Written examination: (26. June 2024)

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

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

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

The exam paper contains 42 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

Two students are counting the number of cars passing by on different stretches of road. They assume that the number of cars passing by in specific time intervals follow Poisson distributions. On the first road (road 1) they assume that the expected number of cars passing by is $\lambda_1 = 10/\text{hour}$, while on the second road (road 2) they assume that the expected number of cars passing by is $\lambda_2 = 15/\text{hour}$.

Now they define two random variables:

- X_1: number of cars passing by on road 1 in 15 minutes
- X_2: number of cars passing by on road 2 in 10 minutes.

You can assume that X_1 and X_2 are independent.

Question I.1 (1)

What is the probability $P(X_1 = 10)$?

1 ☐ 0.125

2 ☐ 0.417

3* ☐ 0.000216

4 ☐ 0.875

5 ☐ 0.583

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

The number of cars passing by in 1 hour follow a Poisson distribution with $\lambda_{1hour} = 10$ and therefor $X_1 \sim Pois(\lambda_1)$ with $\lambda_1 = 10/4 = 2.5$, this probability can be calculated by

```
dpois(10,2.5)

## [1] 0.0002157252
```

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

Continue on page 3

Question I.2 (2)

Which of the following statements about the expected values of the two random variables is correct?

1 ☐ $\frac{E[X_1]}{E[X_2]} = 1.5$

2 ☐ $\frac{E[X_1]}{E[X_2]} = \frac{2}{3}$

3 ☐ $\frac{E[X_1]}{E[X_2]} = \frac{1}{3}$

4 ☐ $\frac{E[X_1]}{E[X_2]} = 3$

5* ☐ $\frac{E[X_1]}{E[X_2]} = 1$

``` 
----------------------------------- FACIT-BEGIN -----------------------------------
```

With the assmuptions of the model then $X_1 \sim Pois(\lambda_1)$ and $X_2 \sim Pois(\lambda_2)$, with $\lambda_1 = 10/4 = 2.5$ and $\lambda_2 = 15/6 = 2.5$ and therefore $E[X_1] = 10/4 = 2.5$ and $E[X_2] = 15/6 = 2.5$ and hence the ratio is 1.

```
------------------------------------ FACIT-END ------------------------------------
```

Question I.3 (3)

What is the probability that the time between two cars passing by is greater than 2 minutes on road 2?

1 ☐ 0.5

2 ☐ 0.184

3* ☐ 0.607

4 ☐ 0.368

5 ☐ 0.816

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

The time between cars passing by is Exponantially distributed and on road 2 the mean value of the time between 2 cars passing by is $60/15 = 4$ minutes, in R we can calculate the number by

```r
lambda2 <- 15 / 60
1 - pexp(2, lambda2)

## [1] 0.6065307
```

```
------------------------------------ FACIT-END ------------------------------------
```

Continue on page 6

Exercise II

A farm made a study in which 225 chickens were randomly divided into 3 treatment groups of 75 animals each. Each group were fed with fodder from different feed producers during a period. For each chicken the weight change over the period of time was measured and the final data set consists of 225 observations of weight changes. The objective of the study is to determine if there is statistical evidence for difference in mean weight change for at least one of the groups. It may be assumed that the variance is the same in all treatment groups.

Question II.1 (4)

What kind of statistical analysis is most suitable for this?

1 ☐ Multiple linear regression analysis

2 ☐ Test for independence in a $r \times c$ frequency table (Contingency table)

3 ☐ Paired t-tests

4* ☐ One-way analysis of variance

5 ☐ t-tests

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

With the description this is clearly 3 independent samples of quantitative data, so the oneway anova is the right choice, so answer 4).

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

Continue on page 7

Exercise III

The engineers at an international airport have conducted a survey, in which they have timed 40 randomly selected security checks. The average duration of the security checks included in the survey was 34.66 seconds, and the sample standard deviation was 10.12 seconds, it is assumed that the times are normally distributed.

Question III.1 (5)

Based on the survey, what is the 99% confidence interval for the mean duration of the security checks?

1 ☐ $[7.26; 62.06]$

2 ☐ $[14.19; 55.13]$

3* ☐ $[30.33; 38.99]$

4 ☐ $[31.42; 37.90]$

5 ☐ $[33.06; 36.26]$

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

The engineers apply method 3.9 with $\alpha = 1\%$, which yields

```
34.66 + c(-1,1)*qt(1-0.01/2,df=40-1)*10.12/sqrt(40)

## [1] 30.32703 38.99297
```

Thus, the 99% confidence interval becomes $[30.33; 38.99]$.

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

Question III.2 (6)

What is the p-value for the usual test of the null hypothesis $H_0 : \mu = 30$ against a two-sided alternative hypothesis?

1 ☐ 0.30%

2* ☐ 0.59%

3 ☐ 4.72%

4 ☐ 94.23%

5 ☐ 99.70%

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

The engineers apply method 3.23. First, the observed test statistic is calculated as

$$t_{\text{obs}} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{34.66 - 30}{10.12/\sqrt{40}} = 2.9123.$$

```
tobs <- (34.66-30)/(10.12/sqrt(40))
```

Then the p-value is then calculated as

$$p = 2P(T > |t_{\text{obs}}|) = 0.5908\%.$$

```
p <- 2*(1-pt(abs(tobs),df=40-1))
```

Thus, the p-value rounded to two decimals is 0.59%.

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

Continue on page 9

Exercise IV

A company wants to estimate the cost of producing solar panels. Therefore, the engineers have designed an experiment to evaluate the costs of producing batches of different sizes and ensure that the observations are completely independent. The results are given in the table below.

