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

Written examination: 3. June 2024

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 14 exercises. To answer the questions, you need to fill in the "multiple choice" form on exam.dtu.dk.

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

Exercise	VI.1	VI.2	VII.1	VIII.1	VIII.2	VIII.3	IX.1	IX.2	X.1	XI.1
Question	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Answer	4	4	4	3	1	4	4	5	2	3

Exercise	XI.2	XI.3	XII.1	XII.2	XIII.1	XIII.2	XIII.3	XIV.1	XIV.2	XIV.3
Question	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Answer	3	3	2	4	3	4	2	4	5	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

Consider the following hypothesis test:

$H_0 : \mu = 18$
$H_1 : \mu \neq 18$

A sample with $n = 48$ observations provided a sample mean $\bar{x} = 17$ and a sample standard deviation $s = 4.5$. It is assumed that the sample is collected from a normal distributed population.

Question I.1 (1)

What is the value of the test statistic (t_{obs}) and what are the critical values with significance level $\alpha = 0.05$ (both answers must be correct)?

1* ☐ $t_{\text{obs}} = -1.54$. $t_{0.025} = -2.012$ and $t_{0.975} = 2.012$

2 ☐ $t_{\text{obs}} = -2.01$. $t_{0.025} = -2.685$ and $t_{0.975} = 2.685$

3 ☐ $t_{\text{obs}} = -1.96$. $t_{0.025} = -1.678$ and $t_{0.975} = 1.678$

4 ☐ $t_{\text{obs}} = -1.89$. $t_{0.025} = -2.012$ and $t_{0.975} = 2.012$

5 ☐ Cannot be calculated without more information.

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

```r
x_bar = 17
mu = 18
s = 4.5
n = 48
tobs <-(x_bar-mu)/(s/sqrt(n))
tobs

## [1] -1.539601
```

With df=47, $t_{0.025}$ and $t_{0.975}$ can be looked up by:

```r
qt(c(.025, 0.975), df=47)

## [1] -2.011741  2.011741
```

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

Question I.2 (2)

From another sample of $n = 45$ observations, the value of the test statistic (t_{obs}) is -1.74. Compute the p-value and draw a conclusion using significance level $\alpha = 0.05$ (both must be correct).

1* ☐ The p-value is 0.0889. We do not reject the null hypothesis.

2 ☐ The p-value is 0.0805. We reject the null hypothesis.

3 ☐ The p-value is 0.0560. We do not reject the null hypothesis.

4 ☐ The p-value is 0.1339. We do not reject the null hypothesis.

5 ☐ We cannot draw a conclusion.

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

```r
n = 45
tobs <- -1.74
tobs

## [1] -1.74

p_value = (1 - pt(abs(tobs), df = n-1))*2
round(p_value,4)

## [1] 0.0889
```

Rejection Rule: Reject H_0 if p-value $\leq \alpha$
$\alpha = 0.05$
Conclusion: Do not reject the null hypothesis.

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

Continue on page 4

Exercise II

A random sample of $n = 30$ observations is collected and the sample mean is estimated to be $\bar{x} = 1.01$ and the sample standard deviation $s = 0.09$.

Question II.1 (3)

What is the 95% confidence interval for the standard deviation?

1 ☐ [0.06, 0.13]

2 ☐ [0.24, 0.40]

3 ☐ [0.09, 0.13]

4 ☐ [0.05, 0.09]

5* ☐ [0.072, 0.121]

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

See Method 3.19. The confidence interval for the standard deviation is:

```
n=30
s=0.09
alpha = 0.05
numerator = (n-1)*s^2
interval <- c(sqrt(numerator/qchisq((1-alpha/2),df=n-1)),
              sqrt(numerator/qchisq((alpha/2),df=n-1)))
round(interval,3)

## [1] 0.072 0.121
```

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

Continue on page 5

Exercise III

The outcome of an experiment is described by the random variable X, where X has the following density function:

x	0	1	2	3	4
$f(x)$	0.17	0.22	0.28	0	0.33

Question III.1 (4)

What is the probability $P(X < 3)$

1 ☐ This cannot be deduced from the information provided

2 ☐ 0.28

3 ☐ 0.33

4 ☐ 0.40

5* ☐ 0.67

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

$P(X < 3) = 0.17 + 0.22 + 0.28 = 0.67$.

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

Question III.2 (5)

The mean of X is 2.10. What is the variance of X?

