Technical University of Denmark *Page 1 of 41 pages.*

Written examination: 13. December 2016

Course name and number: **Introduction to Statistics (02323 and 02402)**

Aids and facilities allowed: All

The questions were answered by

_____ _____ _____

(student number) (signature) (table number)

There are 30 questions of the "multiple choice" type included in this exam divided on 18 exercises. To answer the questions you need to fill in the prepared 30-question multiple choice form (on three seperate pages) in CampusNet

5 points are given for a correct 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 another answer is given, then it does not count (i.e. "0 points"). Hence, if more than one answer option is given to a single question, which in fact is technically possible in the online system, it will not count (i.e. "0 points"). The number of points corresponding to specific marks or needed to pass the examination is ultimately determined during censoring.

The final answers should be given in the exam module in CampusNet. The table sheet here is ONLY to be used as an "emergency" alternative (remember to provide your study number if you hand in the sheet).

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

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

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

The questionnaire contains 41 pages.

Continues on page 2

Multiple choice questions: *Note that not all the suggested answers are necessarily meaningful. In fact, some of them are very wrong but under all circumstances there is one and only one correct answer to each question.*

Exercise I

Archaeopteryx is a genus of bird-like dinosaurs that is transitional between non-avian feathered dinosaurs and modern birds. Assume that we have data from 6 fossils of Archaeopteryx including measurements of the length of the thigh bone (femur) and the upper arm bone (humerus) as shown in the table below.

Femur	38	46	56	59	64	74
Humerus	41	50	63	70	71	76

Data can be loaded into R by:

```
femur = c(38,46,56,59,64,74)
humerus = c(41,50,63,70,71,76)
```

Question I.1 (1)

Archaeologists have long believed that there should be a linear relationship between the length of the femur and length of humerus of extinct animals such as Archaeopteryx. What conclusion can be made by analyzing the above data when the significance level $\alpha = 0.05$ is used?

1 ☐ There is reason to assume a linear relationship, as the length of bones in animals are always positively correlated.

*2 ☐ There is reason to assume a linear relationship, with p-value for the relevant test being 0.0013.

3 ☐ There is reason to assume a linear relationship, with p-value for the relevant test being 0.0780.

4 ☐ There is no reason to assume a linear relationship, with p-value for the relevant test being 0.0033.

5 ☐ There is no reason to assume a linear relationship, with p-value for the relevant test being 0.0780.

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

We need to test if there is a significant correlation, which we can do by typing in the values in R and fit a simple linear regression model. This we can do by testing if the slope (β_1) is significantly different from zero, i.e.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

which is equivalent to the hypothesis

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

where ρ is the correlation, see Section 5.6. We type in the numbers in R and fit a simple linear regression model

```
femur <- c(38,46,56,59,64,74)
humerus <- c(41,50,63,70,71,76)
summary(lm(humerus ~ femur))

##
## Call:
## lm(formula = humerus ~ femur)
##
## Residuals:
##      1      2      3      4      5      6
## -2.106 -1.353  1.338  5.246  1.092 -4.217
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9332     7.3951   0.532  0.62299
## femur         1.0309     0.1289   7.998  0.00133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.693 on 4 degrees of freedom
## Multiple R-squared:  0.9411,Adjusted R-squared:  0.9264
## F-statistic: 63.96 on 1 and 4 DF,  p-value: 0.001325
```

We find that the p-value for the test is 0.00133, which is much below $\alpha = 0.05$. Thus the null hypothesis is rejected and we conclude that there is a linear relationship.

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

Continues on page 4

Exercise II

A study aims to investigate whether intake of a natural product affects weight. The study should include 10 subjects (men with similar weight). The weight change D_i $(i = 1, \ldots, 10)$ after one month of use of the natural product is recorded. It is of interest to test if the weight change can be assumed to be zero, i.e. to test the hypothesis $H_0 : \mu_D = 0$ against the alternative $H_1 : \mu_D \neq 0$. It is decided to apply the significance level $\alpha = 0.05$.

Question II.1 (2)

Assuming that the standard deviation of weight change is $\sigma = 1$ kg, what is the power for detecting an actual weight change of at least 1 kg? (Hint: the function `power.t.test` in R can be useful here.)

1 ☐ 50.0%

2 ☐ 69.3%

*3 ☐ 80.3%

4 ☐ 89.7%

5 ☐ 99.3%

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

In order to find the power of the test to detect an actual change of at least 1 kg, then the recommended R function can be used, we just need to give it the four parameters, see Example 3.67:

- Sample size $n = 10$
- Change to detect $\delta_0 = 1$
- Assumed standard deviation of the population $\sigma = 1$
- Significance level $\alpha = 0.05$

Further, we have to tell it that it is a one-sample test and the alternative is two-sided.

Then the function calculates the power

```
power.t.test(n=10, delta=1, sd=1, sig.level=0.05, type="one.sample", alternative="two.si
```

```
##
##      One-sample t test power calculation
##
##               n = 10
##           delta = 1
##              sd = 1
##       sig.level = 0.05
##           power = 0.8030962
##     alternative = two.sided
```

which is 80.3%.