Batch size (units)	50	100	150	200	250	300	350	400	450	500
Costs (M DKK)	2.33	4.21	6.01	7.51	8.46	8.93	9.45	10.70	10.55	10.74

The data can be read in R using the below code:

```
Batch<-1:10 * 50
Costs<-c(2.33,4.21,6.01,7.51,8.46,8.93,9.45,10.70,10.55,10.74)
```

The engineers initially believe that the data can be described by a linear model on the form

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i \in \{1, ..., 10\},$$

where the errors, ϵ_i, are independent and identically distributed (i.i.d.) with a $N(0, \sigma^2)$ distribution. In the model, the response variable is the cost (in M DKK) and the explanatory variable is the batch size (in units). The engineers therefore fit a linear regression model using the least squares method.

Question IV.1 (7)

What proportion of the variation in the costs is explained by the regression model?

1 ☐ 89.4%

2* ☐ 90.6%

3 ☐ 93.9%

4 ☐ 95.2%

5 ☐ 96.9%

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

The proportion of explained variation is the coefficient of determination (R^2), cf. definition 5.25. The R^2 can be found as the "Multiple R-squared" in the output from the summary command:

```
fit <- lm(Costs~Batch)
summary(fit)

##
## Call:
## lm(formula = Costs ~ Batch)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4733 -0.5129  0.2950  0.5756  1.0250
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.895333   0.641737   4.512  0.00197 **
## Batch       0.018159   0.002069   8.779 2.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9394 on 8 degrees of freedom
## Multiple R-squared:  0.906,Adjusted R-squared:  0.8942
## F-statistic: 77.07 on 1 and 8 DF,  p-value: 2.225e-05
```

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

Continue on page 11

Question IV.2 (8)

What is the 99% confidence interval for the slope, β_1?

1* ☐ $[0.011, 0.025]$

2 ☐ $[0.013, 0.023]$

3 ☐ $[0.016, 0.020]$

4 ☐ $[0.742, 5.049]$

5 ☐ $[1.415, 4.375]$

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

Method 5.15 gives the confidence intervals for the model parameters in a simple linear regression. From the output in the previous question, the estimate and the sample standard deviation of the estimate is found as

$$\hat{\beta}_1 = 0.018 \quad \wedge \quad \hat{\sigma}_{\beta_1} = 0.002.$$

The 99% confidence interval is thus calculated as

$$\hat{\beta}_1 \ \pm \ t_{0.995}\hat{\sigma}_{\beta_1} = 0.018 \ \pm \ 3.355 \cdot 0.002,$$

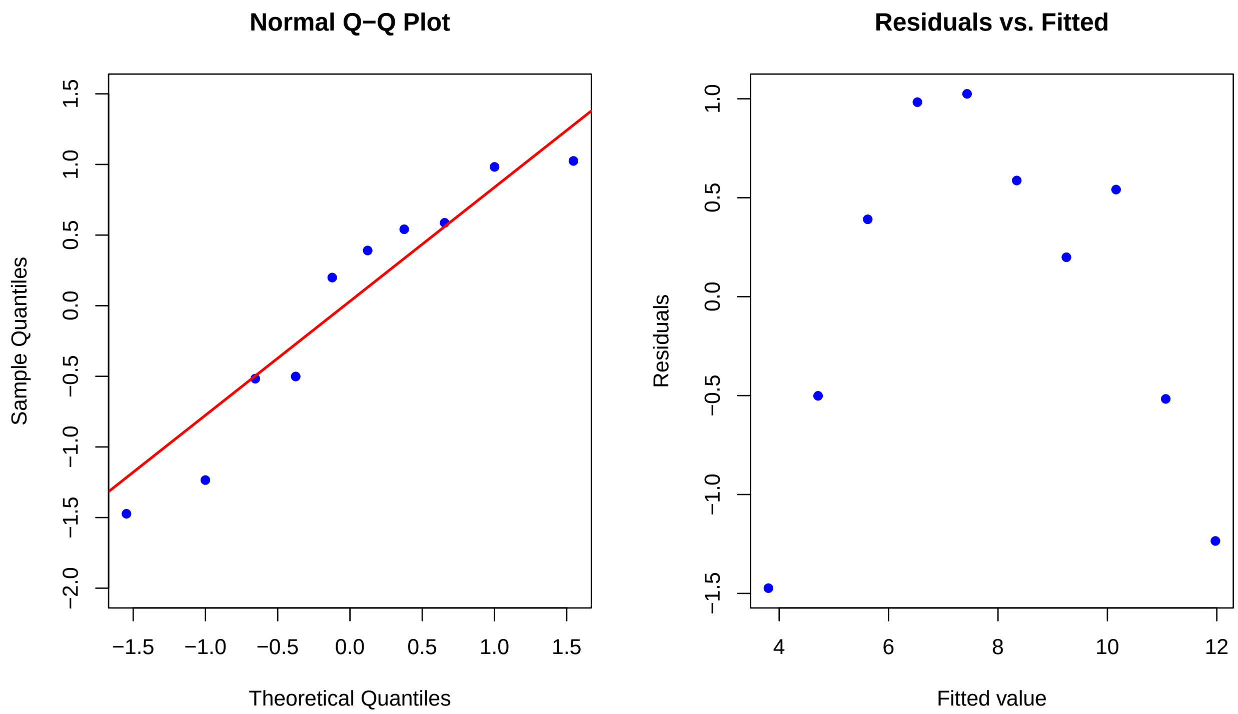
where $t_{0.995}$ is the 99.5% quantile in a t-distribution with $10 - 2 = 8$ degrees of freedom. Alternatively, the confidence intervals can be found using R as:

```
confint(fit,level=0.99)

##                   0.5 %     99.5 %
## (Intercept) 0.74205577 5.04861090
## Batch       0.01121815 0.02509943
```

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

To validate the model, the engineers make two diagnostic plots: A normal Q-Q plot and a plot of the residuals against the fitted values. The two plots are seen below:



Continue on page 13

Question IV.3 (9)

Which of the following statements regarding the validity of the model is most correct?