1 ☐ 1.49

2 ☐ 2.10

3* ☐ 2.21

4 ☐ 4

5 ☐ 6.62

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

```r
# Mean
x <- c(0:2,4)
f <- c(0.17,0.22,0.28,0.33)
sum(x*f)

## [1] 2.1

# Variance
sum((x-2.1)^2*f)

## [1] 2.21
```

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

Continue on page 7

Exercise IV

A survey has been conducted to assess the concentration (we shall disregard the unit) of a chemical compound in the soil in four different locations. The measured concentrations from the survey are given in the below table:

Location 1	Location 2	Location 3	Location 4
253.7	261.1	257.9	244.1
241.2	244.2	263.5	244.9
255.8	250.5	258.6	243.9
249.3	264.9		247.1
	259.3		

The overall average concentration found across all four locations is 252.5 and the average concentrations for each location are listed below:

Location	1	2	3	4
Average concentration	250.0	256.0	260.0	245.0

We then fit a one-way ANOVA model on the form

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where the errors ε_{ij} are independent and follow normal distributions with mean 0 and variance σ^2. The overall mean is μ and the effect of Location i is α_i. The ANOVA table associated with the model shows that the total sum of squares, SST, is 915.92 and the location sum of squares, $SS(Location)$, is 480.00.

Question IV.1 (6)

What is the estimate of the effect of Location 1?

1* ☐ -2.5

2 ☐ 2.5

3 ☐ 6.0

4 ☐ 250.0

5 ☐ 252.5

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

The estimate of the effect of Location 1 is found as the difference between the group mean and the overall mean:

$$\hat{\alpha}_1 = 250.0 - 252.5 = -2.5.$$

See Example 8.1.

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

Question IV.2 (7)

What is the estimate of the error standard deviation?

1* ☐ $\hat{\sigma} = 6.027$

2 ☐ $\hat{\sigma} = 6.325$

3 ☐ $\hat{\sigma} = 7.814$

4 ☐ $\hat{\sigma} = 20.879$

5 ☐ $\hat{\sigma} = 36.327$

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

The estimate of the error variance is the mean square error, MSE, which is found as

$$MSE = \frac{SST - SS(Location)}{n - k} = \frac{915.92 - 480.00}{16 - 4} = 36.327.$$

Thus, the estimate of the error standard deviation is $\hat{\sigma} = \sqrt{36.327} = 6.027$.

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

Continue on page 9

Exercise V

12 observations of perfluorooctanesulfonic acid (PFOS) at six different concentrations were analysed using a new experimental testing method SPETT.

PFOS concentrations are measured in mg/kg. The data was read into R by:

```
# PFOS concentrations
x <- c(0, 0, 2, 2, 4, 4, 6, 6, 8, 8, 12, 12)
# SPETT values
y <- c(16, 116, 1170, 841, 2287, 2012, 2653, 3333, 4270, 3999, 5750, 5407)
```

The linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ where } \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d. for } i = 1, \ldots, 12,$$

was set up, where Y_i is the SPETT value and x_i the PFOS concentration of the i'th observation.

Note that in the remaining of the exercise, the normal distribution and i.i.d. assumptions of the errors are implicit (hence not written with the model).

Question V.1 (8)

What is the estimate of β_1?

1 ☐ 160.7

2* ☐ 467.6

3 ☐ 511.0

4 ☐ 1020.7

5 ☐ 259.0

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

Either use Theorem 5.4 of the book, or in R (first copy reading data from the pdf):

```
lm(y ~ x)

##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##       160.7        467.6
```

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

Question V.2 (9)

We wish to test the hypothesis $H_0 : \beta_0 = 0$, as this would indicate whether the expected SPETT value is zero for a PFOS concentration of zero. We use a significance level of $\alpha = 0.05$.

Which of the following statements is correct?

1 ☐ We do not reject the null hypothesis, since $|\hat{\beta}_0| > 1.96$, where 1.96 is the 95% quantile in a normal distribution.

2* ☐ We do not reject the null hypothesis, since the p-value is 0.23.

3 ☐ We do not reject the null hypothesis, since the p-value is 0.0027.

4 ☐ We reject the null hypothesis, since the p-value is 0.0027.

5 ☐ We reject the null hypothesis, since the p-value is 0.23.

