Written examination: 16. December 2023

Course name and number: **Statistics (Polytechnical Foundation) (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 12 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	II.2	III.1	III.2	III.3	III.4	III.5	IV.1
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer	5	2	2	4	3	5	4	5	3	5

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

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

The exam paper contains 35 pages.

Continue on page 2

Multiple choice questions: *Note that in each question, one and only one of the answer options is correct. Furthermore, not all the suggested answers are necessarily meaningful. Always remember to round your own result to the number of decimals given in the answer options before you choose your answer. Also remember that there may be slight discrepancies between the result of the book's formulas and corresponding built-in functions in R.*

Exercise I

A population is exponentially distributed with rate $\lambda = 2$.

Question I.1 (1)

Which of the following statements is true?

1 ☐ The probability of obtaining an observation between 1 and 2 in a random draw can be calculated in R by: `pexp(1, rate=2) - pexp(2, rate=2)`

2 ☐ The probability of obtaining an observation below 1 in a random draw can be calculated in R by: `pexp(2, rate=2)`

3 ☐ The probability of obtaining an observation below 1 in a random draw can be calculated in R by: `1 - pexp(1, rate=2)`

4 ☐ The probability of obtaining an observation above 3 in a random draw can be calculated in R by: `dexp(3, rate=2)`

5* ☐ None of the above statements are correct

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

The four answers are wrong:

- It the wrong order of subtraction, so a negative probability
- It's below 2
- It's above 1
- It's the density function at 3

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

Question I.2 (2)

According to the central limit theorem (CLT) what type of distribution approximates the mean of a random sample with $n = 100$ observations from the population (note that CLT does not say anything particular about the sample size $n = 100$)?

1 ☐ The standard normal distribution

2* ☐ A normal distribution (that is not also a standard normal distribution)

3 ☐ An exponential distribution

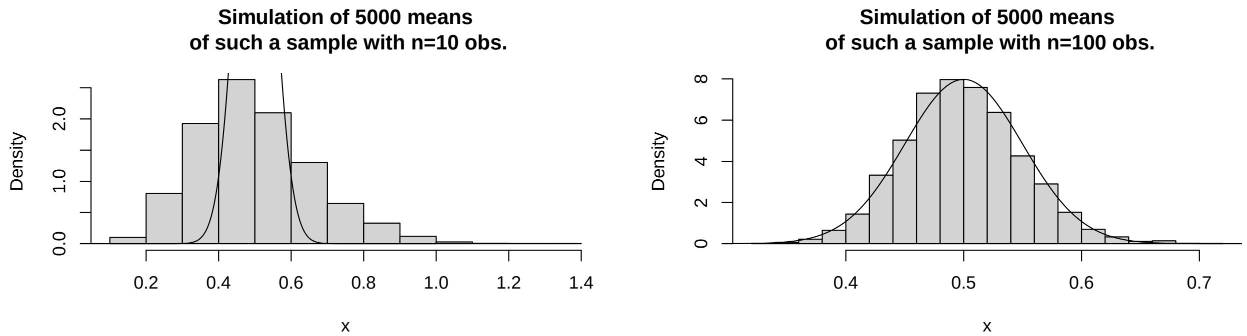
4 ☐ A Poisson distribution

5 ☐ An F-distribution

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

According to the CLT the sum, thus the also the mean, of random variables with similar distribution can be well approximated with a normal distribution. As a thumb rule when $n > 30$. Of course if the population is extremely skewed this might not hold, but for an exponential distribution it works well.

Below follows a simulation: to the left with $n = 10$ observations in the sample the approximation is not so well, but to the right with $n = 100$ the approximation is very good.



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

Continue on page 4

Exercise II

A one-way ANOVA model has been fitted to some data from a balanced experiment (an equal number of observations for each treatment). The ANOVA table from the analysis is given below, where some numbers are replaced by letters.

Source	DF	SS	MS	Test statistic	p-value
Treatment	9	207	D	E	0.03
Residual	50	B	C		
Total	A	707			

Question II.1 (3)

Which set of values is consistent with the ANOVA table?

1 ☐ $A = 59$, $B = 914$, and $D = 23$

2* ☐ $A = 59$, $C = 10$, and $E = 2.3$

3 ☐ $A = 450$, $D = 23$, and $E = 2.3$

4 ☐ $B = 500$, $C = 23$, and $D = 10$

5 ☐ $B = 914$, $C = 10$, and $E = 23$

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

The missing values are calculated as

$$A = 50+9 = 59, \quad B = 707-207 = 500, \quad C = \frac{B}{50} = \frac{500}{50} = 10, \quad D = \frac{207}{9} = 23, \quad E = \frac{D}{C} = \frac{23}{10} = 2.3.$$

Therefore, answer 2 is correct.

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

Question II.2 (4)

Two specific treatments are then compared in the post hoc analysis. What is the least significant difference between the two treatment means using a 5% significance level?

1 ☐ 2.841

2 ☐ 3.060

3 ☐ 3.199

4* ☐ 3.667

5 ☐ 4.130

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

The formula from remark 8.13 is used:

$$\text{LSD}_\alpha = t_{1-\alpha/2}(n-k)\sqrt{2 \cdot \text{MSE}/m}.$$

Since there are $n = 60$ observations across the $k = 10$ treatments and the experiment is balanced, there are $m = 6$ observations for each treatment. Thus,

$$\text{LSD}_{0.05} = t_{0.975}(50)\sqrt{2 \cdot 10/6} = 3.667,$$

which means answer 4 is correct.