1 ☐ The diagnostic plots do not indicate any violations of the model assumptions

2 ☐ The normal Q-Q plot indicates that the residuals are not normally distributed

3 ☐ The residuals vs. fitted values plot indicates that the residuals are not normally distributed

4* ☐ The residuals vs. fitted values plot indicates that the residuals are not independent of the fitted values

5 ☐ The model assumptions must be satisfied as the R^2-value is high.

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

Considering that there are only 10 data points, the small deviations from the straight line in the normal Q-Q plot do not warrant concern. However, the residuals vs. fitted values plot clearly shows a non-linear trend (approximately quadratic), which is a violation of the model assumptions as the residuals must be independent of the fitted values.

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

Continue on page 14

Exercise V

The number of persons living in Danish dorms in 2023 is provided by Statistics Denmark. Here we focus only on a few age-categories:

	males	females	Total
18-24	14048	14128	28176
25-29	8215	6028	14243
30-39	2735	1397	4132
Total	24998	21553	46551

Question V.1 (10)

What proportion of the residents of Danish dormitories are males (among the 18-39 year olds)?

1 ☐ 0.463

2 ☐ 0.500

3 ☐ 0.521

4* ☐ 0.537

5 ☐ 0.409

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

The proportion is simply the ratio between the nomber of males and the total number of persons:

```
(phat_males = 24998 / 46551)

## [1] 0.537
```

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

Question V.2 (11)

We would like to know whether the proportions of males in different age-groups is significantly different (using a 5% significance level). Which of the following statements is true?

1 ☐ We should use a paired t-test with $\alpha = 0.05$ to test if there is a significant difference between age-groups. The result is that we do observe a significant difference.

2 ☐ We should use an unpaired t-test with $\alpha = 0.025$ to test if there is a significant difference between age-groups. The result is that we do observe a significant difference.

3 ☐ We should use an unpaired t-test with $\alpha = 0.05$ to test if there is a significant difference between age-groups. The result is that we do NOT observe a significant difference.

4 ☐ We should use a χ^2 test with 6-degrees of freedom to test if there is a significant difference between age-groups. The result is that we do observe a significant difference.

5* ☐ We should use a χ^2 test with 2-degrees of freedom to test if there is a significant difference between age-groups. The result is that we do observe a significant difference.

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

There are multiple ways to solve this, the easy one are using `prop.test` or `chisq.test`, the solution below use `chisq.test`

```
kollegier <- as.table(rbind(c(14048, 14128), c(8215, 6028), c(2735, 1397)))
dimnames(kollegier) <- list(c("18-24", "25-29", "30-39"),
                        c("men","women"))
chisq.test(kollegier)

##
##  Pearson's Chi-squared test
##
## data:  kollegier
## X-squared = 517, df = 2, p-value <2e-16
```

The p-value is below the significance level, so the null hypothesis ("no difference between groups") is rejected.

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

Question V.3 (12)

Under the hypothesis that the distribution between for male and females is the same for all age groups, what is the expected number of males in the age group 18-24 living in dorms (to be compared with the table above and used for calculating the appropriate test-statistic)?

1 ☐ 14088

2* ☐ 15131

3 ☐ 15517

4 ☐ 14048

5 ☐ 7759

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

Under the hypotheis the expected number is

$$e_{ij} = \frac{j\text{th column total} \cdot i\text{th row total}}{j\text{grand total}} \qquad (1)$$

or in our case

```
28176 * 24998 / 46551

## [1] 15131
```

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

Question V.4 (13)

We now consider only the 18-24 year-olds. Which of the following statements is true (if relevant, in the answer options, significance level $\alpha = 0.05$ is used)?

1* ☐ Among the 18-24 year-olds the proportion of males is NOT significantly different from 0.5, as the estimated 95% confidence interval for the proportion of males in this age-group is [0.493, 0.504]

2 ☐ Among the 18-24 year-olds the proportion of males is exactly 0.5.

3 ☐ Among the 18-24 year-olds the proportion of males is significantly different from 0.5, as the estimated 95% confidence interval for the proportion of males in this age-group is [0.501, 0.533]

4 ☐ Among the 18-24 year-olds the proportion of males is NOT significantly different from 0.5, as the estimated 95% confidence interval for the proportion of males in this age-group is [0.501, 0.533]

5 ☐ Among the 18-24 year-olds the proportion of males is significantly different from 0.5, as the estimated 95% confidence interval for the proportion of males in this age-group is [0.532, 0.542]

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

The question can be solved in several ways the easy one is to use `prop.test`

```r
prop.test(x=14048, n=28176, p = 0.5, correct = FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  14048 out of 28176, null probability 0.5
## X-squared = 0.2, df = 1, p-value = 0.6
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.493 0.504
## sample estimates:
##     p
## 0.499
```

from which we can see that the confodence intervla is [0.493, 0.504] which include 0.5 and hence the null hypothesis is accepted.