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

We can use Theorem 5.12 or read the p-value directly from the lm-summary:

```
fit <- lm(y ~ x)
summary(fit)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -364.71 -172.27  -20.39  137.20  368.61
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   160.73     125.75   1.278     0.23
## x             467.58      18.96  24.666 2.74e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 259 on 10 degrees of freedom
## Multiple R-squared:  0.9838,Adjusted R-squared:  0.9822
## F-statistic: 608.4 on 1 and 10 DF,  p-value: 2.74e-10
```

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

Question V.3 (10)

Researchers would like to know the uncertainty of the SPETT value for a new observation with PFOS concentration 7 mg/kg. What is the 95% prediction interval at this concentration?

1* ☐ [2829, 4039]

2 ☐ [3253, 3615]

3 ☐ [487, 522]

4 ☐ [1898, 3099]

5 ☐ [2653, 4270]

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

Either use the formula in Method 5.18 or do it in R, see Example 5.20:

```
fit <- lm(y ~ x)
predict(fit, newdata=data.frame(x=7), interval="prediction", level=0.95)

##        fit      lwr      upr
## 1 3433.804 2829.042 4038.565
```

Note, that you have to give `newdata` as a `data.frame`.

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

Continue on page 12

Exercise VI

In a high quality paper factory the production occurs in batches. In a quality check experiment, batches were randomly selected, i.e. the quality is independent between the selected batches. In the first quality check 20 out of 85 batches were found to not live up to the quality requirement.

Question VI.1 (11)

What is the confidence interval calculated at a significance level $\alpha = 10\%$ for the proportion not living up to the quality requirements (note, use formula to get the correct answer)?

1 ☐ $[0.084, 0.387]$

2 ☐ $[0.128, 0.342]$

3 ☐ $[0.153, 0.342]$

4* ☐ $[0.160, 0.311]$

5 ☐ $[0.176, 0.294]$

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

Method 7.3: Use the formula:

```
n <- 85
x <- 20
phat <- x/n
phat + c(-1,1) * qnorm(0.95) * sqrt(phat*(1-phat)/n)

## [1] 0.1596160 0.3109723
```

The R function can also be used, however it gives a slightly different result:

```
prop.test(x=20, n=85, conf.level=0.9, correct=FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  20 out of 85, null probability 0.5
## X-squared = 23.824, df = 1, p-value = 1.056e-06
## alternative hypothesis: true p is not equal to 0.5
## 90 percent confidence interval:
##  0.1685119 0.3184077
## sample estimates:
##         p
## 0.2352941
```

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

Question VI.2 (12)

A new quality check is planned. The factory wants a 95% confidence interval with an expected width of 0.1. As a guess of the population proportion, the observed proportion from the first quality check is used (i.e. 20 out of 85 batches).

What is the minimum number of batches which must be taken for the new check to achieve this accuracy?

1 ☐ $n = 18$

2 ☐ $n = 70$

3 ☐ $n = 195$

4* ☐ $n = 277$

5 ☐ $n = 385$

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

We use the estimate $\hat{p} = \frac{20}{85}$ as guess of the population proportion and we set $ME = 0.05$, i.e. half the width of the wanted confidence interval precision. Applying the formula in Method 7.13:

```
p <- 20/85
ME <- 0.05
p*(1-p)*(qnorm(0.975)/ME)^2

## [1] 276.4787
```

and we round up.

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

Continue on page 14

Exercise VII

At a Danish company, a study was designed to investigate whether the online training modules for learning the concepts of sustainability improve trainees knowledge. The trainees participated in a quiz before and after completing the activities of the online training module. The two quizzes are referred to as pre-test and post-test. Six trainees' scores in the pre-test (denoted `pre`) and post-test (denoted `post`) are shown in the table below.

Trainee	1	2	3	4	5	6
pre	41	46	35	49	33	42
post	42	47	43	55	28	49

The following code is now run in R for input data:

```
pre <- c(41,46,35,49,33,42)
post <- c(42,47,43,55,28,49)
```

Question VII.1 (13)

Assuming the samples are from normal distributed populations and the improvement is measured by a difference in the mean of the populations, which of the following codes will result in the desired hypothesis test after reading in the data?