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

Continue on page 6

Exercise III

Temperature in the indoor environment is an important part of people's well being, and in addition heating is an important part of the energy consumption in houses.

A house owner is considering the indoor temperature in one of the rooms of his house. As a first approach, he decides to analyse the daily average temperature in the room over a period of time. The R-output from his analysis is given below (the vector `temp` contains the daily average temperatures in the room).

```
##
##  One Sample t-test
##
## data:  temp
## t = 160.53, df = 233, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  19.97593 20.47234
## sample estimates:
## mean of x
##  20.22413
```

Question III.1 (5)

How many days did the house owner use for the analysis?

1 ☐ 366

2 ☐ 364

3* ☐ 234

4 ☐ 365

5 ☐ 233

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

Degrees of freedom is equal $n-1$ and hence $n = 233 + 1 = 234$ days included.

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

Question III.2 (6)

The house owner wants to test the hypotesis that the mean temperature in the room is 20 °C against the alternative that the mean temperature is different from 20 °C. What is the usual p-value for this hypothesis test?

1 ☐ $< 2.2 \cdot 10^{-16}$

2 ☐ 0.375

3 ☐ 0.0382

4 ☐ 0.137

5* ☐ 0.0765

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

In order to calculate the p-value we need the stanard error. We have the observed t-test statistics for the hypothesis the $\mu = 0$ that is

$$t_{obs} = \frac{\bar{x}}{se} \tag{1}$$

hence $se = \bar{x}/t_{obs}$ is

```
(se <- 20.22413/160.53)

## [1] 0.1259835
```

and the p-value for the hypothesis is

```
2*(1-pt((20.22413-20)/se,df=233))

## [1] 0.07653644
```

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

The house owner would also like to analyse the variation over time. In order to do so, he decides to test whether or not the mean temperature at a specific time of day is constant over time. Formally, he does this by testing the hypothesis that the temperature at that time of day can be assumed to be the same in two different months. The output of the analysis is given below (the test statistics have been replaced by `Q`):

```
## Welch Two Sample t-test
##
## data:  temp1 and temp2
## t = Q, df = 53.627, p-value = 0.9793
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9278637  0.9040722
## sample estimates:
## mean of x mean of y
##  19.10497  19.11686
```

Here `temp1` and `temp2` are vectors with the temperatures in the two different months.

Question III.3 (7)

Is there a significant difference in the average temperature between the two months on a significance level $\alpha = 0.05$?

1 ☐ Yes, since $0.979 > 0.95$

2 ☐ Yes, since $0 \notin [19.10, 19.11]$

3 ☐ No, since $0.904 > 0.05$

4* ☐ No, since $0.979 > 0.05$

5 ☐ No, since $0 \notin [19.10, 19.11]$

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

From the R-output the p-value can be directly seen as 0.9793, which is greater than 0.05 and hence there is not a significant difference.

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

Question III.4 (8)

Suppose we instead had used the (unprovided) test statistic `Q` for testing if there is a significant temperature difference between the two months. What are the critical values using a significance level $\alpha = 0.01$?

1 ☐ ± 1.832

2 ☐ ± 1.960

3 ☐ ± 2.005

4 ☐ ± 2.398

5* ☐ ± 2.671

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

We should use the 99.5% quantile from the t distribution with the right degrees of freedom, which can be determined from the R output above.

```
qt(0.995, df = 53.627)

## [1] 2.670662
```

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

Question III.5 (9)

The house owner now wants to test if there is a difference between two specific days, while taking the hour of day into account. He therefore considers a paired t-test for the comparison.

If X_i and Y_i denote the outcomes from the two samples used in the paired t-test, which of the following statements about the assumptions of the statistical model is correct?

We use the notation $V[X_i] = \sigma_X^2$, $V[Y_i] = \sigma_Y^2$, and $V[X_i - Y_i] = \sigma_{X-Y}^2$ for the variances, and μ_X, μ_Y for the means of the two samples, and μ for the difference in means.

1 ☐ $X_i \sim N(\mu, \sigma_X^2)$ and $Y_i \sim N(\mu, \sigma_Y^2)$ where both are i.i.d. and independent of each other

2 ☐ $X_i - Y_i \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$ and is i.i.d.

3* ☐ $X_i - Y_i \sim N(\mu, \sigma_{X-Y}^2)$ and is i.i.d.

4 ☐ $X_i - Y_i \sim N(0, \sigma_{X-Y}^2)$ and is i.i.d.

5 ☐ $X_i \sim N(\mu_X, \sigma_X^2)$, $Y_i \sim N(\mu_Y, \sigma_Y^2)$ where both are i.i.d. and independent of each other

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

The assumption is that the differences are Gaussian and i.i.d., this exclude answer options 1 and 5, there is no assumption that the mean equal 0 (that is often the hypothesis), this exclude option 4. The variance is usually not equal $\sigma_X^2 + \sigma_Y^2$, but actually usually smaller, and in any case it is not the assumption and hence we can exclude answer 2. Hence the only option left is 3 where the mean and variance is just defined as some number (the mean μ is actually equal $\mu_X - \mu_Y$), and that is also the correct one in this case, where the difference is considered as a one-sample situation.