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

Continue on page 18

Question V.5 (14)

Assuming independence between individuals. What is the probability, that 100 or more females live in a dorm with 190 people, if the probability of an individual being a female is assumed to be 0.45?

1* ☐ The probability is 0.021

2 ☐ The probability is 0.015

3 ☐ The probability is 0.50

4 ☐ The probability is 0.45

5 ☐ The probability is 0.985

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

Under the stated assumptions the random variable is binomial or formally we consider X with

$$X \sim Binom(190, 0.45) \tag{2}$$

and the probability we are looking for is $P(X \geq 100) = 1 - P(X \leq 99)$ which we can find in R by

```
1 - pbinom(q=99, size=190, prob=0.45)

## [1] 0.0208
```

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

Continue on page 19

Exercise VI

A simple predator-prey model is the Lotka-Volterra model

$$\frac{dx}{dt} = \alpha x - \beta xy$$
$$\frac{dy}{dt} = \delta xy - \gamma y,$$

where x is the size of the prey population and y is size of the predator population. The equation allows for a constant of motion (i.e. the quantity will stay constant through time for given initial conditions) given by

$$K = y^{\alpha} e^{-\beta y} x^{\gamma} e^{-\delta x}.$$

Assume that $\alpha = 2/3$, $\beta = 4/3$, $\gamma = \delta = 1$, and that the predator and prey population sizes have been observed at $y = 1/2$ and $x = 1$ respectively. The uncertainties of the observations are assumed to be $\sigma_y^2 = 1/16^2$ and $\sigma_x^2 = 1/8^2$, and further the observations are assumed independent.

Question VI.1 (15)

Using the non-linear error propagation rule what is the approximation of the variance of K?

1 ☐ 0.312

2* ☐ 0

3 ☐ 0.559

4 ☐ 0.75

5 ☐ 0.889

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

We need the derivative of the constant of motion wrt. to x and y

$$\frac{\partial K}{\partial y} = \alpha y^{\alpha-1} e^{-\beta y} x^{\gamma} e^{-\delta x} - \beta y^{\alpha} e^{-\beta y} x^{\gamma} e^{-\delta x} \tag{3}$$
$$= K\left(\frac{\alpha}{y} - \beta\right) \tag{4}$$
$$\frac{\partial K}{\partial x} = \gamma y^{\alpha} e^{-\beta y} x^{\gamma-1} e^{-\delta x} - \delta y^{\alpha} e^{-\beta y} x^{\gamma} e^{-\delta x} \tag{5}$$
$$= K\left(\frac{\gamma}{x} - \delta\right) \tag{6}$$

inserting the given values we get

$$K = 2^{-2/3} e^{-4/3 \cdot 1/2} 1^1 e^{-1} = 2^{-2/3} e^{-2/3} \tag{7}$$

and further

$$\frac{\alpha}{y} - \beta = \frac{2/3}{1/2} - 4/3 = 0 \tag{8}$$

$$\frac{\gamma}{x} - \delta = 1 - 1 = 0 \tag{9}$$

hence according to the error propagation the variance of K is 0 (this is of course not the true variance and a better answer could be obtained through simulation).

`------------------------------------ FACIT-END ------------------------------------`

Continue on page 21

Question VI.2 (16)

Assume now that the predator population is observed without error ($\sigma_y^2 = 0$), and hence that the only source of uncertainty is the prey population. Using the variance from above ($\sigma_x^2 = 1/8^2$) and assmuming normality of the error in x, i.e. $X = x + \epsilon$ with $\epsilon \sim N(0, \sigma_x^2)$. In what interval does the standard deviation of K fall (the answer should not rely on the non-linear approximations of the error propagation rule and should rather be base on simulation)?

1 ☐ $(0.07, 0.1)$

2 ☐ $(0.12, 0.2)$

3* ☐ $(10^{-4}, 0.01)$

4 ☐ $(0.03, 0.05)$

5 ☐ $(0.3, 0.5)$

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

In order to solve the question we need many realisations of K based on different realisations of X, when $X \sim N(1, 1/8^2)$. Mathamatically we can write it as

$$K_i = y^\alpha e^{-\beta y} x^\gamma e^{-\delta x_i} \tag{10}$$

where x_i is a relisations of X. The standard error can now be calculated by

$$\hat{sd}(K) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(K_i - \bar{K})^2} \tag{11}$$

in R this can be done by (using $N = 1000$)

```
set.seed(3250783)
N <- 1e3
x <- rnorm(N,mean=1,sd=1/8)
y <- 1/2
alpha <- 2/3
beta <- 4/3
gamma <- delta <- 1
K = y ^ alpha * exp(- beta * y) * x ^ gamma * exp(- delta* x)
sd(K)

## [1] 0.001330441
```

which is clearly between 10^{-4} and 0.01.

The above is of course one realisation and the unceartainty could also be assesed repeating the above a large number (L) of times. This is done with $L = 10000$ times in the code below

```
set.seed(3250783)
N <- 1e3
L <- 1e4
x <- matrix(rnorm(N * L,mean=1,sd=1/8),ncol = L)
y <- 1/2
alpha <- 2/3
beta <- 4/3
gamma <- delta <- 1
K = y ^ alpha * exp(- beta * y) * x ^ gamma * exp(- delta* x)
range(apply(K,2,sd))

## [1] 0.001059092 0.001954591
```

I.e. in 10000 relalisations each time using 1000 realisations there is not one estimate of the standard error outside the interval given in answer 3.