1 ☐
```
t.test(mean(pre), mean(post))
```

2 ☐
```
t.test(post, pre, var.equal=TRUE)
```

3 ☐
```
t.test(pre, post)
```

4* ☐
```
t.test(pre-post)
```

5 ☐
```
t.test(sd(pre), sd(post))
```

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

A paired t-test is designed to compare the means of the same group or item under two separate scenarios. An unpaired t-test compares the means of two independent or unrelated groups. In an unpaired t-test, an additional option is to assume the variance of the groups to be equal.

The standard statistical test for this setup is a paired two-sample t-test and can also be computed by taking the difference between the observations for each individual in the form of a one-sample t-test.

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

Continue on page 16

Exercise VIII

It is assumed that the there are 2.5 forest fires on average on a hot summer day, and that the number of forest fires per day follows a Poisson distribution.

Question VIII.1 (14)

What is the probability that there are at least five forest fires on a hot summer day?

1 ☐ 0.004

2 ☐ 0.067

3* ☐ 0.109

4 ☐ 0.175

5 ☐ 0.762

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

It is the probability $P(X \geq 5)$, where $X \sim \exp(\lambda = 2.5)$, which we get by

```
1 - ppois(4,2.5)

## [1] 0.108822
```

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

Question VIII.2 (15)

Let the random variable X represent the number of fores fires in a period of seven consecutive hot summer days. Which distribution does X follow?

1* ☐ A Poisson distribution with mean 17.5.

2 ☐ A binomial distribution with $n = 49$ and $p \approx 0.36$.

3 ☐ A Normal distribution with mean 17.5 and standard deviation 7.

4 ☐ An exponential distribution with rate ≈ 0.57.

5 ☐ None of the above.

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

When a Poisson process period length is changed, the number of events in the scaled period are still Poisson distributed, with the mean rate scaled accordingly. Hence, in the present case it has increased by 7 times, hence

```
2.5*7
```

```
## [1] 17.5
```

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

Question VIII.3 (16)

It is estimated that 78% of forest fires could have been avoided if simple precautions had been met. If the fire department one day reports five fores fires, what is the probability that all of these fores fires could have been avoided if the simple precautions had been met?

1 ☐ 0.0005

2 ☐ 0.156

3 ☐ 0.175

4* ☐ 0.289

5 ☐ 0.711

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

We must use the binomial distribution and calculate the probability by

```
dbinom(x= 5, size= 5, prob = 0.78)
```

```
## [1] 0.2887174
```

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

Continue on page 18

Exercise IX

30 butterflies were captured and wing lengths were measured in cm. Some summary statistics of the sample (stored in `length_cm`) are shown below.

```
round(summary(length_cm),2)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.06    3.72    3.93    4.03    4.32    5.37

round(sd(length_cm),2)

## [1] 0.49
```

Question IX.1 (17)

Perform parametric bootstrapping and simulate 10000 samples assuming that the population from which the sample was taken follows a normal distribution. Which of the following is the correct 99% confidence interval for the median of the wing length?

HINT: If you do the "default" coding from the book, you will get the correct result with `set.seed(2023)`.

1 ☐ [3.46 cm, 3.93 cm]

2 ☐ [3.38 cm, 4.00 cm]

3 ☐ [2.81 cm, 5.48 cm]

4* ☐ [3.76 cm, 4.31 cm]

5 ☐ [3.90 cm, 4.17 cm]

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

```
median_cm <- 3.93
mean_cm <- 4.03
sd_cm <- 0.49
n <- length(length_cm)
k <- 10000

set.seed(2023)
sim_samples <- replicate(k, rnorm(n, mean = mean_cm, sd = sd_cm))
sim_medians <- apply(sim_samples, 2, median)
round(quantile(sim_medians, c(0.005, 0.995)),2)
```

```
##  0.5% 99.5%
##  3.76  4.31
```

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

Question IX.2 (18)

The following code was run:

```
median_cm <- 3.93
mean_cm <- 4.03
sd_cm <- 0.49
n <- length(length_cm)
k <- 10000
```

Which of the following answers makes use of non-parametric bootstrapping to estimate the 95% confidence interval for the first quartile (25th percentile) of the wing lengths in cm?