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

Continue on page 11

Exercise IV

An energy trading company wants to learn about the electricity price in a particular area for a particular period. They download data from the market and calculate the daily electricity price and relevant weather variables. The following variables are in the data set:

- `Price`: the electricity price in the market
- `Cloudcover`: cloud cover (in %)
- `Humid`: relative humidity
- `Temperature`: temperature
- `Windspeed`: wind speed

```
summary(lm(Price ~  Cloudcover + Humid + Temperature + Windspeed))

##
## Call:
## lm(formula = Price ~ Cloudcover + Humid + Temperature + Windspeed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30525 -0.04983  0.02637  0.07770  0.18326
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4419418  0.1080436   4.090 0.000139 ***
## Cloudcover   0.0003513  0.0006310   0.557 0.579901
## Humid        0.0003016  0.0010300   0.293 0.770754
## Temperature  0.0098091  0.0041229   2.379 0.020784 *
## Windspeed   -0.0529552  0.0127183  -4.164 0.000109 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1112 on 56 degrees of freedom
## Multiple R-squared:  0.3757,Adjusted R-squared:  0.3311
## F-statistic: 8.427 on 4 and 56 DF,  p-value: 2.146e-05
```

Question IV.1 (10)

How many days are included in the data set?

1 ☐ 51

2 ☐ 55

3 ☐ 56

4 ☐ 57

5* ☐ 61

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

The number of observations is equal to the degrees of freedom plus the number of parameters in the model.

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

Question IV.2 (11)

What is the result of the first backward selection step on the model when using a significance level $\alpha = 0.05$ (both the conclusion and the argument must be correct)?

1* ☐ `Humid` should be removed, since $0.771 > 0.580 > 0.05$

2 ☐ `Windspeed` should be removed, since it has the highest uncertainty (not counting `Intercept`)

3 ☐ `Windspeed` and `Temperature` should be removed, since $0.00011 < 0.05$ and $0.021 < 0.05$

4 ☐ `Humid` and `Cloudcover` should be removed, since $0.771 > 0.05$ and $0.580 > 0.05$

5 ☐ None of the variables should be removed, since the t values are all numerically greater than $t_{\text{crit}} = 2.003$

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

In a backward selection step in-significant inputs, i.e. with a p-value above the significance level (0.05), one removes the variable with the highest p-value, but only one variable shall be removed in a single step.

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

Question IV.3 (12)

Disregarding any conclusion about a potential model reduction, which of the following conclusions can be drawn for the market at the particular period with the estimated result?

1 ☐ The estimate of the mean price in the period is 0.4419

2 ☐ When the temperature increases, the price decreases, and when the wind speed increases, the price increases

3 ☐ The 99% prediction interval for the mean price has the width $2 \cdot 0.111$

4 ☐ The model can be used to predict the mean value of the wind speed in the period

5* ☐ The model can explain 37.6% of the observed variation in the price in the period

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

The `Multiple R-squared:` measured the explained variation by the model.

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

Continue on page 14

Exercise V

This exercise contains questions related to supermarkets.

Question V.1 (13)

Back in the days, the cashiers in the supermarket entered the prices manually on the cash register. When employees were tired, they would often make errors when entering the prices. Assume that for a particular situation, they randomly made a price error for 5% of the costumers. Assume independence of the price enterings.

What is the probability that 10 or more out of 100 customers would experience a price error?

1 ☐ 0.0015

2 ☐ 0.0043

3* ☐ 0.028

4 ☐ 0.063

5 ☐ 0.55

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

```
1 - pbinom(9, 100, 0.05)

## [1] 0.02818829
```

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

Question V.2 (14)

In a study of a supermarket, the arrival rate of customers is assumed to be 200 customers/hour in the peak hours. Customers arrive according to a Poisson process. If more than 250 customers arrive in an hour, the store's capacity will be exceeded. What is the probability that the store's capacity is not exceeded during a peak hour?