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

Continues on page 6

Exercise III

It is believed that the amount of cholesterol in chicken eggs, X, is normally distributed with mean $\mu = 200$ mg and standard deviation $\sigma = 15$ mg, i.e. $X \sim N(200, 15^2)$.

Question III.1 (3)

What is the proportion of chicken eggs having an amount of cholesterol higher than 205 mg?

1 ☐ $P(X > 205) = 0.631$

*2 ☐ $P(X > 205) = 0.369$

3 ☐ $P(X > 205) = 0.491$

4 ☐ $P(X > 205) = 0.394$

5 ☐ $P(X > 205) = 0.605$

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

We need to calculate the proportion (or the probability of drawing a random egg from the population) above 205. Remember

$$P(X > 205) = 1 - P(X \leq 205)$$

where $P(X \leq 205)$ is the cumulated distribution function (cdf) for a normal distribution and that we can get from R by

```
1 - pnorm(q=205, mean=200, sd=15)
```

```
## [1] 0.3694413
```

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

Question III.2 (4)

Industrial kitchens may buy cartons of eggs, where the content, Y, in a carton corresponds to the combined content of 100 eggs, i.e. the total content of cholesterol in a carton is $Y = \sum_{i=1}^{100} X_i$. The content of the 100 eggs can be assumed independent from each other.

You buy a carton of eggs, corresponding to buying 100 eggs. Which of the following R commands gives the probability that the total cholesterol, Y, is higher than 20.5 g (note that 200 mg is 0.2 g)?

1 ☐ `pnorm(q=100*0.205, mean=100*0.200, sd=100*0.015)`

2 ☐ `1-pnorm(q=100*0.200, mean=100*0.205, sd=sqrt(100*0.015*0.015))`

3 ☐ `pnorm(q=100*0.205, mean=100*0.200, sd=100*100*0.015)`

*4 ☐ `1-pnorm(q=100*0.205, mean=100*0.200, sd=sqrt(100*0.015*0.015))`

5 ☐ `pnorm(q=100*0.205, mean=100*0.200, sd=sqrt(100*0.015*0.015))`

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

We need to use the Theorem 2.54: Mean and variance of linear combinations. We need to find the mean

$$\begin{aligned}\mathrm{E}(Y) &= \mathrm{E}\left(\sum_{i=1}^{100} X_i\right) \\ &= \mathrm{E}(X_1 + X_2 + \cdots + X_n) \\ &= \mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n) \\ &= 100 \cdot \mathrm{E}(X) = 100 \cdot 200 \text{ mg} = 100 \cdot 0.200 \text{ g}\end{aligned}$$

and the variance

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{Var}\left(\sum_{i=1}^{100} X_i\right) \\ &= \mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) \\ &= 100 \cdot \mathrm{Var}(X) = 100(\cdot 15 \text{ mg})^2 = 100 \cdot (0.015 \text{ g})^2\end{aligned}$$

which then gives the standard deviation

$$\sigma_Y = \sqrt{100 \cdot (0.015 \text{ g})^2} = \sqrt{100 \cdot 0.015 \cdot 0.015} \text{ g}$$

Finally, the probability we need to calculate is

$$P(Y > 20.5) = 1 - P(Y \leq 20.5)$$

which in R is written as: `1-pnorm(q=100*0.205, mean=100*0.200, sd=sqrt(100*0.015*0.015))`.

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

Continues on page 8

Exercise IV

We consider a binomial random variable Y where $n = 100$ and $p = 0.45$.

Question IV.1 (5)

Calculate $P(Y > 40)$:

1 ☐ 0.183

2 ☐ 0.971

3 ☐ 0.420

*4 ☐ 0.817

5 ☐ 0.866

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

We need first to remember

$$P(Y > 40) = 1 - P(Y \leq 40)$$

Since we know the parameters of the distribution, we can find this probability in R as

```
1 - pbinom(q=40, size=100, prob=0.45)

## [1] 0.8169431
```

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

Question IV.2 (6)

We define a new random variable X so that $X = k \cdot Y$, where the constant k is given by $k = 2$ and Y is binomial distributed random variable with $n = 100$ and $p = 0.45$. Please state the variance of the random variable X:

1 ☐ $\text{Var}(X) = \text{Var}(k \cdot Y) = k + n \cdot p(1-p) = 26.75$

2 ☐ $\text{Var}(X) = \text{Var}(k \cdot Y) = k^2 \cdot n^2 \cdot p^2(1-p)^2 = 49.50^2$

3 ☐ $\text{Var}(X) = \text{Var}(k \cdot Y) = k^2 \cdot n^2 \cdot p(1-p) = 9900$

*4 ☐ $\text{Var}(X) = \text{Var}(k \cdot Y) = k^2 \cdot n \cdot p(1-p) = 99.00$

5 ☐ $\text{Var}(X) = \text{Var}(k \cdot Y) = k \cdot n \cdot p(1-p) = 49.50$

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

We need to use the Theorem 2.54, combined with Theorem 2.21. This gives us a formula for variance of a linear function of a random variable and the variance of a binomial distributed random variable

$$\text{Var}(X) = \text{Var}(k \cdot Y) = k^2\text{Var}(Y) = k^2 \cdot n \cdot p(1-p) = 2^2 \cdot 100 \cdot 0.45 \cdot (1-0.45) = 99.$$

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

Continues on page 10

Exercise V

We consider an exponentially distributed random variable X with parameter β. The distribution function is given by $F(X \leq x) = 1 - e^{-x/\beta}$, where $x > 0$ and $\beta > 0$. Please note that the mean value of X equals β.