1 ☐
```
fun <- function(x) quantile(x, 0.25, type=2)
sim_samples <- replicate(k, sample(length_cm, n, replace = TRUE))
sim_stats <- apply(sim_samples, 2, fun)
quantile(sim_stats, c(0.005, 0.995))
```

2 ☐
```
fun <- function(x) quantile(x, 0.25, type=2)
sim_samples <- replicate(k, rnorm(n, mean_cm, sd_cm))
sim_stats <- apply(sim_samples, 2, fun)
quantile(sim_stats, c(0.025, 0.975))
```

3 ☐
```
fun <- function(x) quantile(x, 0.75, type=2)
sim_samples <- replicate(k, sample(length_cm, n, replace = TRUE))
sim_stats <- apply(sim_samples, 2, fun)
quantile(sim_stats, c(0.01, 0.99))
```

4 ☐
```
fun <- function(x) quantile(x, 0.75, type=2)
sim_samples <- replicate(k, rnorm(n, mean_cm, sd_cm))
sim_stats <- apply(sim_samples, 2, fun)
quantile(sim_stats, c(0.01, 0.99))
```

5* ☐

```r
fun <- function(x) quantile(x, 0.25, type=2)
sim_samples <- replicate(k, sample(length_cm, n, replace = TRUE))
sim_stats <- apply(sim_samples, 2, fun)
quantile(sim_stats, c(0.025, 0.975))
```

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

Options 2 and 5 use the probability values appropriate for the 95% confidence interval. Among these two, only option 5 simulates 25th percentiles and uses non-parametric sampling.

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

Continue on page 21

Exercise X

A study involves two populations: a population of systems analysts using a current technology and a population of systems analysts using a new software package. The assumption of normal distributed population is fulfilled for both populations. In terms of the time required to complete a system design project, the population means are as follows:

μ_1 is the mean project completion time for systems analysts using the current technology.
μ_2 is the mean project completion time for systems analysts using the new software package.

The researcher in charge of the new software package would like to know if the new software package will provide a shorter mean project completion time. The hypothesis which must be tested is

H_0: $\mu_1 - \mu_2 = 0$
H_a: $\mu_1 - \mu_2 \neq 0$

The following summary statistics are provided:

- Sample sizes: $n_1 = 12$, $n_2 = 12$
- Sample means: $\bar{x}_1 = 325$ hours, $\bar{x}_2 = 286$ hours
- Sample standard deviations: $s_1 = 40$, $s_2 = 44$

Question X.1 (19)

What is value of the t-test statistic (t_{obs}) and the degree of freedom (both answers must be correct)?

1 ☐ $t_{\text{obs}} = 1.80$ and $df = 16.8$

2* ☐ $t_{\text{obs}} = 2.27$ and $df = 21.8$

3 ☐ $t_{\text{obs}} = 2.91$ and $df = 20.8$

4 ☐ $t_{\text{obs}} = 3.12$ and $df = 19.8$

5 ☐ $t_{\text{obs}} = 2.02$ and $df = 21.8$

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

Use Theorem 3.50.

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

Continue on page 22

Exercise XI

In a laboratory, one wants to dilute a stock solution A of concentration C_A to obtain a solution B of concentration C_B. The following rule is applied:

$$C_B = \frac{C_A V_A}{V_B}$$

Here, V_A and V_B represent the volumes of the initial stock solution A and the resulting solution B, respectively. C_A, C_B, V_A and V_B are all random variables.

Question XI.1 (20)

Which of the following formulas can be utilized to approximate the standard deviation of solution B's concentration (σ_{C_B}) via the non-linear error propagation rule?

1 ☐ $\sigma_{C_B} = \sqrt{\frac{\left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right)^2}{\left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)^2}}$

2 ☐ $\sigma_{C_B} = \left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)^2$

3* ☐ $\sigma_{C_B} = \sqrt{\left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)^2}$

4 ☐ $\sigma_{C_B} = \frac{\left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right)^2}{\left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)^2}$

5 ☐ $\sigma_{C_B} = \left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right) + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right) + \left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)$

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

Apply the Method 4.3 The non-linear approximative error propagation rule.

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

Question XI.2 (21)

Let X_i be a random variable. The following code is run in R to draw k realizations of X_i from a distribution:

```
x <- rnorm(k)^2 + rnorm(k)^2
```

Which of the following statements is correct?

1 ☐ X_i follows a χ^2-distribution with 1 degree of freedom, and thus has $q_{0.25} = 0.102$.

2 ☐ X_i follows a standard normal distribution with mean 0 and variance 1, and thus has $q_{0.25} = -0.674$.

3* ☐ X_i follows a χ^2-distribution with 2 degrees of freedom, and thus has $q_{0.25} = 0.575$.