1 ☐ 0.00028

2 ☐ 0.00061

3 ☐ 0.51879

4 ☐ 0.92470

5* ☐ 0.99972

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

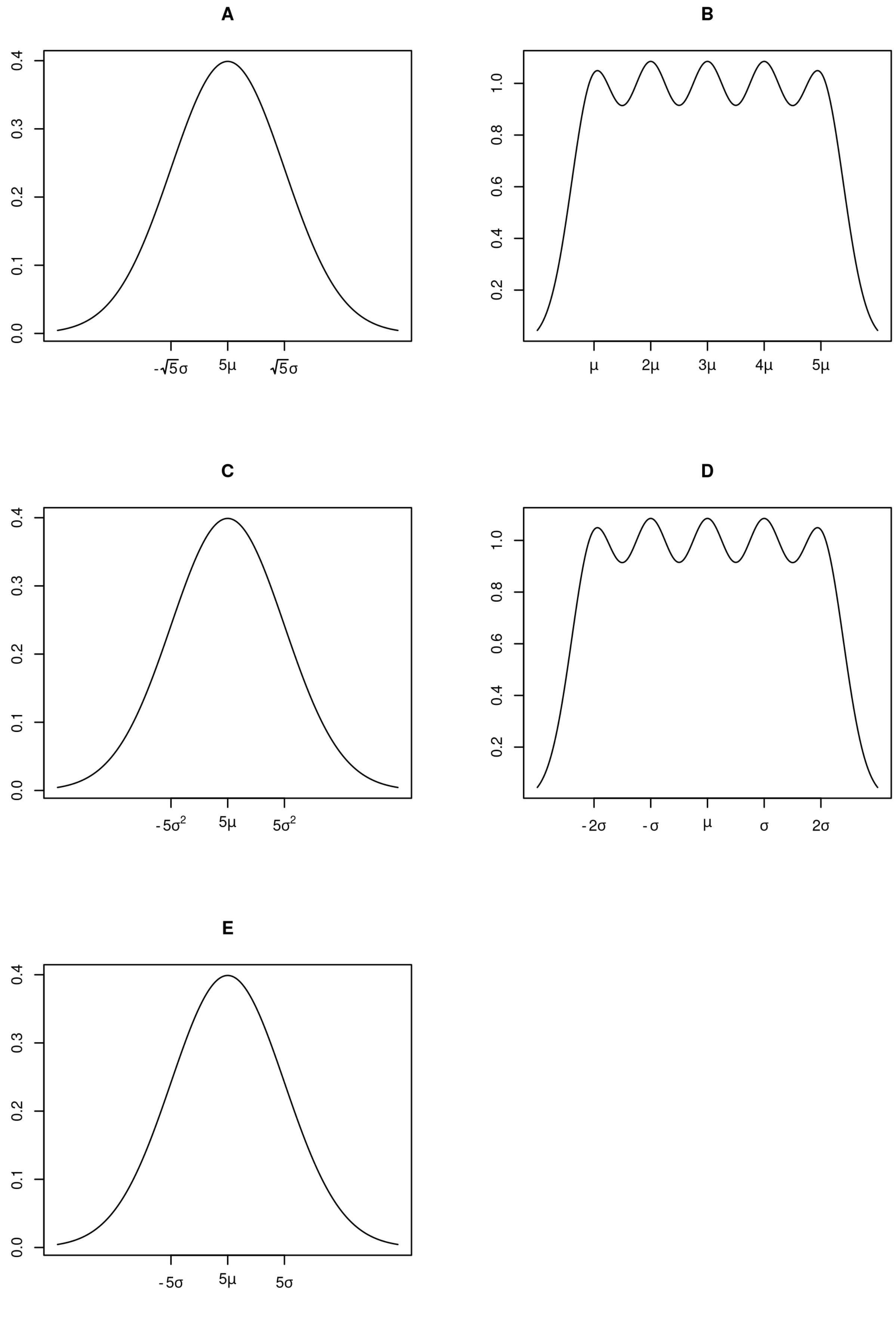
```r
ppois(250, 200)

## [1] 0.9997154
```

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

Question V.3 (15)

Let $X \sim N(\mu, \sigma^2)$ denote the average daily turnover in a particular supermarket store. The store was open 5 days a week, and it can be assumed that the daily turnovers are independent between days.

One of the following plots show the probability density of the weekly turnover. Which one?


A
0.0
0.1
0.2
0.3
0.4
-√5σ
5μ
√5σ
B
0.2
0.4
0.6
0.8
1.0
μ
2μ
3μ
4μ
5μ
C
0.0
0.1
0.2
0.3
0.4
-5σ²
5μ
5σ²
D
0.2
0.4
0.6
0.8
1.0
-2σ
-σ
μ
σ
2σ
E
0.0
0.1
0.2
0.3
0.4
-5σ
5μ
5σ

1* ☐ A

2 ☐ B

3 ☐ C

4 ☐ D

5 ☐ E

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

The weekly turn over is the sum of 5 daily turn overs and is therefore also normal distributed.

Find the mean and standard deviation of the weekly turn over using the identities of mean and variance

$$\mathrm{E}(X+X+X+X+X) = \mathrm{E}(X)+\mathrm{E}(X)+\mathrm{E}(X)+\mathrm{E}(X)+\mathrm{E}(X) = \mu+\mu+\mu+\mu+\mu = 5\mu$$
$$\mathrm{V}(X+X+X+X+X) = \mathrm{V}(X)+\mathrm{V}(X)+\mathrm{V}(X)+\mathrm{V}(X)+\mathrm{V}(X) = \sigma^2+\sigma^2+\sigma^2+\sigma^2+\sigma^2 = 5\sigma^2$$

Then we find the plot where the standard deviation is marked at $\sqrt{5}\sigma$

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

Continue on page 18

Exercise VI

Let X and Y be two independent exponentially distributed random variables with rates 1.2 and 1.7, respectively.

Question VI.1 (16)

We are interested in the probability that $X + Y$ is greater than 3. Use simulation to assess which of the values below is the correct result. We recommend that you use at least 10000 simulations.