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

Continue on page 23

Exercise VII

Let the function $f(x)$ be defined by

$$f(x) = \alpha\phi_1(x) + \beta\phi_2(x),$$

where $\phi_i(x)$ is the probability density function of a normal random variable with mean μ_i and variance σ_i^2.

Question VII.1 (17)

Under what conditions is $f(x)$ a probability density function (the answer should apply for any value of $\sigma_i > 0$ and $\mu_i \in \mathbb{R}$)?

1 ☐ $\alpha = \beta = 1$

2 ☐ $\alpha \in [0, 2]$ and $\beta = 2 - \alpha$

3 ☐ $\alpha = \frac{\sigma_1^2}{\sigma_2^2}$ and $\beta = \frac{\sigma_2^2}{\sigma_1^2}$

4* ☐ $\alpha \in [0, 1]$ and $\beta = 1 - \alpha$

5 ☐ $\alpha = \frac{\mu_1^2}{\sigma_1^2}$ and $\beta = \frac{\mu_2^2}{\sigma_2^2}$

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

For $f(x)$ to be a probability density function we need

$$f(x) \geq 0 \tag{12}$$

$$\int f(x)dx = 1. \tag{13}$$

Since $\phi_1(x)$ and $\phi_2(x)$ are both probability functions they are both larger than 0, and hence the first condition is fulfilled if $\alpha \geq 0$ and $\beta \geq 0$. For the second condition we can write the integral

$$\int f(x)dx = \int (\alpha\phi_1(x) + \beta\phi_2(x))dx \tag{14}$$

$$= \alpha \int \phi_1(x)dx + \beta \int \phi_2(x)dx, \tag{15}$$

again using that $\phi_1(x)$ and $\phi_2(x)$ are both probability density functions, the second condition amounts to

$$1 = \alpha + \beta \tag{16}$$

or $\beta = 1 - \alpha$ and hence $\beta \geq 0$ inply $\alpha \leq 1$, so combining we get $\alpha \in [0, 1]$ and $\beta = 1 - \alpha$, which is answer no 4. Of course an equivalent answer is $\beta \in [0, 1]$ and $\alpha = 1 - \beta$, but that is not one of options.

`----------------------------------- FACIT-END -----------------------------------`

Continue on page 24

Exercise VIII

An aircraft manufacturer uses an expensive type of screws in the production of a certain model. To reduce production costs, the manufacturer considers replacing the expensive screws with a cheaper type of screws. Therefore, the manufacturer tests the tensile strength (MPa) of the two types of screws, and the results are shown in the below table.

Tensile strengt	Cheap	Expensive
Sample mean (MPa)	1250	1300
Sample standard deviation (MPa)	54.24	28.54
Sample size	25000	15000

Question VIII.1 (18)

Assuming the samples were completely random, what is the 95% confidence interval for the difference in mean tensile strengths (mean of the cheap type minus mean of the expensive type) based on the test results?

1 ☐ $[-50.07; -49.93]$

2 ☐ $[-50.13; -49.87]$

3 ☐ $[-50.34; -49.66]$

4 ☐ $[-50.68; -49.32]$

5* ☐ $[-50.81; -49.19]$

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

The two samples are large and assumed independent. Therefore, method 3.47 applies, and the 95% confidence interval can be found as

$$\overline{x}-\overline{y} \ \pm \ t_{1-0.05/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = 1250-1300 \ \pm \ t_{0.975}\sqrt{\frac{54.24^2}{25000}+\frac{28.54^2}{15000}},$$

where $t_{0.975}$ is the 97.5%-quantile in a t-distribution with ν degrees of freedom (see below for a simple of $t_{0.975}$):

$$\nu = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}} = \frac{\left(\frac{54.24^2}{25000}+\frac{28.54^2}{15000}\right)^2}{\frac{(54.24^2/25000)^2}{25000-1}+\frac{(28.54^2/15000)^2}{15000-1}} = 39407.7444.$$

Thus, the 95% confidence interval is:

```
m1 <- 1250 ; m2 <- 1300
s1 <- 54.24; s2 <- 28.54
n1 <- 25000; n2 <- 15000
v <- ((s1^2/n1+s2^2/n2)^2)/(((s1^2/n1)^2)/(n1-1)+((s2^2/n2)^2)/(n2-1))
m1-m2+c(-1,1)*qt(1-0.05/2,df=v)*sqrt(s1^2/n1+s2^2/n2)

## [1] -50.81283 -49.18717
```

which rounded to two decimal points give $[-50.81; -49.19]$.

Note that with the large sample size the degrees of freedom (we know it must be between 15000 and 25000+15000-2=39998) is not really needed, as quantiles of the t-distributioon will be almost equal quantiles in the standard normal distribution when ν is large, and the interval can be calculated by

```
m1-m2+c(-1,1)*qnorm(1-0.05/2)*sqrt(s1^2/n1+s2^2/n2)

## [1] -50.81281 -49.18719
```

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

Question VIII.2 (19)

Under the null hypothesis $H_0 : \mu_{\text{cheap}} - \mu_{\text{expensive}} = -50$, what is the observed test statistic based on the test results?