4 ☐ X_i follows a χ^2-distribution with 3 degrees of freedom, and thus has $q_{0.25} = 1.213$.

5 ☐ X_i follows a normal distribution with mean 0 and variance 3, and thus has $q_{0.25} = -1.168$.

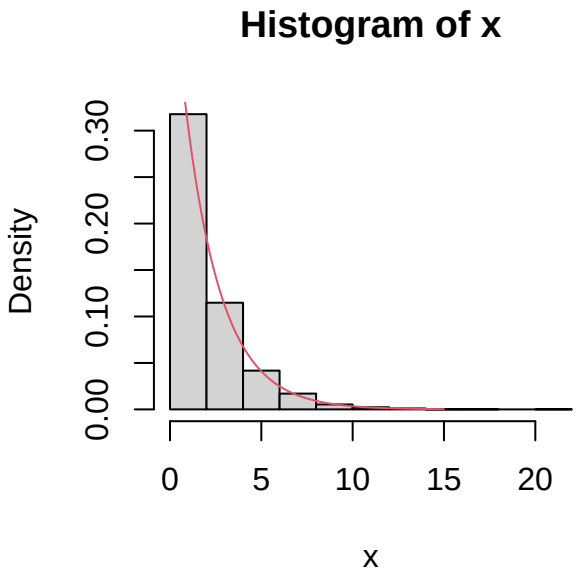
---------- FACIT-BEGIN ----------

We draw k standard normal numbers in a vector and square each. We do that twice and sum the two vectors, such that we have k numbers which are χ^2-distributed. So, according to 2.79, they have 2 df. Check it by:

```
k <- 10000
x <- rnorm(k)^2 + rnorm(k)^2
length(x)

## [1] 10000

hist(x, prob=TRUE)
xseq <- seq(0,15,by=0.1)
lines(xseq, dchisq(xseq, df=2), col=2)
```



The theoretical 25% quantile can be found by:

```
qchisq(0.25, df=2)

## [1] 0.5753641
```

Alternatively, it can be solved by simulation by approximating the quantile:

```
k <- 100000
x <- rnorm(k)^2 + rnorm(k)^2
quantile(x, 0.25)

##       25%
## 0.5769594
```

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

Question XI.3 (22)

Which of the following statements is NOT true?

1 ☐ Non-parametric bootstrapping is a re-sampling technique used to estimate the variability of a parameter without making any assumptions about the underlying population.

2 ☐ Non-parametric bootstrapping involves repeatedly sampling with replacement from the original sample to create many new, simulated samples of the same size as the original sample.

3* ☐ Non-parametric bootstrapping should be preferred over parametric bootstrapping if you know the distribution of the population.

4 ☐ Non-parametric bootstrapping can be applied to estimate the 95% confidence interval of a sample mean.

5 ☐ Non-parametric bootstrapping can be used to estimate confidence intervals for a parameter and to test hypotheses.

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

If the distribution of the population is known, then the parametric bootstrapping should be the preferred method. See more in Section 4.2 and 4.3.

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

Continue on page 27

Exercise XII

A certain industrial process depends on the pH value and the amount of catalyst in a solution. The relationship between output per hour (kg/h) (`efficiency`), pH (`ph`), and the amount of catalyst used (`catalyst`) is to be investigated using the following multiple linear regression model

$$\mathit{efficiency}_i = \beta_0 + \beta_1 \cdot \mathit{ph}_i + \beta_2 \cdot \mathit{catalyst}_i + \varepsilon_i \,,$$

where the ε_i are independent and $N(0, \sigma^2)$-distributed. R output from fitting the model to the data available is shown below:

```
##
## Call:
## lm(formula = efficiency ~ ph + catalyst)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6360  -2.4232  -0.5899   1.7566  16.8564
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.660     14.652   2.707  0.01496 *
## ph            -4.059      1.532  -2.649  0.01687 *
## catalyst       4.593      1.247   3.683  0.00184 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.34 on 17 degrees of freedom
## Multiple R-squared:  0.5492,Adjusted R-squared:  0.4962
## F-statistic: 10.36 on 2 and 17 DF,  p-value: 0.001145
```

Question XII.1 (23)

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

1 ☐ The effect of catalyst on efficiency is not significant, because the p-value is less than 0.01.

2* ☐ The amount of catalyst have a significant effect on efficiency, while the pH does not.