1* ☐ 0.078

2 ☐ 0.120

3 ☐ 0.344

4 ☐ 0.645

5 ☐ 0.920

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

The correct answer is 0.07827. The probability can be estimated using the following r code (result should be close to 0.07827)

```
x <- rexp(10000, rate = 1.2)
y <- rexp(10000, rate = 1.7)
mean(x + y > 3)

## [1] 0.0792
```

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

Continue on page 19

Exercise VII

The Yellow Duck racing team is testing the performance of different tyre compounds on a specific race track. The team's two drivers and the reserve driver have each driven a single lap on each of the six different tyre compounds, and all the laps were completed using the same car under identical weather conditions. The lap times can be found in the below table (Note: the lap time '1:40.391' reads 1 minute and 40.391 seconds).

Tyre compound	C_0	C_1	C_2	C_3	C_4	C_5
Driver 1 lap time	1:40.391	1:41.506	1:42.241	1:43.058	1:43.766	1:44.801
Driver 2 lap time	1:40.495	1:41.455	1:42.468	1:43.350	1:44.230	1:45.391
Reserve driver lap time	1:40.617	1:41.623	1:42.750	1:43.617	1:44.411	1:45.346

The lap times (measured in seconds) can be read into R using the following code chunk:

```
Time_Driver_1 <- c(100.391,101.506,102.241,103.058,103.766,104.801)
Time_Driver_2 <- c(100.495,101.455,102.468,103.350,104.230,105.391)
Time_Reserve  <- c(100.617,101.623,102.750,103.617,104.411,105.346)
```

The engineers at Yellow Duck racing team fit a two-way ANOVA model to the data:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

with the usual assumptions. In the model, the α-parameters are driver effects, and the β-parameters are tyre compound effects.

Question VII.1 (17)

What are the parameter estimates $\hat{\alpha}_{\text{Reserve}}$ and $\hat{\beta}_{C_1}$ according to the model?

1 ☐ $\hat{\alpha}_{\text{Reserve}} = -0.235$ and $\hat{\beta}_{C_1} = -2.361$

2 ☐ $\hat{\alpha}_{\text{Reserve}} = -0.235$ and $\hat{\beta}_{C_1} = -1.334$

3 ☐ $\hat{\alpha}_{\text{Reserve}} = 0.036$ and $\hat{\beta}_{C_1} = 2.317$

4 ☐ $\hat{\alpha}_{\text{Reserve}} = 0.199$ and $\hat{\beta}_{C_1} = -2.361$

5* ☐ $\hat{\alpha}_{\text{Reserve}} = 0.199$ and $\hat{\beta}_{C_1} = -1.334$

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

The data is set up in R as follows:

```r
Driver   <- as.factor(c(rep(1,6),rep(2,6),rep("R",6)))
Compound <- as.factor(rep(0:5,3))
Time     <- c(Time_Driver_1,Time_Driver_2,Time_Reserve)
```

We can find the estimated effects of the drivers and the compounds as:

```r
mu <- mean(Time)
alpha <- tapply(Time,Driver,mean) - mu
beta <- tapply(Time,Compound,mean) - mu
```

The estimated effects are thus $\hat{\alpha}_{\text{Reserve}} = 0.199$ and $\hat{\beta}_{C_1} = -1.334$, which corresponds to answer 5.

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

Question VII.2 (18)

The model supports which of the following conclusions at a 5% significance level?

1 ☐ Neither the driver effect nor the compound effect is significant

2 ☐ The driver effect is significant, while the compound effect is not significant

3 ☐ The driver effect is not significant, while the compound effect is significant