1 ☐ -241.14

2 ☐ -120.57

3 ☐ -2.31

4* ☐ 0.00

5 ☐ 241.14

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

Method 3.51 is applied to calculate the observed test statistic (method 3.49 can also be used, but there is a mistake in method 3.49):

$$t_{\text{obs}} = \frac{\hat{\mu}_{\text{cheap}} - \hat{\mu}_{\text{expensive}} - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{1250 - 1300 - (-50)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = 0.$$

`--------------------------------- FACIT-END ----------------------------------`

The manufacturer also considers buying a new and more fuel-efficient aircraft model, and therefore they have measured the fuel-consumption (in kg) of the two models on 10 popular routes under similar weather and weight conditions. The manufacturer is only interested in comparing the logarithm of the fuel consumption as given in the below table:

Log of fuel consumption	Current model	New model
Sapporo - Tokyo	7.964	7.932
Sydney - Melbourne	7.813	7.762
Mumbai - Delhi	8.299	8.243
Beijing - Shanghai	8.219	8.174
Paris - Montreal	9.832	9.782
Dubai - London	9.829	9.775
London - New York	9.842	9.794
New York - Los Angeles	9.498	9.445
Kuala Lumpur - Singapore	7.023	6.942
Cancun - Mexico City	8.408	8.347

The data can be read in R using the below code:

```r
log_c <- c(7.964,7.813,8.299,8.219,9.832,9.829,9.842,9.498,7.023,8.408)
log_n <- c(7.932,7.762,8.243,8.174,9.782,9.775,9.794,9.445,6.942,8.347)
```

Question VIII.3 (20)

What is the p-value for the appropriate test of the null hypothesis $H_0 : \delta = \mu_{\text{current}} - \mu_{\text{new}} = 0.05$ against a two-sided alternative hypothesis? (Here μ_{current} and μ_{new} refer to the mean of the logarithm of the fuel consumption.)

1 ☐ $p < 0.001$

2 ☐ $p = 0.442$

3* ☐ $p = 0.452$

4 ☐ $p = 0.908$

5 ☐ $p = 0.995$

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

Since the observations are paired (one pair for each route), the methods described in section 3.2.3 applies, and a paired t-test is performed. This is equivalent to performing a one sample t-test using the differences in the log fuel consumption. Since the data is available, the test can be performed in R using the code:

```
t.test(log_c,log_n,mu=0.05,paired=TRUE)

##
##  Paired t-test
##
## data:  log_c and log_n
## t = 0.78574, df = 9, p-value = 0.4522
## alternative hypothesis: true mean difference is not equal to 0.05
## 95 percent confidence interval:
##  0.04417506 0.06202494
## sample estimates:
## mean difference
##          0.0531
```

Reading from the output, the p-value is 0.452.

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

Continue on page 29

Exercise IX

A car manufacturer wants to find out if there is a difference in breaking strength in beams made with metal from different suppliers. Let Y represent the breaking strength of beams. In the following breaking strength are measured on beams each made with metal from a single supplier. Metal from four different suppliers were included in the study and the breaking strength was measured for 5 similar beams from each supplier:

Supplier A	Supplier B	Supplier C	Supplier D
92.0	131.0	74.1	90.4
111.6	103.5	52.8	95.2
98.4	100.0	82.5	87.6
87.7	84.7	94.7	63.2
134.9	134.5	107.3	119.5

Question IX.1 (21)

The engineers in the company have conducted the following analysis in R. What is the conclusion, at significance level $\alpha = 5\%$, about the difference in breaking strength of the test beams made with metal from the different suppliers (both conclusion and argument must be correct)?

```
anova(lm(y ~ Supplier))

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value Pr(>F)
## Supplier   3 2508.8  836.25   2.027 0.1507
## Residuals 16 6601.0  412.56
```

1* ☐ A significant difference in breaking strength is not found, since the p-value is greater than the significance level.

2 ☐ A significant difference in breaking strength is found, since the p-value is greater than the significance level.

3 ☐ A significant difference in breaking strength is not found, since the p-value is less than the significance level.

4 ☐ A significant difference in breaking strength is found, since the p-value is less than the significance level.

5 ☐ None of the above conclusions are correct.

`----------------------------------- FACIT-BEGIN -----------------------------------`

From the ANOVA output we can read that the $p - value$ is 0.1507 and as 0.1507 is larger than the significance level there is a significant difference between the suppliers (See the second half of Chapter 8).

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

Continue on page 31

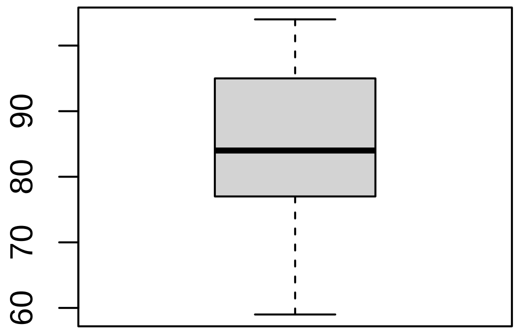
Exercise X

A sample has been taken from a population. It is read into R by:

```
x <- c(82,  70,  77,  59,  86,  81, 102,  95,  89, 104)
```

Question X.1 (22)

The modified boxplot of the sample is shown below:



Considering the definition of an extreme observation in the modified boxplot and the information given, which of the following statements is then correct?

1* ☐ None of the observations are identified as extreme.

2 ☐ One of the observations is identified as extreme.

3 ☐ Two of the observations are identified as extreme.

4 ☐ Three of the observations are identified as extreme.

5 ☐ With the given information it cannot be concluded if any of the observations are identified as extreme.

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

Clearly, none of the observations are identified as extreme, since there are no marked point(s) above or below the upper or lower whisker.

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

Question X.2 (23)

Assuming a normal distribution of the population. What is the 99.9% confidence interval for the population mean?