3 ☐ Both the amount of catalyst and the pH are significant, because the p-values are less than 0.05.

4 ☐ Neither the amount of catalyst nor the pH are significant, because the p-values are less than 0.05.

5 ☐ The model intercept is significant, because the p-value of 0.0150 is greater than 0.01.

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

Given the significance level, p-values have to be smaller than 0.01 for significance. Therefore, catalyst is significant, whereas pH is not.

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

Question XII.2 (24)

Look at the same R output as above. What effect does an increase of two units of catalyst have on the expected hourly output assuming pH level is kept constant at 4?

1 ☐ The expected hourly output increases by 3.87 kg per hour.

2 ☐ The expected hourly output decreases by 4.06 kg per hour.

3 ☐ The expected hourly output increases by 4.59 kg per hour.

4* ☐ The expected hourly output increases by 9.19 kg per hour.

5 ☐ The expected hourly output remains constant.

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

The estimated slope for catalyst is $\hat{\beta}_2 = 4.593$. A two unit increase of catalyst gives an expected increase of expected hourly output at $2 \cdot 4.593 = 9.186$ (ph is kept constant, hence no contribution to the change from that).

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

Continue on page 29

Exercise XIII

In an experiment, five different chemical substances were mixed with soil in equivalent concentrations. The same type of plants were grown next to each other in each of the five mixes, hence the plants were exposed to the same growing conditions except for the different chemicals in the soil.

When the plants were matured, they were harvested for each of the five different chemical-mixed soil conditions and it was counted how many of them had a trace of the chemicals.

The recorded counts for each of the chemicals were:

	A	B	C	D	E
Trace	12	14	18	5	15
No trace	38	35	35	42	35

The R code for reading in the data is given by:

```
tbl <- matrix(c(12, 14, 18, 5, 15,
                38, 35, 35, 42, 35), nrow=2, byrow=TRUE)
```

Question XIII.1 (25)

Under the null hypothesis of no difference in proportion of plants that had no trace of chemicals

$$H_0 : p_i = p, i = 1, \ldots, 5$$

what is the expected number of plants with no trace for chemical C?

1 ☐ 12.1

2 ☐ 12.6

3* ☐ 39.4

4 ☐ 50.2

5 ☐ 53.1

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

Under the null hypothesis of no difference the best estimate of the proportion is calculated by the totals

```r
# The total sum of plants with no trace
(xNoTrace <- sum(tbl[2, ]))

## [1] 185

# divided by the total number
(phat <- xNoTrace / sum(tbl))

## [1] 0.7429719
```

The number of observations with chemical C is $18 + 35 = 53$. So, the expected number of plants with no trace for chemical C is

```r
phat * 53

## [1] 39.37751
```

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

Question XIII.2 (26)

What is the conclusion of the hypothesis test using significance level of $\alpha = 0.05$ for the null hypothesis that there is no difference in proportion of plants

$$H_0 : p_i = p, i = 1, \ldots, 5$$

with trace of chemical (both argument and conclusion must be correct)?

1 ☐ The relevant p-value is 0.041, however the rule of thumb for validity is not ok. Therefore, no conclusion can be drawn.

2 ☐ The relevant p-value is 0.041 and the rule of thumb for validity is checked ok. Therefore, a significant difference in uptake of the five chemicals is detected.

3 ☐ The relevant p-value is 0.083, however the rule of thumb for validity is not checked ok. Therefore, no conclusion can be drawn.

4* ☐ The relevant p-value is 0.083 and the rule of thumb for validity is checked ok. Therefore, it is accepted that no difference in uptake of the five chemicals is detected.

5 ☐ The relevant p-value is 0.021 and the rule of thumb for validity is checked ok. Therefore, a significant difference in uptake of the five chemicals is detected.

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

See Method 7.20. We must carry out the χ^2-test:

```
chisq.test(tbl, correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 8.2493, df = 4, p-value = 0.08286
```

and check the rule of thumb for validity, hence check that all the expected values under the null hypothesis are above 5:

```
  chisq.test(tbl)$expected

##          [,1]     [,2]     [,3]     [,4]     [,5]
## [1,] 12.85141 12.59438 13.62249 12.08032 12.85141
## [2,] 37.14859 36.40562 39.37751 34.91968 37.14859
```

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

Question XIII.3 (27)

In another experiment, seeds from 200 different plant types were selected. For each plant type, a particular gene modification was made to some seeds. The regular and the modified seeds were grown pairwise next to each other and thus exposed to the same conditions. After the plants were matured, their quality for sale was assessed. The resulting counts were

	Low	Medium	High
Low	25	9	19
Medium	23	21	23
High	28	33	19

where the regular seed categories are the rows and the modified seed categories are the columns.