4* ☐ Both the driver effect and the compound effect are significant

5 ☐ The significance of the effects cannot be determined with this data

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

The ANOVA table associated with the model is found in R as:

```r
Results  <- data.frame(Driver,Compound,Time)
Model    <- lm(Time~Driver+Compound,data=Results)
anova(Model)

## Analysis of Variance Table
##
## Response: Time
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Driver     2  0.576  0.2878  15.278 0.0009115 ***
```

```
## Compound   5 44.152  8.8304 468.810 1.614e-11 ***
## Residuals 10  0.188  0.0188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the p-values from the output, the analysis supports the conclusion that both the driver effect and the compound are significant at a 5% significance level. Hence, answer 4 is correct.

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

Question VII.3 (19)

The racing team performs post hoc pairwise comparisons between all three drivers by calculating confidence intervals using an overall significance level of 5%. What is the confidence interval for the difference between the reserve driver and Driver 2? (The effect of the reserve driver minus the effect of Driver 2)

1 ☐ $[-0.557, 0.882]$

2 ☐ $[-0.159, 0.484]$

3 ☐ $[-0.118, 0.443]$

4* ☐ $[-0.065, 0.390]$

5 ☐ $[-0.014, 0.339]$

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

The approach described in section 8.3.3 is invoked. Since the overall significance level is 5%, the Bonferroni corrected significance level that should be applied for the three differences is 5%/3. Using the parameter estimates and the mean square error from the ANOVA table yields:

$$\hat{\alpha}_{\text{Reserve}} - \hat{\alpha}_{\text{Driver 2}} \pm t_{1-(0.05/3)/2}((3-1)(6-1))\sqrt{0.0188\left(\frac{1}{6}+\frac{1}{6}\right)} = [-0.065, 0.390] \,.$$

In conclusion, answer 4 is correct.

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

Continue on page 22

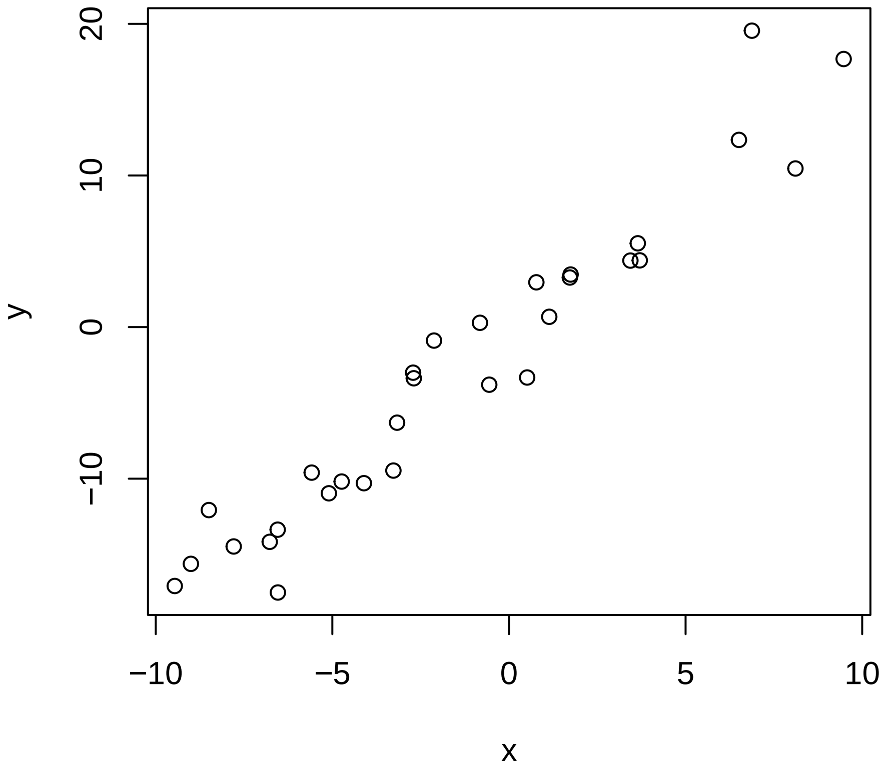
Exercise VIII

The simple linear regression model is given by

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \, ,$$

where $\varepsilon_i \sim N(0, \sigma^2)$ and are independent, $i = 1, \ldots, n$.

A sample of the two paired variables are stored in R in the vectors `x` and `y`.A scatter plot of the variables is seen below:



The simple linear regression model is fitted, and the result is printed below. Note that some of the values have been replaced by letters:

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9599 -1.4571  0.1936  1.4127  7.2499
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.43369    0.49844   -0.87    0.392
## x                  A    0.09284   19.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.636 on 28 degrees of freedom
## Multiple R-squared:  0.9342,Adjusted R-squared:  0.9319
## F-statistic: 397.8 on 1 and 28 DF,  p-value: < 2.2e-16
```

Question VIII.1 (20)

Which of the following values should replace `A` in the result (hint: looking at the figure can also be of help)?

1 ☐ -0.73

2 ☐ 0.73

3* ☐ 1.85

4 ☐ 9.46

5 ☐ 20.15

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

We can clearly see from the points that the line that will be fitted with the simple linear regression model will have a slope around 2, since when we e.g. go from 0 to 10 on the x-axis we to around 20 on the y-axis. Hence, 1.8 is the only meaningful value.

It also possible to calculate backwards from the t value using Method 5.12.

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

Question VIII.2 (21)

Which of the following calls in R calculates the width of the 99% confidence interval for β_0?

1* ☐ `2*qt(0.995,28) * 0.49844`

2 ☐ `2*qt(0.995,28) * 0.09284`

3 ☐ `qt(0.995,27) * 0.49844`

4 ☐ `qt(0.95,28) * 0.09284`

5 ☐ `qt(0.99,28) * 0.43369`

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

We use the formula from the parameter confidence intervals method in Chapter 5

$$\hat{\beta}_0 \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_0}$$

where the $\hat{\sigma}_{\beta_0}$ is the value from the result under `Std Error` and multiply the $\pm$ term with 2 to get the width.

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

Continue on page 25

Exercise IX

The Danish Health Authority (DHA) is designing a survey to examine the drinking habits of young adults in Denmark. Specifically, the DHA wants to estimate the proportion of young adults in Denmark that drink more than the maximum recommended units of alcohol in an average week. The DHA wants the estimate to be within 0.01 of the true proportion with 95% probability.

Question IX.1 (22)

What is the minimum number of young adults that should be included in the survey to achieve the desired precision (we refrain from making any assumptions about true proportion)?

1 ☐ 2401

2 ☐ 4147

3* ☐ 9604

4 ☐ 16588

5 ☐ 38415

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

We apply formula (7-25) from method 7.13. Since there is no information about the true proportion, we apply the formula for the worst case scenario with $\alpha = 0.05$ and $\mathrm{ME} = 0.01$:

$$n \geq \frac{1}{4}\left(\frac{z_{0.975}}{0.01}\right)^2 = 9603.647.$$

Therefore, the DHA should include at least 9604 young adults in the sample.

