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.1, 0.90))
```

5* ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.025, 0.975))
```

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

The differences in the answers are:

- replace: must be TRUE
- quantiles calculated: they must be 2.5% and 97.5% to have the right significance level with 95% percent in between

So

```
simsamples <- replicate(10000, sample(x, replace = TRUE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.025, 0.975))

##  2.5% 97.5%
##   4.6  10.5
```

See Section 4.3.

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

Continue on page 8

Exercise III

An NGO has 15 callers employed to recruit new members. Let X represent the number of members a single caller recruits during one working day. The number of new members each caller recruits in a day can be assumed to be independent of each other. From experience it is known that a good model for X is a binomial distribution, where the probability of getting a new member in a call is 7%. It is assumed that each caller can do 120 calls in one day.

Question III.1 (6)

What is the probability that a caller on a single day recruits more than 5 new members?

1 ☐ 0.12

2* ☐ 0.85

3 ☐ 0.45

4 ☐ 0.17

5 ☐ 0.96

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

It's a binomial setup, so discrete, and the probability is

$$P(X > 5) = 1 - P(X \leq 5)$$

which is found in R by

```
n <- 120
p <- 0.07
1 - pbinom(5, n, p)

## [1] 0.8522782
```

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

Question III.2 (7)

If Y is the total number of new members the 15 callers can recruit in a day, what is the mean and variance of Y?

1 ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 10.8$

2 ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 41.9$

3 ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 43.5$

4 ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 102.4$

5* ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 117.2$

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

The mean and variance is given for the binomial distribution in Theorem 2.21. Hence $\mu_X = np = 8.4$ and $\sigma_X^2 = np(1-p) = 7.812$.

We have the total recruitment of the 15 callers by the sum

$$Y = \sum_{1}^{15} X$$

Applying the rules in Theorem 2.56, we get $\mathrm{E}(Y) = \sum_1^{15} 8.4 = 126$ and $\mathrm{V}(Y) = \sum_1^{15} 7.812 = 117.18$.

The trick is <u>not</u> to make

$$Y = 15X$$

which it would be only the recruitment of a single caller times 15, which would give a wrong variance $\mathrm{V}(Y) = \mathrm{V}(15X) = 15^2 \cdot 7.812 = 1757.7$! Well, it's not among the answers, to that made it less of a pitfall.

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

Continue on page 10

Exercise IV

In Denmark, people often discuss whether it has been a good or a bad summer. The table below shows the average temperatures for the months May to September during the years 2014-2018:

	2014	2015	2016	2017	2018	Average
May	11.7	9.7	12.9	12.0	15.0	12.26
June	14.9	12.7	16.0	14.7	16.5	14.96
July	19.5	15.5	16.4	15.5	19.2	17.22
August	16.0	17.4	16.1	16.0	17.5	16.60
September	14.6	13.2	16.2	13.3	14.1	14.28
Average	15.34	13.70	15.52	14.30	16.46	15.01