Question V.1 (7)

Please state the median of X:

1 ☐ The median of X becomes $0.5 \cdot 2 \cdot \beta$

2 ☐ The median of X becomes $0.5^2 \cdot \beta$.

*3 ☐ The median of X becomes $\log(2) \cdot \beta$ (where log is the natural logarithm)

4 ☐ The median of X becomes $\log(\frac{1}{2}) \cdot \beta^2$ (where log is the natural logarithm)

5 ☐ The median of X becomes $2 \cdot \beta$

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

The median is the quantile where exactly half of the probability mass is below, so we can set the cdf equal to 0.5 and then solve for x

$$\begin{aligned} P(X \leq x) &= 0.5 \Leftrightarrow \\ 1 - e^{-x/\beta} &= 0.5 \Leftrightarrow \\ e^{-x/\beta} &= 0.5 \Leftrightarrow \\ \frac{1}{e^{x/\beta}} &= 0.5 \Leftrightarrow \\ e^{x/\beta} &= \frac{1}{0.5} \Leftrightarrow \\ \frac{x}{\beta} &= \log 2 \Leftrightarrow \\ x &= \log 2 \cdot \beta. \end{aligned}$$

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

Continues on page 11

Exercise VI

A biologist is interested in examining the effects of three different diets (A, B, C) for cultivating tiger shrimps. She purchases 24 uniform larvae from a hatchery for the experiment. Each larva is placed in its own container, and it is determined by random which diet that should be given, such that each diet is tested on 8 different larvae. The larvae grow to become tiger shrimps, and after completing the study period the weight of the shrimps is measured, Y_{ij} (in grams). Since the weight can be assumed to follow a normal distribution, the following model is applied

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

In the model α_i describes the effect of diet i ($i = 1, 2, 3$). Finally, μ is the average and ε_{ij} is the model residuals which are assumed normally distributed with mean 0 and standard deviation σ_ε. An ANOVA of the above model is given below, and it is seen that diet is statistically significant.

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
Diet       2  44.67 22.3350  8.9221 0.001568 **
Residuals 21  52.57  2.5033
```

Question VI.1 (8)

Beforehand there was an interest in comparing the mean value of diet A and diet C. Their estimated mean values are $\hat{\mu}_A = 12.7251$ and $\hat{\mu}_C = 15.7251$, respectively. Please provide a 95 % confidence interval for the mean difference in weight between diet A and diet C.

1 ☐ $-3.000 \pm 2.119\sqrt{52.488^2(\frac{1}{12} + \frac{1}{12})}$

2 ☐ $-3.000 \pm 1.960\sqrt{64.624(\frac{1}{12} + \frac{1}{12})}$

3 ☐ $-3.000 \pm 1.960\sqrt{44.670(\frac{1}{3} + \frac{1}{3})}$

*4 ☐ $-3.000 \pm 2.080\sqrt{2.503(\frac{1}{8} + \frac{1}{8})}$

5 ☐ $-3.000 \pm 2.080\sqrt{52.570(\frac{1}{4} + \frac{1}{4})}$

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

We must calculate a pre-planned confidence interval as described in Method 8.9:

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)},$$

We can read most values from the ANOVA R output, and the t-quantile can be found as

```
qt(p=0.975, df=21)

## [1] 2.079614
```

When inserting values into the formula it becomes the one shown in answer 4.

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

Continues on page 13

Exercise VII

The exercise is no longer a part of the curriculum

Question VII.1 (9)

The question is no longer a part of the curriculum

Continues on page 14

Exercise VIII

A study aims at comparing the wear resistance of two different kinds of rubber (A and B) used as material for shoe soles. The study includes 100 school children aged 8-10 years. Each child receives a pair of shoes where the sole of one shoe is made of material A, while the sole of the other shoe is made of material B. For each pair of shoes, it is decided by randomization whether material A should be on shoe to the right or to the left. The children use the shoes every day for 3 months, and after the experiment has been completed the wear (in mm) on each shoe is measured.

Question VIII.1 (10)

If it can be assumed that the measured wear is continuous and normally distributed for each kind of rubber, please specify which of the following statistical tests should be applied, if you want to test whether the materials A and B are equal with respect to wear:

1 ☐ A contingency table analysis

2 ☐ An F test comparing two variances

3 ☐ A usual (non-paired) t-test

*4 ☐ A paired t-test

5 ☐ A one-way ANOVA

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

We must test for a difference in mean between two groups, hence a two-sample t-test. Now the question is if it is a paired setup or not. Since each children have both a sole in Material A and a sole in Material B, and each pair is exposed to the same wear (although there could be a difference for each child between right and left, however this is compensated by randomizing the left and right material). Thus, this makes a paired setup and we can take the difference between the soles for each child and use a one-sample test. Hence, we should use a paired t-test. See Section 3.2.3 for more information.

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

Question VIII.2 (11)

It turns out that the null hypothesis is accepted, i.e. it is concluded that the two materials wear out equally. Instead the researchers calculate for each child in the study the average wear for the pair of shoes. It is of interest to analyze, using a standard t-test, if boys and girls wear the shoes equally, or alternatively, if there is a difference in wear between gender (two-sided test). A total of 50 girls and 50 boys were included in the experiment.

As the wear can be assumed normally distributed within each gender with equal variance, we get the usual test statistics $t_{obs} = 2.23$ with 98 degrees of freedom. The p-value becomes:

1 ☐ The p-value becomes 0.014

*2 ☐ The p-value becomes $2 \cdot 0.014$

3 ☐ The p-value becomes $2 \cdot 0.05$

4 ☐ The p-value becomes 0.23

5 ☐ The p-value becomes $1 - 0.23$

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

We have the observed statistic t_{obs}, which under the null hypothesis is t-distributed, and the degrees of freedom are given, so we can simply calculate the p-value by