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

Question IX.2 (23)

In a previous study including 400 young adults, statisticians from the DHA accepted the null hypothesis $\mathcal{H}_0 : p = 0.25$ at a 10% significance level. What is the least possible estimate of the proportion that the statisticians could have found in the study?

1 ☐ $0.00\% = 0/400$

2 ☐ $20.75\% = 83/400$

3* ☐ $21.50\% = 86/400$

4 ☐ $23.25\% = 93/400$

5 ☐ $25.00\% = 100/400$

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

Since the null hypothesis $\mathcal{H}_0 : p = 0.25$ was accepted at a 10% significance level, the p-value has to be greater than or equal to 10%. The p-value can be calculated from formula (7-21) in method 7.11, and hence:

$$2\mathbb{P}(Z > |z_{obs}|) \geq 0.1 \Leftrightarrow \mathbb{P}(Z \leq |z_{obs}|) = \Phi(|z_{obs}|) \leq 0.95 \Leftrightarrow |z_{obs}| \leq \Phi^{-1}(0.95) \approx 1.645.$$

Using (7-16) from method 7.11 gives the following equation and values for x:

$$|z_{obs}| = \left| \frac{x - 0.25 \cdot 400}{\sqrt{0.25 \cdot 0.75 \cdot 400}} \right| = \left| \frac{x - 100}{\sqrt{75}} \right| \leq \Phi^{-1}(0.95) \Leftrightarrow x \in [85.755, 114.245] \, .$$

Thus, the lowest possible number of respondents who said they drink more than the maximum recommended units of alcohol in an average week must be 86. Consequently, the least possible estimate that the statisticians could have found is $86/400 = 21.5\%$. Thus, answer 4 is correct.

Notice that if you calculate a 90% confidence interval around 0.25, you would reach the same conclusion (when rounding up to the nearest fraction of 400).

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

Continue on page 27

Exercise X

As part of a study on adaptive learning platforms, 47 students volunteered to try a new teaching method for the entire semester. The students' performances were tested by a pretest before the semester and a posttest after the semester.

Pretest scores are stored in `pretest` and posttest scores are stored in `posttest`. Both are ordered by student number.

Question X.1 (24)

The following code was run:

```
sum(pretest)

## [1] 1620.042

quantile(pretest, probs = c(0.25, 0.5,  0.75))

##      25%      50%      75%
## 16.66667 30.00000 53.33333
```

Which of the following statements can be concluded about the pretest scores:

1 ☐ The mean of the pretest scores is 30

2 ☐ The median of the pretest scores is 34.5

3* ☐ The IQR of the pretest scores is 36.7

4 ☐ The standard deviation of the pretest scores is 16.7

5 ☐ None of the above

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

The IQR of a sample is $q_{0.75} - q_{0.25}$, which in this case is 36.7 when rounded.

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

Question X.2 (25)

We wish to compare the students' pretest and posttest performances by using the mean change in test scores (posttest minus pretest) as a target.

Which of the following code snippets correctly computes a 95% confidence interval for this quantity using non-parametric bootstrapping?

1 ☐
```
sim_mean_diff <- replicate(1000,
                             mean(sample(posttest, 20, replace = TRUE)) -
                             mean(sample(pretest, 20, replace = TRUE)))
quantile(sim_mean_diff, c(0.025, 0.975))
```

2* ☐
```
sim_mean_diff <- replicate(1000,
                   mean(sample(posttest - pretest, 20, replace = TRUE)))
quantile(sim_mean_diff, c(0.025, 0.975))
```

3 ☐
```
t.test(posttest, pretest , paired = FALSE, conf.level = 0.95)$conf.int
```

4 ☐
```
t.test(posttest, pretest, paired = TRUE, conf.level = 0.95)$conf.int
```

5 ☐
```
t.test(posttest, pretest, paired = TRUE, conf.level = 0.975)$conf.int
```

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

Snippets 3-5 use a t-test and hence do not use non-parametric bootstrap. The setup in this experiment is two samples, paired, which the second snippet does (the first snippet is an unpaired two-sample).

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

Question X.3 (26)

As a result of the previous question, the researchers got the confidence interval $[7.9, 17.2]$.

Which of the following statements can be concluded?

1* ☐ The mean posttest result is significantly higher than the mean pretest result on a 5% significance level

2 ☐ The mean pretest result is significantly higher than the mean posttest result on a 5% significance level

3 ☐ There is not a significant difference between the mean pretest and posttest results on a 5% significance level

4 ☐ There is a linear relationship between prestest and posttest results

5 ☐ None of the above

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

Since 0 is not contained in the 95% confidence interval, there is a significant difference in mean test results. As this difference is positive, it is option 1 that is correct.

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

Continue on page 30

Exercise XI

A hospital took blood samples from 469 randomly selected people of different age and screened the samples for a specific chemical. The results of the screenings are given in Table 1 below:

Table 1	Age group 1	Age group 2	Age group 3	Age group 4	Total
Chemical not detected	17	28	21	15	81
Chemical detected	73	138	105	72	388
Total	90	166	126	87	469

The data used to construct table 1 can be read into R using:

```
table1 <- matrix(c(17,28,21,15,73,138,105,72),nrow=2,byrow=TRUE)
```

Question XI.1 (27)

Under the null hypothesis that the probability of a sample having traces of the chemical is the same across the different age groups, what is the expected number of samples without traces of the chemical taken from people in age group 3?