To investigate if there is a difference between the years, the following R code has been run, where `month` is the indicator for the months May to September, `year` for the years 2014 to 2018 and `temp` is the average temperature:

```
anova(fit <- lm(temp ~ year + month))

## Analysis of Variance Table
##
## Response: temp
##           Df Sum Sq Mean Sq F value  Pr(>F)
## year       4   23.4    5.85    3.87   0.022 *
## month      4   77.5   19.37   12.82 7.3e-05 ***
## Residuals 16   24.2    1.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question IV.1 (8)

At significance level $\alpha = 0.05$ what is the conclusion from the R code above (by difference is meant significant difference in mean. Remember all parts of the answer must be correct)?

1 ☐ There is neither a difference in temperature between months and years, since $7.3 \cdot 10^{-5} < 0.05$ and $0.022 < 0.05$.

2 ☐ There is both a difference in temperature between months and years, since $2 \cdot 7.3 \cdot 10^{-5} < 0.05$ and $2 \cdot 0.022 < 0.05$.

3 ☐ There is a difference in temperature between months, since $7.3 \cdot 10^{-5} < 0.05$, however there is not a difference between years, since $0.022 > 7.3 \cdot 10^{-5}$.

4* ☐ There is both a difference in temperature between months and years, since $7.3 \cdot 10^{-5} < 0.05$ and $0.022 < 0.05$.

5 ☐ There is neither a difference in temperature between months and years, since $2 \cdot 7.3 \cdot 10^{-5} < 0.05$ and $2 \cdot 0.022 < 0.05$.

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

The effect of both factors is significant, since both p-values for the F-test in Theorem 8.22 are below the significance level - they are under `Pr(>F)`.

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

Question IV.2 (9)

The model used for the test above can be written as

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

What is the estimate of σ^2?

1 ☐ $\hat{\sigma}^2 = (23.4 + 77.5 + 24.2)/(4 + 4 + 16)$

2* ☐ $\hat{\sigma}^2 = 24.2/16$

3 ☐ $\hat{\sigma}^2 = \sqrt{24.2}$

4 ☐ $\hat{\sigma}^2 = (5.85 + 19.37 + 1.51)/(4 + 4 + 16)$

5 ☐ $\hat{\sigma}^2 = \sqrt{1.51}$

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

It's the value listed under `Mean Sq`. It is $MSE = \frac{SSE}{(k-1)(l-1)}$, see the table on the page after Theorem 8.22.

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

Question IV.3 (10)

What is

$$\sum_{\text{all } i,j} (y_{ij} - \bar{y})^2$$

where y_{ij} is the individually observed temperatures and $\bar{y}$ is the average af all the observed temperatures?

1 ☐ 77.4

2 ☐ 26.7

3 ☐ 24.2

4 ☐ 16.7

5* ☐ 125.1

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

You have to recognize this formula as the total variance: SST. It's given in Theorem 8.20 and it is calculated by summing all the variances in the table:

```
23.4+77.5+24.2
```

```
## [1] 125
```

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

Question IV.4 (11)

In order to make pairwise comparisons of the average temperature in the individual years the Bonferonni corrected Least Significant Difference (LSD) can be used. With a significance level of $\alpha = 0.05$, what is the value of it in the present case?

1 ☐ 1.27

2 ☐ 2.27

3 ☐ 1.61

4 ☐ 1.36

5* ☐ 2.53

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

We have to look in Section 8.3.3. There are five years, so in total $M = 5 \cdot 4/2 = 10$ comparisons can be made, and there are 5 observations for each year (m in Remark 8.13):

```
qt(1-0.025/10, df=16) * sqrt(2*1.51*1/5)
```

```
## [1] 2.53
```

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

Question IV.5 (12)

What would the p-value be if one had chosen to test whether there is a difference between years in a one-way analysis (i.e. ignoring the difference between months)?

1* ☐ 0.362

2 ☐ 0.638

3 ☐ 0.499

4 ☐ 0.463

5 ☐ 0.537

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

Now we really have to do it "manually". We have to find the values to calculate F in Theorem 8.6.

Further, we have to realize that the variances are calculated the same way for each factor, no matter if it is one or two way, so with month is out of the model we get:

```
SSE <- 77.5 + 24.2
SSTr <- 23.4

F <- (SSTr/(5-1)) / (SSE/(25-5))
```

and with that we can calculate the p-value by:

```
1 - pf(F, df1=4, df2=20)

## [1] 0.362
```

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

Continue on page 14

Exercise V

In a research project on energy consumption in schools, it was investigated how students and teachers set the radiator thermostats in classrooms. Thermostat settings were recorded during a period with cold weather at a number of schools in randomly selected classrooms. It was predetermined that thermostats are well set if the setting is between 2 and 3 on all radiators in a room, as it otherwise indicates under or oversized radiators. Besides, it's not desirable for thermostats to be set differently in a room, as this results in inferior comfort and poorer return water cooling.

The following observations were made during the period:

	School 1	School 2	School 3	School 4
Not-well set	18	11	22	9
Well set	38	36	15	12

Hence, at School 2 there were 11 out of 47 rooms in which the thermostats were not well set.

Question V.1 (13)

What is the 95% confidence interval for the proportion of thermostats that were not set well at School 1 (note that the result from the R functions and the books formula may be slightly different, but both results are always closest to the correct answer)?

1 ☐ $[0.07, 0.17]$

2* ☐ $[0.20, 0.44]$

3 ☐ $[0.26, 0.53]$

4 ☐ $[0.05, 0.22]$

5 ☐ $[0.18, 0.51]$

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

We use the CI formula in Method 7.3:

```
p <- 18/(18+38)