It is desired to carry out a test for independence of the two variables on significance level $\alpha = 0.05$. The validity of the test has been checked ok. What is the result of the usual test (both conclusion and arguments must be correct)?

1 ☐ The relevant p-value is 0.048, thus it is accepted that that the variables are independent.

2* ☐ The relevant p-value is 0.048, thus it is concluded that that the variables are not independent.

3 ☐ The relevant p-value is 0.063, thus it is concluded that that the variables are not independent.

4 ☐ The relevant p-value is 0.063, thus it is accepted that that the variables are independent.

5 ☐ None of the answers above are correct.

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

See Theorem 7.24 and the example below the theorem.

A test for independence of two categorical variables observed on the same units can be carried out with the χ^2-test. The R code for reading in the table is:

```
tbl <- matrix(c(25, 9, 19,
                23, 21, 23,
                28, 33, 19), byrow=TRUE, nrow=3)
tbl
```

```
##      [,1] [,2] [,3]
## [1,]   25    9   19
## [2,]   23   21   23
## [3,]   28   33   19
```

so we simply use the R function:

```
chisq.test(tbl)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 9.5757, df = 4, p-value = 0.04821
```

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

Continue on page 33

Exercise XIV

Some researchers examine the effects of different types of software and different types of hardware on the total runtime of a computer program. The researchers have applied a two-way ANOVA model and have obtained the following ANOVA table, where some numbers have been replaced with letters:

Source	DF	Sum Sq	Mean Sq	Test statistic	p-value
Software	5	$SS(S)$	$MS(S)$	F_1	0.432
Hardware	4	$SS(H)$	$MS(H)$	F_2	0.036
Residual	20	SSE	MSE		

Question XIV.1 (28)

What can be deduced about the relationship between F_1 and F_2 from the information provided in the ANOVA table above?

1 ☐ $F_1 > F_2$

2 ☐ $0.036 \cdot F_1 = 0.432 \cdot F_2$

3 ☐ $F_1 = F_2$

4* ☐ $F_1 < F_2$

5 ☐ Nothing can be deduced about the relationship between F_1 and F_2 with provided information.

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

Since hardware has fewer degrees of freedom than software and the p-value is higher for software, F_2 must be larger than F_1. Alternatively, we can find the exact values as

```
F1 <- qf(1-0.432,df1=5,df2=20)
F1

## [1] 1.020443

F2 <- qf(1-0.036,df1=4,df2=20)
F2

## [1] 3.169142
```

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

Question XIV.2 (29)

Which of the following statements is correct?

1 ☐ Using a significance level of 1%, we find that both the software type and the hardware type have significant effect on the runtime of the computer program.

2 ☐ Using a significance level of 5%, we find that both the software type and the hardware type have significant effect on the runtime of the computer program.

3 ☐ Using a significance level of 10%, we find that neither the software type nor the hardware type have significant effect on the runtime of the computer program.

4 ☐ Using a significance level of 5%, we find that the software type has significant effect on the runtime of the computer program, while the hardware type does not.

5* ☐ Using a significance level of 1%, we find that neither the software type nor the hardware type have significant effect on the runtime of the computer program.

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

Since both the p-values are higher than 0.01, we may conclude that there is no evidence supporting that either the software type or the hardware type has significant effect on the runtime.

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

Question XIV.3 (30)

The researchers now want to perform post-hoc analysis and have selected a significance level α. The researchers want to make pairwise comparisons of the mean runtime for all the different hardware types, and in order to control the type I error rate, the researchers want to use the Bonferroni correction. What Bonferroni corrected significance level should the researchers use?

1 ☐ $\alpha/4$

2 ☐ $\alpha/5$

3 ☐ $\alpha/6$

4* ☐ $\alpha/10$

5 ☐ $\alpha/15$

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

There are $4 + 1 = 5$ different types of hardware used in the experiments, which implies that we make $5 \cdot 4/2 = 10$ different comparisons. Hence, the researchers should divide the selected significance level by 10.

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

The exam is finished. Enjoy the summer!