1 ☐ 20.25

2* ☐ 21.76

3 ☐ 26.30

4 ☐ 28.67

5 ☐ 104.24

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

The estimate of the common proportion of samples having traces of the chemical is found using (7-42) as

$$\hat{p} = \frac{388}{469},$$

which implies that the expected number of samples without traces of the chemical is

$$e_{13} = n_3(1 - \hat{p}) = 126\left(1 - \frac{388}{469}\right) = 21.76,$$

cf. (7-43). Answer 2 is therefore correct.

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

Question XI.2 (28)

Which of the following is the correct conclusion when testing the null hypothesis that the probability of a sample having traces of the chemical is the same across the different age groups at a 5% significance level (both the argument and the conclusion must be correct)?

1 ☐ The p-value is 0.025 and the null hypothesis is therefore rejected

2 ☐ The p-value is 0.025 and the null hypothesis is therefore accepted

3 ☐ The p-value is 0.975 and the null hypothesis is therefore rejected

4* ☐ The p-value is 0.975 and the null hypothesis is therefore accepted

5 ☐ The p-value is 0.975 and the test is therefore inconclusive

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

Method 7.22 applies, which can be done in R as:

```
table1 <- matrix(c(17,28,21,15,73,138,105,72),nrow=2,byrow=TRUE)
chisq.test(table1,correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  table1
## X-squared = 0.21606, df = 3, p-value = 0.975
```

Since the p-value is much larger than the chosen significance level, the null hypothesis is accepted. Hence, answer 4 is correct.

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

Question XI.3 (29)

The samples that had traces of the chemical were further subdivided as shown in Table 2 below:

Table 2	Age group 1	Age group 2	Age group 3	Age group 4	Total
Type A detected	35	64	42	20	161
Type B detected	30	60	55	45	190
Type C detected	8	14	8	7	37
Total	73	138	105	72	388

The data used to construct table 2 can be read into R using:

```
table2 <- matrix(c(35,64,42,20,30,60,55,45,8,14,8,7),nrow=3,byrow=TRUE)
```

Consider now only the samples with traces of the chemical. The hospital staff would like to test for independence between the type of chemical detected in a sample and the age group of the person who submitted the sample. Assuming the hospital invokes a 90% confidence level, which of the following statements is correct?

1* ☐ The observed test statistic is 10.177 and it should be compared with χ_{crit}, where χ_{crit} is the 90% quantile of a χ^2 distribution with 6 degrees of freedom

2 ☐ The observed test statistic is 10.177 and it should be compared with χ_{crit}, where χ_{crit} is the 90% quantile of a χ^2 distribution with 8 degrees of freedom

3 ☐ The test rejects the null hypothesis of independence at the chosen significance level

4 ☐ The test is invalid as some of the calculated expected values are less than 5

5 ☐ Under the null hypothesis, the probability of observing a test statistic less than 10.177 is 11.74%

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

Method 7.24 applies, which can be done in R as:

```
table2 <- matrix(c(35,64,42,20,30,60,55,45,8,14,8,7),nrow=3,byrow=TRUE)
chisq.test(table2,correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  table2
## X-squared = 10.177, df = 6, p-value = 0.1174

chisq.test(table2,correct=FALSE)$expected

##          [,1]     [,2]     [,3]      [,4]
## [1,] 30.29124 57.26289 43.56959 29.876289
## [2,] 35.74742 67.57732 51.41753 35.257732
## [3,]  6.96134 13.15979 10.01289  6.865979
```

The above output shows that all the calculated expected values are greater than five, which means that the test is valid. The output also shows that the observed test statistic is 10.177, which should be compared with the critical value

$$\chi^2_{1-\alpha}((r-1)(c-1)) = \chi^2_{1-0.1}((3-1)(4-1)) = \chi^2_{0.9}(6) = 10.64.$$

As the observed test statistic is less than the critical value, the null hypothesis of independence is accepted. The same conclusion could have been reached by noticing that the p-value (which shows that under the null hypothesis, the probability of observing a test statistic greater than 10.177 is 11.74%), is above the significance level of 10%. The only correct statement is therefore answer 1.

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

Continue on page 34

Exercise XII

Question XII.1 (30)

Bertil and Karin have collected data as part of their bachelor thesis, and as part of this, they are studying the relationship between two variables, `height` and `time`.

They wish to apply a linear regression, but cannot agree on how to correctly check the model assumptions. Only one of the statements below is correct. Which one?

1 ☐ Non-parametric bootstrapping of the residuals would reveal if the assumptions of linear regression are met

2 ☐ A histogram of the `height` values would reveal if the normality assumption is met

3 ☐ The value of the coefficient of determination (R^2) would reveal if the linearity assumption is met

4 ☐ A boxplot of the `time` values would reveal if the normality assumption is met

5* ☐ A QQ plot of the residuals would reveal if the normality assumption is met

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

If the assumptions of linear regression is met, the residuals should be normally distributed, which can be assessed using a QQ plot.

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

Continue on page 35

The exam is finished. Enjoy the Christmas break!