## With the formula of the book
p - qnorm(0.975) * sqrt(p*(1-p)/(18+38))

## [1] 0.1991

p + qnorm(0.975) * sqrt(p*(1-p)/(18+38))

## [1] 0.4437
```

Note that this gives a slightly different result than the R function, see Remark 7.8. It's still closest to the correct answer, since it gives:

```
## With the R function
prop.test(x = 18, n = 18+38, correct=FALSE)$conf.int

## [1] 0.2140 0.4518
## attr(,"conf.level")
## [1] 0.95
```

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

Question V.2 (14)

It was planned to compare schools to investigate if there were differences in practice of thermostat setting at the schools. Under the null hypothesis of no difference, then what is the expected number of not-well set thermostats at School 3?

1* ☐ $e_{13} = 37 \cdot \frac{60}{161} = 13.8$

2 ☐ $e_{13} = 15 \cdot \frac{22}{37} = 8.9$

3 ☐ $e_{13} = 60 \cdot \frac{124}{161} = 46.2$

4 ☐ $e_{13} = 22 \cdot \frac{22}{37} = 13.1$

5 ☐ $e_{13} = 15 \cdot \frac{124}{161} = 11.6$

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

We find the formula for expected counts under the null hypothesis that all proportions are equal in Section 7.4.

So we sum the row of Not-well set:

```
18+11+22+9

## [1] 60
```

and the total number is

```r
60 + 38 + 36 + 15 + 12
```

```
## [1] 161
```

so we have 37 counts at School 3,

```r
37 * 60/161
```

```
## [1] 13.79
```

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

Question V.3 (15)

You want to investigate whether there was a difference in the proportion of not-well set thermostats at the four schools. Use a significance level of 1%. What will be the conclusion (both conclusion and argument must be correct)?

1 ☐ A difference between the schools cannot be detected, since the relevant test statistic is below the critical level of 15.4.

2 ☐ A difference between the schools can be detected, since the relevant test statistic is below the critical level of 11.3.

3 ☐ A difference between the schools cannot be detected, since the relevant test statistic is above the critical level of 0.0057.

4 ☐ A difference between the schools cannot be detected, since the relevant test statistic is below the critical level of 0.0057.

5* ☐ A difference between the schools can be detected, since the relevant test statistic is above the critical level of 11.3.

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

If we don't want to do all the calculations, then we can put in R (we could have done that already) by:

```r
M <- as.table(rbind(c(18,11,22,9),c(38,36,15,12)))
M
```

```
##    A  B  C  D
## A 18 11 22  9
## B 38 36 15 12
```

```r
## Chi^2 test
(Xsq <- chisq.test(M))

##
##  Pearson's Chi-squared test
##
## data:  M
## X-squared = 13, df = 3, p-value = 0.006

## Observed statistic
Xsq$statistic

## X-squared
##     12.57

## p-value
1 - pchisq(Xsq$statistic, df=ncol(M)-1)

## X-squared
##  0.005671

## Critical value
qchisq(0.99, df=length(x)-1)

## [1] 11.34
```

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

Continue on page 18

Exercise VI

The following 3 questions deal with various statistical problems that may arise when treating water.

Question VI.1 (16)

When controlling drinking water the quality is measured by regular water analysis. There is, of course, legislation on quality among other things requirements for concentrations of different substances. One requirement is that the conductivity of the water at the consumer's tap must not be more than 2500 μS/cm at 20 °C. If the conductivity is above this level, the concentration of salts is too high and the water is referred to as aggressive.

The conductivity of water at randomly selected consumers' taps has been measured. Let a measurement be represented by the stochastic variable X_i, which can be assumed to be normally distributed. Twenty independent measurements have been collected to determine the conductivity and the observed values are stored in the vector `x` in R. You now want to test if the water on average is aggressive, and the following null hypothesis about the mean value, μ, of the conductivity in the drinking water at the consumer's tap, is formulated

$$H_0 : \mu = 2500$$

with the alternative hypothesis

$$H_1 : \mu \neq 2500$$

The following from R is given:

```
t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 4.8527, df = 19, p-value = 0.0001106
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   704.1022 1772.1090
## sample estimates:
## mean of x
##  1238.106
```

Based on the result above and with a 5% significance level, which of the following statements is correct (both conclusion and reasoning must be correct)?

1 ☐ We reject H_0 as the relevant confidence interval does not contain 0.

2 ☐ We accept H_0, since the sample average is within the relevant confidence interval.

3 ☐ We accept H_0, since the test statistic is greater than 1.96.

4 ☐ We reject H_0, since the p-value is 0.0001106.

5* ☐ We reject H_0, since 2500 is not within the relevant confidence interval.

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

From the R-output we see that the 95% confidence interval is 704-1772, i.e. does not contain 2500. Thus we reject the null-hypothesis, since the 2500 is outside the acceptance region (=the confidence interval). Answer 4 is not correct because the p-value of the `t.test` result above was calculated for a different Null hypothesis, i.e. $H_0 : \mu = 0$.

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

Question VI.2 (17)

At waterworks the water is purified by one of two methods, A or B, and the remaining concentration of a substance is measured. Measurements of remaining concentration are given in the stochastic variables $Y_{A,i}$ and $Y_{B,i}$ for methods A and B, respectively. It is of interest to investigate which of the two methods best purifies the water. $Y_{A,i}$ and $Y_{B,i}$ can be assumed to be normally distributed and their variances, σ_A^2 and σ_B^2, can be assumed to be equal. For each of the two methods a sample of 20 observations is taken. We want to test the null hypothesis

$$H_0 : \quad \mu_A = \mu_B$$

where the alternative hypothesis is

$$H_1 : \quad \mu_A \neq \mu_B$$

Which of the following procedures is a correct approach?

1 ☐ A paired t-test with 19 degrees of freedom.

2 ☐ A paired t-test with 18 degrees of freedom.

3* ☐ A two-sample t-test with 38 degrees of freedom.

4 ☐ A two-sample t-test with 39 degrees of freedom.

5 ☐ A F-test testing variance homogeneity.

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

We use Welch's t-tests where the number of degrees of freedom is given by (3.50):

$$\nu = (\sigma_A^2/20 + \sigma_B^2/20)^2)/((\sigma_A^2/20)^2/(20-1) + (\sigma_B^2/20)^2/(20-1))$$

since $\sigma_A = \sigma_B$ we have

$$\nu = (2\sigma_A^2/20)^2/(2(\sigma_A^2/20)^2/(20-1)) = 2(20-1) = 38$$

Alternatively, we could have used the pooled t-test because the variances of both samples are similar (3.53.

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

Question VI.3 (18)

A new study is planned on the concentration of a substance in a drinking water drilling. We want to achieve a power of 90% to detect a mean value difference of 2 units from a given value. From experience it is known that the standard deviation is 3.5 units and you want to perform the test at a 5% significance level (remember that results from R functions may differ slightly from the result obtained with the book's formulas). How many observations must be taken to fulfil these requirements?

1 ☐ At least 15 observations.

2* ☐ At least 35 observations.

3 ☐ At least 48 observations.

4 ☐ At least 67 observations.

5 ☐ At least 102 observations.

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

We can either use the formula in Method 3.65 or the function in R:

```
## The formula
zb <- qnorm(0.9)
za <- qnorm(1-0.05/2)
delta <- 2
sigma <- 3.5

(sigma * (zb+za)/delta)^2
```

```
## [1] 32.17898

## The R function
power.t.test(power=0.9, delta=2, sd=3.5, sig.level=0.05, type="one.sample")

##
##      One-sample t test power calculation
##
##               n = 34.15781
##           delta = 2
##              sd = 3.5
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
```

The results of the formula and from the R output differ slightly due to rounding. However, both indicate that only answer 2 can be correct.

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

Continue on page 22

Exercise VII

On 30 randomly selected summer days, corresponding values of the temperature at noon, x measured in degrees Celsius, and the number of ice creams sold in an ice cream chain, Y, have been recorded. The following model has been fitted in R:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{where } \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.}$$

The result of this is shown below:

```
summary(fit1)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -161.24  -81.60  -46.14  103.83  249.85
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2382.001    116.620  -20.43   <2e-16 ***
## x             230.703      5.083   45.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 124.7 on 28 degrees of freedom
## Multiple R-squared:  0.9866,Adjusted R-squared:  0.9861
## F-statistic:  2060 on 1 and 28 DF,  p-value: < 2.2e-16
```

Question VII.1 (19)

Based on the above R output, what is the estimate of the variance of the errors $\hat{\sigma}^2$?

1 ☐ 116.6

2 ☐ 116.6^2

3 ☐ 124.7

4* ☐ 124.7^2

5 ☐ $(230.7/28)^2$

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

$\hat{\sigma}$ is seen directly from "Residual standard error". Thus the variance is obtained by squaring this value.

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

Question VII.2 (20)

Based on the above R output, what is the prediction of the mean value of ice creams sold, $\hat{y}_{\text{new}}$, at $x_{\text{new}} = 25°$ C?

1 ☐ 231 ice creams.

2 ☐ 2382 ice creams.

3* ☐ 3386 ice creams.

4 ☐ 5768 ice creams.

5 ☐ 11535 ice creams.

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

```
-2382 + 230.7 * 25
```

```
## [1] 3385.5
```

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

Question VII.3 (21)

Based on the above R output, what is the critical value for the test

$$H_0 : \beta_1 = 0$$

using a 1% significance level?

1* ☐ 2.76

2 ☐ 2.05

3 ☐ 1.96

4 ☐ 1.70

5 ☐ 2.56

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

We know that the standardized parameters follow a t-distribution under H_0 from Theorem 5.12. Hence, as in Example 5.13, we find the critical values of the test by:

```
qt(0.995, df=28)

## [1] 2.763262
```

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

Question VII.4 (22)

Which of the following statements about a prediction interval for $Y_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} + \varepsilon_{\text{new}}$ is not correct?

1 ☐ The prediction interval is wider than a corresponding confidence interval.

2 ☐ The width of the prediction interval depends on the sample size.

3 ☐ The prediction interval is symmetrical around the predicted value.

4 ☐ The width of the prediction interval depends on the value of x_{new}.

5* ☐ If the sample size becomes large enough, then the width of the prediction interval becomes 0.

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

The prediction interval is an interval for a new value, thus the width will not be smaller than $2 \cdot t_{1-\alpha/2}\sigma$, where σ represents the standard error of the predicted value $\hat{y}_{new}$ (5.18).

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

Question VII.5 (23)

It is suspected that the linear model is not a correct model, so now you instead fit the model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$. The model is fitted by (note that `x2` is `x` squared):

```r
x2 <- x^2
fit2 <- lm(y ~ x + x2)
summary(fit2)