1 ☐ [76.41, 92.59]

2 ☐ [74.51, 94.49]

3* ☐ [63.39, 105.61]

4 ☐ [45.52, 123.47]

5 ☐ [33.12, 135.88]

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

```
t.test(x, conf.level=0.999)

##
##  One Sample t-test
##
## data:  x
## t = 19.138, df = 9, p-value = 1.339e-08
## alternative hypothesis: true mean is not equal to 0
## 99.9 percent confidence interval:
##   63.39107 105.60893
## sample estimates:
## mean of x
##      84.5
```

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

Continue on page 33

Exercise XI

A person (investor 1) decides to invest in stocks. The investor invest 10000 kr. in 10 different stocks (1000 kr. in each), after one year the investor sells all the stocks. The returns (measured in kr.) of the 10 stocks is given in the table below.

Stock	1	2	3	4	5	6	7	8	9	10
Return	1144	1218	1480	747	1178	-121	-382	-24	-652	-32

The investor wants to investigate if the investment has been a success. The success criterion is that the return is significantly different from (using a two-sided test) and bigger than a 2% (200 kr. in total, or 20 kr. per stock) return using significance level $\alpha = 5\%$.

The investor has dicided to use a test with no distributional assumption on the population, it is however assumed that the returns are independent. Some of the R-code below should be used for the next question.

```
x <- c(1144,  1218,   1480,    747,  1178, -121, -382,  -24, -652, -32)

k <- 10000
sim.samp <- replicate(k, sample(x, replace = TRUE))

quantile(apply(sim.samp, 2, mean), prob = c(0.025, 0.05, 0.95, 0.975),
         type = 2)

##   2.5%     5%    95%  97.5%
##   0.15  68.60 832.95 905.40

t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 1.8531, df = 9, p-value = 0.09687
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -100.5569 1011.7569
## sample estimates:
## mean of x
##     455.6

t.test(x, mu = 200)

##
##  One Sample t-test
##
```

```
## data:  x
## t = 1.0396, df = 9, p-value = 0.3256
## alternative hypothesis: true mean is not equal to 200
## 95 percent confidence interval:
##  -100.5569 1011.7569
## sample estimates:
## mean of x
##     455.6
```

Question XI.1 (24)

What is the result of the test (both the conclusion and the argument must be correct)?

1 ☐ The null-hypothesis is accepted as the p-value is 0.3

2* ☐ The null-hypothesis is accepted as $20 \in [0.15, 905.4]$

3 ☐ The null-hypothesis is rejected as the p-value is 0.1

4 ☐ The null-hypothesis is rejected as $0 < 0.15$

5 ☐ The null-hypothesis is rejected as $0 \in [-101, 1012]$

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

The results from `t.test` cannot be used since these rely on the normal distribution assumption of the relealised values. The first quantiles that are given are based on nonparametric bootstrap, and hence does not rely on distributional assumptions. Based on these quantiles a 95% confidence intreval for the return is [0.15, 905.4], and since 20 is inside that interval we concluded that the null hypothesis is accepted (this is answer no. 2).

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

Another investor has invested in 10 different stocks, also with 1000 kr. in each stock. The two investors want to compare the returns of their investments. In the R-code below `x` is the returns for investor 1 and `y` is the returns for investor 2.

```r
k <- 10000
sim.x <- replicate(k, sample(x, replace = TRUE))
sim.y <- replicate(k, sample(y, replace = TRUE))
sim.diff <- replicate(k, sample(x - y, replace = TRUE))

quantile(apply(sim.x, 2, mean) - apply(sim.y, 2, mean), prob = c(0.025,0.975),
         type = 2)
```

```
##        2.5%       97.5%
## -1562.82480    99.26984

quantile(apply(sim.diff, 2, mean), prob = c(0.025,0.975))

##       2.5%      97.5%
## -1186.2022  -285.1908

t.test(x, y, paired=TRUE)

##
##  Paired t-test
##
## data:  x and y
## t = -3.0451, df = 9, p-value = 0.0139
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -1298.7101  -191.5961
## sample estimates:
## mean difference
##       -745.1531
```

The two investors want to compare the returns of their investments using a statistical method with as few assumptions as possible, e.g. with no distributional assumptions at all, it is however assumed that all observations are independent.

Question XI.2 (25)

Is there, based on the R-output above, a significant (using significance level $\alpha = 5\%$) difference between the two investors returns (both the conclusion and the argument should be correct)?

1 ☐ Yes, since $0 \notin [-1186, -285]$

2 ☐ Yes, since $0.014 < 0.05$

3 ☐ Yes, since $0 \notin [-1299, -192]$

4 ☐ No, since $0.014 < 0.05$

5* ☐ No, since $0 \in [-1562, 99]$

`----------------------------------- FACIT-BEGIN ----------------------------------`

The situation here is that the data are indepedent and therefore it is an unpaired situation and hence the `t.test` results does not make sense here.

In the unpaired situation, which we have here, we should sample for `x` and `y` independently and hence the 95% confidence intrerval for the difference in mean is [-1186, 99.3] and with 0 inside the interval we do not find a significant difference (answer no. 5).

`------------------------------------ FACIT-END ------------------------------------`

Continue on page 37

Exercise XII

The engineers at a construction company is developing a model for the lifespan of buildings. The model can be written as

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i,$$

where Y_i (`Life`) is the lifespan (in years) of building i, x_{1i} (`Econ_RF`) the economical risk factor of building i, x_{2i} (`Matr_RF`) the material risk factor of building i, and x_{3i} (`Dsgn_RF`) the design risk factor of building i. Furthermore, the model assumes that the errors, ϵ_i, are independent and identically distributed with a $N(0, \sigma^2)$ distribution.