```
2 * (1 - pt(2.23, df=98))

## [1] 0.02802943
```

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

Continues on page 16

Exercise IX

A course at a university is offered each semester typically with more than 300 students taking the exam. Examination results for 280 students who have passed the course at the previous exam is given in the table below. For example, the tables shows that 24 students got the grade 12. The distribution of the 280 grades is considered in the next 4 questions.

Grade	02	4	7	10	12	In total
Count	22	78	84	72	24	280

The data (grades) can be loaded into R by:

```
grades = rep(x=c(2,4,7,10,12), times=c(22,78,84,72,24))
```

Question IX.1 (12)

Use the central limit theorem to determine a 95% confidence interval for the mean grade based on the students who have passed the exam. (It is important in this question that the grades are perceived numerically, eg. 02 corresponds to the number 2, etc.).

1 ☐ [6.51 ; 7.43]

*2 ☐ [6.62 ; 7.32]

3 ☐ [4 ; 10]

4 ☐ [5.12 ; 8.67]

5 ☐ [5.99 ; 8.72]

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

According to the central limit theorem (Theorem 3.14), we know that even if this data is not normal distributed, then if we have $n > 30$ observations, the standardized sample mean follows a standard normal distribution and we can use the t-distribution to calculate a confidence interval. Therefore we load the sample into R and either use the in-built function or calculate the confidence interval using the formula

```
## Read the data
grades = rep(x=c(2,4,7,10,12), times=c(22,78,84,72,24))
## Use the inbuilt function
t.test(grades)
```

```
##
##  One Sample t-test
##
## data:  grades
## t = 38.972, df = 279, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  6.619297 7.323560
## sample estimates:
## mean of x
##  6.971429

## Use the formula
mean(grades) + c(-1,1) * qt(0.975, df=280-1) * sqrt(sd(grades)^2/280)

## [1] 6.619297 7.323560
```

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

Continues on page 18

Question IX.2 (13)

You now want to test whether the proportion of students who have passed the exam with a grade of '7' or higher can be assumed to be 65%, which has been an objective in designing the grading scale. We denote this proportion p_{7+}. From the table in the previous question it is seen that 180 students out of 280 got the grade '7' or higher.

Determine the p-value when we wish to test $H_0 : p_{7+} = 0.65$ against $H_1 : p_{7+} \neq 0.65$:

*1 ☐ 0.8021

2 ☐ $1.745 \cdot 10^{-6}$

3 ☐ $8.725 \cdot 10^{-7}$

4 ☐ 0.5989

5 ☐ 0.4011

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

This is a single proportion hypothesis test. Using Theorem 7.10

$$z_{\text{obs}} = \frac{180 - 280 \cdot 0.65}{\sqrt{280 \cdot 0.65 \cdot 0.35}} = -0.2506,$$

with which we can calculate the p-value by

$$p\text{-value} = 2P(Z > |-0.2506|) = 2 \cdot (1 - P(Z \leq 0.2506)),$$

using R

```
2 * (1 - pnorm(0.2506))

## [1] 0.8021234
```

Or we could simply calculate it directly in R

```
prop.test(x=180, n=280, p=0.65, alternative="two.sided", correct=FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  180 out of 280, null probability 0.65
## X-squared = 0.062794, df = 1, p-value = 0.8021
## alternative hypothesis: true p is not equal to 0.65
```

```
## 95 percent confidence interval:
##  0.5851475 0.6967000
## sample estimates:
##         p
## 0.6428571
```

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

Question IX.3 (14)

A student is interested in analyzing the data in more detail, and run the following code using the 280 grades stored in the vector `grades`