##
## Call:
## lm(formula = y ~ x + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -193.67  -59.32  -25.73   64.96  263.87
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -159.3055   470.1439  -0.339    0.737
## x             24.9850    42.9277   0.582    0.565
## x2             4.5715     0.9502   4.811 5.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 93.15 on 27 degrees of freedom
## Multiple R-squared:  0.9928,Adjusted R-squared:  0.9922
## F-statistic:  1856 on 2 and 27 DF,  p-value: < 2.2e-16
```

Based on this R output and with a 1% significance level, what can now be concluded about the relationship between ice cream sales and temperature (both conclusion and argument must both be correct)?

1 ☐ The relationship differs statistically significantly from a linear relationship, since $\hat{\beta}_2$ is positive.

2 ☐ The relationship does not differ statistically significantly from a linear relationship, since the p-value for $\hat{\beta}_1$ is above 1%.

3* ☐ The relationship differs statistically significantly from a linear relationship, since the p-value for $\hat{\beta}_2$ is below 1%.

4 ☐ We cannot reject that the relationship is linear, since $\hat{\beta}_2 < \hat{\beta}_1$.

5 ☐ We cannot reject that the relationship is linear, since $R^2 \approx 1$.

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

From the output we see that the p-value corresponding to the second order term is below 1%.

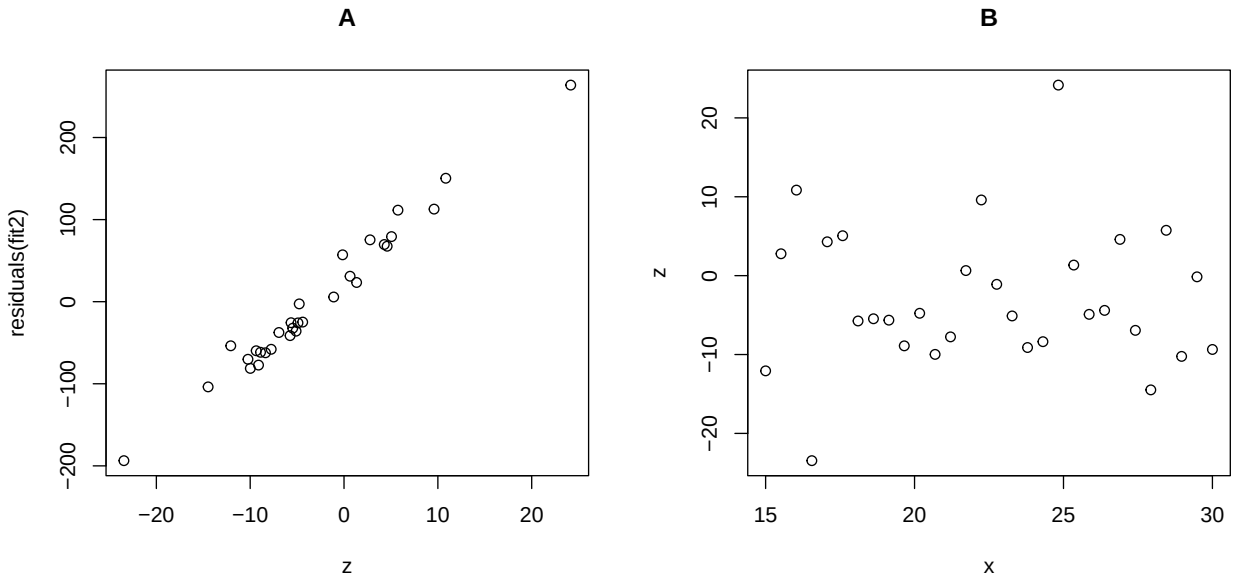
---------- FACIT-END ----------

Question VII.6 (24)

In this exercise the result from the fit of the model from the previous question is investigated.

Below are shown two plots (A and B), where:

- z is a new variable, which represent measurements of sunshine on a day (the unit doesn't matter).
- x is the temperature on a day.
- `residuals(fit2)` is the residuals from the fit.



Which of the following statements is correct?

1 ☐ Figure B is used to examine whether the assumption of variance homogeneity is met.

2 ☐ Figure A is used to investigate whether the normal distribution assumption is met.

3 ☐ Figure B indicates a strong relationship between z og x.

4* ☐ Figure A can be used to investigate whether z should be included in the model.

5 ☐ Figure B can be used to check whether the relationship between x and Y is modelled correctly.

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

In the residual plot (A) we see a clear pattern. Because there is a clear relationship between z and the residuals, it indicates that z should be included in the model (5.28).

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

Continue on page 28

Exercise VIII

The number of shooting stars per hour, X, is given by $X \sim Po(3)$, i.e. Poisson distributed with mean 3 shooting stars per hour.

Question VIII.1 (25)

If one counts shooting stars for four hours, how many shooting stars might one expect to see (remember that the expectation value is equal to the mean)?

1 ☐ 8

2 ☐ 9

3* ☐ 12

4 ☐ 16

5 ☐ 24

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

We scale λ by 4 and obtain: $\lambda_4 = 4 \cdot 3 = 12$

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

Question VIII.2 (26)

Suppose one has just observed a shooting star. What is then the probability of waiting more than 10 minutes for the next shooting star?

1 ☐ $P(X = 0)$, $X \sim Po(3)$

2 ☐ $P(X > 0)$, $X \sim Po(3)$

3 ☐ $P(Y > 10)$, $Y \sim Exp(3)$

4* ☐ $P(Y > \frac{10}{60})$, $Y \sim Exp(3)$

5 ☐ $P(Y > \frac{10}{6})$, $Y \sim Exp(3)$

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

Here we use the connection between Poisson process and the exponential distribution, i.e. the waiting times between poisson events are exponentially distributed. If Y is the waiting time

between two events, we need to find the probability $P(Y > 10 \text{ min}) = P(Y > 10/60 \text{ hours})$, where $Y \sim Exp(3)$.

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

Continue on page 30

Exercise IX

In a production of steel pipes one is interested in the diameter of the pipes. Therefore, a sample of 30 tubes is taken and the sample variance is calculated to be $s^2 = 531$ mm^2.

Question IX.1 (27)

What is the 99% confidence interval for the standard deviation of the diameter of the tubes?

1 ☐ [18.4, 31.0]

2 ☐ [18.1, 30.6]

3 ☐ [310, 1079]

4* ☐ [17.2, 34.3]

5 ☐ [294, 1175]

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

We use Method 3.19 to calculate the confidence interval. First the $1 - \alpha/2$ and $\alpha/2$ quantiles of the χ^2-distribution with 29 degrees of freedom are found in R:

```
qchisq(0.995, df=29)

## [1] 52.3

qchisq(0.005, df=29)

## [1] 13.1
```

$$\left[\sqrt{\frac{29 \cdot 531}{52.3}}, \sqrt{\frac{29 \cdot 531}{13.1}}\right] = [17.2, 34.3]$$

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

Question IX.2 (28)

The diameter of the steel pipes is usually measured by one of two different measurement methods. It is now suspected that the two methods don't measure identically. 11 pipes are therefore randomly selected. Each pipe is now measured by both methods. The observation made with measurement Method 1 on pipe i is at position i in the vector `x` and corresponding measurement with Method 2 at position i in the vector `y`. The measurements by both methods can be assumed to be normally distributed.

The following analyses are now carried out in R:

```
t.test(x-y)

##
##  One Sample t-test
##
## data:  x - y
## t = -2.541, df = 10, p-value = 0.0293
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2.409246 -0.158027
## sample estimates:
## mean of x
##  -1.28364

t.test(x,y)

##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -0.1353, df = 20, p-value = 0.894
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.0683  18.5010
## sample estimates:
## mean of x mean of y
##   96.9018   98.1855
```

You want to test the null hypothesis that the two measurement methods have the same mean

$$H_0 : \mu_X = \mu_Y$$

against the alternative hypothesis that they are different

$$H_1 : \mu_X \neq \mu_Y$$

At a 5% significance level, which of the following possibilities is correct (both conclusion and argument must be correct)?

1* ☐ We reject H_0, since $p < 0.05$.

2 ☐ We cannot reject H_0, since $p = 0.89$.

3 ☐ We reject H_0, since the alternative hypothesis is different from 0.

4 ☐ We cannot reject H_0, since the test statistic, t_{obs}, is -0.14.

5 ☐ We reject H_0, since the difference between the sample averages is greater than 1.

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

It is a paired setup, and the measurements for each pipe is located at the same position in both vectors. Thus we use the output from the first call.

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

Question IX.3 (29)

The company that produces the steel pipes is getting a new and faster machine for producing steel pipes. Regardless of the answer in the previous question, only measurement Method 1 is used below. The desired diameter of the steel pipes is 100 mm and the machine will be properly calibrated. You will plan a new experiment where you want to test the null hypothesis

$$H_0 : \mu_X = 100$$

against the alternative hypothesis

$$H_1 : \mu_X \neq 100$$

We use $\hat{\sigma}_x^2 = 502.23$, a significance level of 5% and the following R-code has been executed, since a sample size of $n = 40$ is wanted:

```
power.t.test(n=40, sd=sqrt(502.23), power=0.9, type="one.sample")

##
##      One-sample t test power calculation
##
##               n = 40
##           delta = 11.8
##              sd = 22.4
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
```

Based on the R code above, what can now be concluded before the experiment is performed (both argument and conclusion must be correct)?

1 ☐ If the true mean is 88.2 or lower, we have more than 90% chance of rejecting H_1.

2* ☐ If the true mean is 88.2 or lower, we have more than 90% chance of rejecting H_0.

3 ☐ If the true mean is 88.2 or lower, we will reject H_0.

4 ☐ If the true mean differs with less than 11.8 we accept H_0.

5 ☐ If the true mean differs with less than 11.8 we reject H_0.

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

Delta is 11.8. To obtain a power of at least 90% we need to be at least as far away from the hypotesized mean as given by delta. This corresponds to detecting a true mean of 88.2 or below or 111.8 and above. Therefore only answer 2 can be correct.

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

Continue on page 34

Exercise X

In the production of a particular type of plate, each plate has a 20% probability of having a flaw. A random sample of 10 plates is now taken.

Question X.1 (30)

What is the probability that at most 3 plates in the sample have a flaw?

1 ☐ 0.95

2 ☐ 0.32

3 ☐ 0.68

4* ☐ 0.88

5 ☐ 0.60

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

$X \sim B(10, 0.2)$. The Probability in question is $P(X \leq 3)$. It can be found using the following R code:

```
pbinom(3,size=10,prob=0.2)

## [1] 0.8791261
```

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

The exam is finished. Have a great summer!