The engineers have fitted the model using the method of least squares, and the output from R is given below:

```
##
## Call:
## lm(formula = Life ~ Econ_RF + Matr_RF + Dsgn_RF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5709  -3.3908  -0.0053   3.3407  15.3690
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 151.76054    1.20254 126.199   <2e-16 ***
## Econ_RF      -0.35638    0.01270 -28.055   <2e-16 ***
## Matr_RF      -0.90717    0.01398 -64.878   <2e-16 ***
## Dsgn_RF      -0.11806    0.01289  -9.156   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.992 on 165 degrees of freedom
## Multiple R-squared:  0.9681,Adjusted R-squared:  0.9675
## F-statistic:  1669 on 3 and 165 DF,  p-value: < 2.2e-16
```

You may assume that the model assumptions are satisfied.

Continue on page 38

Question XII.1 (26)

One particular building had a lifespan of 81.3 years. The building had an economical risk factor of 64.7, a material risk factor of 55.2, and a design risk factor of 28.1. What is the residual associated with this building?

1 ☐ The residual cannot be determined without further information

2* ☐ 6.0 years

3 ☐ -6.0 years

4 ☐ 5.0 years

5 ☐ -5.0 years

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

The residual of the building is the difference between the observed lifespan and the lifespan predicted by the model. The predicted lifespan in this case is found as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 64.6 + \hat{\beta}_2 55.2 + \hat{\beta}_3 28.1 = 75.34.$$

Thus, the residual associated with the building becomes

$$e = y - \hat{y} = 81.3 - 75.34 = 5.96(\text{years}).$$

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

Question XII.2 (27)

How many buildings (observations) were included in the data set used to fit the model?

1 ☐ 165

2 ☐ 166

3 ☐ 167

4 ☐ 168

5* ☐ 169

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

The output states that the residual standard error has been estimated using 165 degrees of freedom. From Theorem 6.2, it is given that $n-(p+1)$ degrees of freedom should be used, i.e. $n-(p+1)=165$ from which it can be derived that $n=169$ (as $p=3$).

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

Continue on page 40

Exercise XIII

A young professional chess player is scheduled to play 20 games the coming month. The games can be assumed to be independent, and the player is assumed to have the same probability of winning each game. Let X denote the number of games the player wins the coming month.

Question XIII.1 (28)

What is the appropriate statistical model/distribution for X?

1* ☐ A binomial distribution

2 ☐ A χ^2 distribution

3 ☐ An F distribution

4 ☐ A hypergeometric distribution

5 ☐ A Poisson distribution

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

Recall that the binomial(n, p) distribution counts the number of successes in n independent trials, where the probability of a success in each trial is p. Since there are 20 games in the coming month, and the games are assumed independent with an identical probability of the player winning each game, the appropriate model is a binomial model with $n = 20$ and p equal to the common probability of the player winning in a game.

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

Question XIII.2 (29)

The young player has achieved the following results over the last three years:

Year	Wins	Draws	Losses
2023	43	78	33
2022	25	55	22
2021	34	46	41

The player now wants to test whether the result distribution has changed over the years, i.e. test the null hypothesis that the proportions of wins, draws, and losses are the same across the three years.

When performing the usual hypothesis test of the null hypothesis, what distribution should be used to calculate the critical value?

1* ☐ A $\chi^2(4)$ distribution

2 ☐ A $\chi^2(9)$ distribution

3 ☐ An $\mathrm{F}(2, 2)$ distribution

4 ☐ An $\mathrm{F}(3, 3)$ distribution

5 ☐ An $\mathrm{F}(2, 6)$ distribution

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

The data is given in a 3×3 frequency table, and the usual test of the null hypothesis is presented in method 7.22. In equation (7-56), the critical value is given as $\chi^2_{1-\alpha}((r-1)(c-1))$, and since $r = c = 3$ in this example, the critical value comes from a $\chi^2(4)$ distribution.

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

Question XIII.3 (30)

The player and a friend compare their performance ratings across seven tournaments they both participated in. At each tournament they both get a performance rating based on their results, and they assume that their performances at different tournaments are independent. However, looking at their performance ratings given in the below table, they notice that their performances seem to depend on the level of the tournaments (local, national, or international).

Tournament	Local 1	Local 2	Local 3	Local 4	National 1	National 2	International 1
Player	2219	2248	2311	2256	2175	2140	2025
Friend	2144	2169	2341	2222	2088	2055	1979

Assuming normality of observations, which of the following is the most appropriate statistical test for comparing the mean performances of the player and the friend?

1 ☐ A Welch two-sample t-test

2* ☐ A paired t-test

3 ☐ A pooled two-sample t-test

4 ☐ A one-way ANOVA with two treatments (groups)

5 ☐ A χ^2-test

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

Since their performance ratings depend on the level of the specific tournaments (for instance, they both perform the worst at international 1 and perform the best at local 3), the appropriate test must take into account that the level of the tournaments vary. This can be achieved by considering their performances ratings as paired observations and conducting a paired t-test or similarly a two-way ANOVA.

Both the Welch two-sample t-test and the pooled two-sample t-test (which is equivalent to a one-way ANOVA with two treatments) consider the samples as completely independent and fail to correct for the varying levels of the tournaments. Finally, the χ^2-test is used when comparing proportions and not means.

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

The exam is finished. Enjoy the summer!