```
k = 100000
samples = replicate(k, sample(grades, replace = TRUE))
simval = apply(samples, 2, sd)
resultater = quantile(simval, c(0.025,0.975))
```

Please state which numerical result that has been calculated in the vector `resultater`:

*1 ☐ A 95% confidence interval for the standard deviation of the grades (non-parametric bootstrap)

2 ☐ A 95% confidence interval for the distribution of the grades (parametric bootstrap)

3 ☐ A 95% prediction of the median of the grades (non-parametric bootstrap)

4 ☐ A 95% prediction for the standard error of the grades (parametric bootstrap)

5 ☐ A 95% confidence interval for 75% percentile of the grades (parametric bootstrap)

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

It is clear that it is a simulation results, namely found using a bootstrapping method. First, the sample is re-sampled simply by drawing randomly from the sample with replacement, thus there is no assumption about the distribution (i.e. non-parametric). Second, the standard deviation is calculated for all the resampled samples, and from these the 2.5% and 97.5% quantiles are found, therefore: A non-parametric 95% confidence interval for the standard deviation has been bootstrapped.

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

Continues on page 20

Question IX.4 (15)

You now want to examine if the distribution of the grades is the same for men and women. The distribution of grades by sex is shown in the table below.

Grades	02	4	7	10	12	In total
Men	14	47	59	47	18	185
Women	8	31	25	25	6	95

Please calculate the expected number of men with the grade '7' in the case where the grade distribution is assumed equal for men and women (i.e. assuming the null hypothesis):

*1 ☐ 55.5

2 ☐ 59

3 ☐ 47.57

4 ☐ 28.5

5 ☐ 42

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

We must calculate the expected value in cell (1,3) under the null hypothesis that the distribution is the same between the genders. According to Method 7.3, the expected proportion of men is

$$\frac{x}{n} = \frac{185}{185 + 95} = 0.6607,$$

which we multiply with the total number of observations with the grade 7

$$(59 + 25) \cdot \frac{185}{185 + 95} = 55.5.$$

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

Continues on page 21

Exercise X

A discrete random variable X, is used to describe the number of events during a time interval. X has the density function on the familiar form: $P(X = x) = \frac{2^x}{x!}e^{-2}$, for $x \geq 0$.

Question X.1 (16)

What is the mean of X?

1 ☐ $\frac{1}{2}$

2 ☐ $\log(2)$ (where log is the natural logarithm)

*3 ☐ 2

4 ☐ π

5 ☐ 2^2

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

We recognize that the distribution used for characterizing number of events per time interval is the Poisson distribution, and we recognize the pdf from Definition 2.27. Then we can see that $\lambda = 2$ and we are asked about the mean, which (see Theorem 2.28) is simply equal to λ.

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

Continues on page 22

Exercise XI

A study aims at comparing cognitive abilities of 3 groups of children. The groups consist of a) children with Tourette's Syndrome (TS), b) children with ADHD and c) children without any of these diagnoses (Control).

In the study, each child is asked to solve a sequence of tasks on a computer and the average reaction time, R_i (milliseconds) is recorded for each child. The study included 17 children with TS, 13 with ADHD and 20 controls, i.e. a total of $n = 50$ children.

When analyzing the data from the experiment it has been assumed that the reaction time R_i is normally distributed for each group with constant variance, σ_E^2. In order to compare if the mean reaction time is the same for the three groups (TS, ADHD and controls) the following ANOVA table is provided

```
Analysis of Variance Table

Response: reactiontime
          Df Sum Sq Mean Sq F value    Pr(>F)
group      A 485848  242924 D          .976e-07 ***
Residuals  B 542563  C
---
```

It is seen, however, that not all numbers are given in the ANOVA table, but some are only shown by the symbols A, B, C and D. These 4 symbols are part of the solution to the next question.

Question XI.1 (17)

Which distribution does the value D follow if the mean reaction time is the same for all three groups (TS, ADHD and Control)?

1 ☐ $F(A, A+B)$ i.e. an F-distribution with degrees of freedom A and $A+B$ from the ANOVA table

2 ☐ $F(C, B)$ i.e. an F-distribution with degrees of freedom C and B from the ANOVA table

*3 ☐ $F(A, B)$ i.e. an F-distribution with degrees of freedom A and B from the ANOVA table

4 ☐ $F(A, C)$ i.e. an F-distribution with degrees of freedom A and C from the ANOVA table

5 ☐ $F(B, A)$ i.e. an F-distribution with degrees of freedom B and A from the ANOVA table

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

It is recognized as a one-way ANOVA, since there is one factor `group`. According to Theorem 8.6 we know that the test-statistic follows an F-distribution under the null hypothesis with the degrees of freedoms in the R output.

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

Continues on page 24

Question XI.2 (18)

What is the conclusion from the ANOVA table if the significance level $\alpha = 0.05$ is applied?

1 ☐ We must reject the hypothesis that the mean reaction time of the control children equals the mean reaction time for children with TS or ADHD.

2 ☐ We must reject the hypothesis that the variance of the reaction time of the control children equals the variance of the reaction time for children with TS or ADHD.

3 ☐ We can prove that the variance of the reaction time is the same for all three groups since the p-value equals $0.976 \cdot 10^{-7}$.

*4 ☐ We must reject the hypothesis that the average reaction time is the same for all three groups since the p-value equals $0.976 \cdot 10^{-7}$.

5 ☐ We can not reject the hypothesis that the average reaction time is the same for all three groups.

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

The null hypothesis is that the mean in equal in the three groups, hence that the mean reaction time is equal in the three groups. The p-value is $0.976 \cdot 10^{-7}$, which is way under $\alpha = 0.05$, and therefore the null hypothesis is rejected.

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

Question XI.3 (19)

Subsequently it is decided to examine whether the age of the children $x_{1,i}$ influences the reaction time, Y_i. Another experiment was conducted in which $n = 12$ children with no diagnosis (control), but with different ages solved the sequence of tasks and the average reaction time was measured. The model $Y_i = \beta_0 + \beta_1 \cdot x_{1,i} + \varepsilon_i$ is applied in order to examine the association between age and reaction time. In the model the residuals ε_i are assumed i.i.d. normally distributed with constant variance, hence $\varepsilon_i \sim N(0, \sigma^2)$. You get the following output for the new experiment:

```
Call:
lm(Reactiontime ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max
-54.520 -35.522   4.268  27.160  51.949
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   933.15     153.23   6.090 0.000117 ***
Age           -41.05      15.36  -2.672 0.023400 *
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 37.72 on 10 degrees of freedom
Multiple R-squared:  0.4166,Adjusted R-squared:  0.3583
F-statistic: 7.141 on 1 and 10 DF,  p-value: 0.0234
```

Using the numbers in the output from the model please provide the test statistics related to the hypothesis $H_0 : \beta_1 = 0$:

*1 ☐ -2.672

2 ☐ 153.23

3 ☐ -41.05

4 ☐ 37.72

5 ☐ 0.4166

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

The observed test statistic t_{obs} for the test if slope β_1 is zero, can be found under `t value` in the summary output in the row of the explanatory variable, here `Age`.

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

Question XI.4 (20)

Using the analysis result from the previous question find the estimate of the correlation coefficient, $\hat{\rho}$, between Reaction time (Y_i) and Age ($x_{1,i}$):

1 ☐ $\hat{\rho} = -\sqrt{0.3583}$

*2 ☐ $\hat{\rho} = -\sqrt{0.4166}$

3 ☐ $\hat{\rho} = \sqrt{0.3583}$

4 ☐ $\hat{\rho} = 0.4166$

5 ☐ $\hat{\rho} = -\sqrt{0.3583/0.4166}$

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

See Section 5.6. We know the relation between the proportion of explained variance and the sample correlation coefficient.

We take the root of the $\hat{r}^2$ value (explained variance) and the sign of the estimated slope, and get

$$\hat{\rho} = \text{sign}(\hat{\beta}_1)\sqrt{\hat{r}^2} = -\sqrt{0.4166}.$$

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

Continues on page 27

Question XI.5 (21)

A critique of the model $Y_i = \beta_0 + \beta_1 \cdot x_{1,i} + \varepsilon_i$ used in the previous question is that it does not account for whether the answer is correct or not on the individual questions, but simply the reaction time is recorded.

It is decided to expanded model to the following: $Y_i = \beta_0 + \beta_1 \cdot x_{1,i} + \beta_2 \cdot x_{2,i} + \varepsilon_i$, where $x_{2,i}$ is the number of correct answers in the sequence of questions (the remaining variables are defined as they were in the previous question). Based on a new study including 12 different children we obtain the results:

```
Call:
lm(Reactiontime ~ Age + Correct)

Residuals:
    Min      1Q  Median      3Q     Max
-39.958 -24.407   6.917  12.897  42.297

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  892.633    123.203   7.245 4.84e-05 ***
Age          -48.104     15.081  -3.190   0.0110 *
Correct        5.310      1.765   3.009   0.0147 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 28.73 on 9 degrees of freedom
Multiple R-squared:  0.5908,Adjusted R-squared:  0.4999
F-statistic: 6.498 on 2 and 9 DF,  p-value: 0.01793
```

Provide the estimates $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\sigma}^2$:

1 ☐ $\hat{\beta}_1 = -48.104$, $\hat{\beta}_2 = -5.310$ and $\hat{\sigma}^2 = 28.73$

2 ☐ $\hat{\beta}_1 = 892.633$, $\hat{\beta}_2 = -48.104$ and $\hat{\sigma}^2 = 28.73$

3 ☐ $\hat{\beta}_1 = 892.633$, $\hat{\beta}_2 = -48.104$ and $\hat{\sigma}^2 = 5.310$

4 ☐ $\hat{\beta}_1 = -48.104$, $\hat{\beta}_2 = 5.310$ and $\hat{\sigma}^2 = 892.633$

*5 ☐ $\hat{\beta}_1 = -48.104$, $\hat{\beta}_2 = 5.310$ and $\hat{\sigma}^2 = 28.73^2$

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

We find the two parameter estimates $\hat{\beta}_1$ (`Age`) and $\hat{\beta}_2$ `Correct` under `Estimate` in the printed results. The estimate of the variance of the error ($\varepsilon_i \sim N(0, \sigma^2)$) at `Residual standard error`.

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

Continues on page 29

Exercise XII

An engineer is studying a process Y which can be expressed by $Y = U/B$. It can be assumed that U and B are independent random variables. The engineer has 20 pairwise measurements of U and B stored as vectors in the statistical program R and these are referred to as `uobs` and `bobs`,respectively.

Question XII.1 (22)

The engineer would like to calculate a 95% confidence interval for the variance of Y, i.e. σ_Y^2 using non-parametric bootstrap. Which of the following chunks of code in R is most appropriate?

1 ☐
```
samples = replicate(10000,sample(uobs/bobs,replace=FALSE))
results = apply(samples,1, var)
quantile(results, c(0.025, 0.975))
```

*2 ☐
```
samples = replicate(10000,sample(uobs/bobs,replace=TRUE))
results = apply(samples,2, var)
quantile(results, c(0.025, 0.975))
```

3 ☐
```
samples = replicate(10000,sample(var(uobs)/var(bobs),replace=TRUE))
results = apply(samples,2, var)
quantile(results, c(0.025, 0.975))
```

4 ☐
```
samples = replicate(10000,sample(uobs/bobs,replace=TRUE))
results = apply(samples,1, var)
quantile(results, c(0.95))
```

5 ☐
```
samples = replicate(10000,sample(uobs/bobs,replace=FALSE))
results = apply(samples,2, var)
quantile(results, c(0.025, 0.975))
```

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

We want to find the code which is right and first we see that they are all non-parametric (since they all use the `sample` command instead of specifying a distribution).

Then we check if `replace=TRUE`, if not then it is not a useful bootstrapping (see Section 4.3): 2, 3 and 4 is fine.

Then we see that 3 is some weird expression with the ratio of variances: so we are left with 2 and 4.

We check 4, and find two problems: the `apply` function is used on dimension 1 (such that it applies the function on the rows of the generated data and not the columns) and only the 95% quantile is calculate on the bootstrapped values.

Finally, we check 2 and find that it calculates the confidence interval correct.

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

Continues on page 31

Question XII.2 (23)

We continue with the problem from the previous question, i.e. we analyze a process Y that can be expressed by $Y = U/B$.

If we assume that $U \sim N(\mu = 35, \sigma^2 = 10^2)$ and $B \sim N(\mu = 50, \sigma^2 = 10^2)$, what is the probability that Y exceeds 1, i.e. please calculate the probability $P(Y > 1)$:

1 ☐ < 0.001

*2 ☐ 0.1444

3 ☐ 0.4701

4 ☐ 0.5298

5 ☐ 0.8556

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

We can write up

$$P(Y > 1) = P(\frac{U}{B} > 1) = P(U > B) = P(U - B > 0) = 1 - P(U - B \leq 0).$$

We know from Theorem 2.40 that a linear function of normal distributed random variables is also normal distributed. With Theorem 2.56 we can calculate the mean of $U - B$

$$\mu_{U-B} = E(U - B) = E(U) - E(B) = 35 - 50 = -15,$$

and the variance

$$\sigma^2_{U-B} = \text{Var}(U - B) = \text{Var}(U) + \text{Var}(B) = 100 + 100 = 200.$$

Hence $U - B \sim N(-15, 200)$ and now we can calculate

$$P(Y > 1) = 1 - P(U - B \leq 0),$$

in R by

```
1 - pnorm(q=0, mean=-15, sd=sqrt(200))

## [1] 0.1444222
```

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

Continues on page 32

Exercise XIII

A research institute wants to estimate a 95% confidence interval for the true proportion p of consumers who are consciously trying to purchase organic foods when shopping. The research institute plans to ask n consumers the question: "Do you consciously chose to buy organic foods when you do your grocery shopping?". Possible answers to this must be "Yes" or "No".

Question XIII.1 (24)

How many independent consumers n must respond to the survey for at 95% confidence interval for the true proportion, p, to not be wider than 0.04 (Hint: As a starting point for the calculations it can be assumed that 50% of consumers will answer 'Yes' to the question)?

1 ☐ $n = \frac{1.96 \cdot \frac{1}{2} \cdot \frac{1}{2}}{0.01} = 49$

2 ☐ $n = (\frac{1.96 \cdot \frac{1}{2} \cdot \frac{1}{2}}{0.02})^2 = 600.25$ i.e. at least 601

*3 ☐ $n = (\frac{1.96^2 \cdot \frac{1}{2} \cdot \frac{1}{2}}{0.02^2}) = 2401$

4 ☐ $n = (\frac{2 \cdot 1.96 \cdot \frac{1}{2} \cdot \frac{1}{2}}{0.02})^2 = 1250.50$ i.e. at least 1251

5 ☐ $n = (\frac{1.96 \cdot \sqrt{\frac{1}{2} \cdot \frac{1}{2}}}{0.01})^2 = 9604$

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

Since we are trying to determine the sample size for a one-proportion test, we can use Method 7.13

$$n = p(1-p)\left(\frac{z_{1-\alpha/2}}{ME}\right)^2 = \frac{1}{2} \cdot \frac{1}{2} \cdot \left(\frac{1.96}{0.02}\right)^2 = 2401,$$

and see that the answer is simply this formula modified with some parts shifted around.

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

Question XIII.2 (25)

The question is no longer a part of the curriculum

Continues on page 33

Exercise XIV

Assume that the number of attempts for the driving test (before it is passed) in a particular municipality can be described by the model $Y = X + 1$, where X is a Poisson distributed random variable with mean $\lambda = 0.4$, i.e. $X \sim Pois(\lambda = 0.4)$.

Question XIV.1 (26)

We now consider the number of attempts among 100 randomly selected individuals who must pass the driving test. What will be the mean μ and variance σ^2 of the total number of attempts $\sum_{i=1}^{100} Y_i$ for everyone passing the exam?

1 ☐ $\mu = 140$ and $\sigma^2 = \sqrt{40}$

2 ☐ $\mu = 140$ and $\sigma^2 = \sqrt{140}$

3 ☐ $\mu = 140$ and $\sigma^2 = 140$

*4 ☐ $\mu = 140$ and $\sigma^2 = 40$

5 ☐ $\mu = 140$ and $\sigma^2 = 40^2$

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

First we calculate the mean of X, which we from Theorem 2.28 is λ. We can then use Theorem 2.54 to find the mean and variance of Y.

$$\mathrm{E}(Y) = \mathrm{E}(X + 1) = \mathrm{E}(X) + 1 = \lambda + 1 = 1.4,$$

and similarly the variance

$$\mathrm{Var}(Y) = \mathrm{Var}(X + 1) = \mathrm{Var}(X) = \lambda = 0.4.$$

Then we use Theorem 2.56 to find the mean and variance of the linear combination of the 100 drivers

$$\mu = \mathrm{E}(\sum_{i=1}^{100} Y_i) = \mathrm{E}(Y_1) + \mathrm{E}(Y_2) + \cdots + \mathrm{E}(Y_{100}) = 100 \cdot 1.4 = 140,$$

and

$$\sigma^2 = \mathrm{Var}(\sum_{i=1}^{100} Y_i) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + \cdots + \mathrm{Var}(Y_{100}) = 100 \cdot 0.4 = 40.$$

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

Continues on page 34

Exercise XV

Assume that the number of traffic accidents per day, X, follows a Poisson distribution. From 200 independent observations, the rate λ has been estimated to $\hat{\lambda} = 1.2$.

Question XV.1 (27)

Please provide a 95% confidence interval for the true rate λ:

1 ☐ $[1.2; 4.8]$

2 ☐ $[0; 4]$

*3 ☐ $1.2 \pm 1.96 \cdot \sqrt{\frac{1.2}{200}}$

4 ☐ $1.2 \pm 1.96 \cdot \frac{1.2}{200}$

5 ☐ $1.2 \pm 1.96 \cdot \frac{1.2^2}{200^2}$

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

Since we have $n = 200 > 30$ then, according to the central limit theorem (Theorem 3.14), we can use the usual confidence interval based on the t-distribution (or standard normal distribution). So we need the sample mean and sample variance, which we using Theorem 2.28 find simply equal to the estimate of the rate for a Poisson distributed random variable

$$\hat{\mu} = \bar{x} = \hat{\lambda} = 1.2,$$
$$\hat{\sigma}^2 = s^2 = \hat{\lambda} = 1.2,$$

which we plug in the formula from Method 3.9

$$\begin{aligned} \bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} &= \bar{x} \pm t_{\alpha/2} \cdot \sqrt{\frac{s^2}{n}} \\ &= 1.2 \pm 1.96 \cdot \sqrt{\frac{1.2}{200}}, \end{aligned}$$

where $t_{\alpha/2} = t_{0.975} \approx z_{0.975}$ is found by

```
qt(p=0.975, df=199)
```

```
## [1] 1.971957
```

```
qnorm(p=0.975)
```

```
## [1] 1.959964
```

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

Continues on page 35

Exercise XVI

Two types of prescription drugs (A and B) to lower blood cholesterol, are compared in a clinical study. In analyzing the data, it was estimated how much drug A reduces cholesterol, denoted Δ_A, and correspondingly how much drug B is reducing cholesterol, denoted Δ_B (both drugs were found to reduce cholesterol and in the following positive values of Δ indicate reduction).

A 95% confidence interval for the difference in reduction ($\Delta_A - \Delta_B$) has been estimated. This interval is $[0.24; 0.50]$ mmol/L.

Question XVI.1 (28)

Which of the following is a reasonable conclusion to the survey?

1 ☐ Drug A reduces cholesterol by 0.24 mmol/L while drug B reduces cholesterol by 0.50 mmol/L

2 ☐ There is 95% probability that drug A is better to lower the cholesterol than drug B for any person

3 ☐ There is 95% probability that drug A will lower cholesterol with at least 39 mmol/L compared to drug B for any person

*4 ☐ There is at least 95% confidence that drug A reduces cholestrol better than drug B

5 ☐ None of the above

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

Lets go through the answers one by one:

- 1: We have a confidence interval for the difference in mean for the two drugs, but we don't know nothing about how much each of drug reduces cholesterol
- 2: We don't know exactly the probability that drug A is better than drug B. Actually, we can only talk about the probability like this before the experiment, i.e. not using the data. We could maybe have an estimate of a probability, but not like this state "there is 95% probability of ..." from values calculated from data
- 3: Same as 2
- 4: This formulation is correct. If we tested the null hypothesis $H_0 : \Delta_A - \Delta_B = 0$, we would find that zero is outside the confidence interval. Therefore we know that the p-value would be below 5% and thus the formulation "There is at least 95% confidence ..." is appropriate (we could also have used 'certainty' instead of 'confidence')
- 5: Since 4 is reasonable conclusion, then this is not correct

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

Continues on page 38

Exercise XVII

An engineer is planning to take a sample from a population. We consider the following three statements:

I. If the sample has variance zero, then the variance in the population is also zero.

II. If the population has variance zero, then the variance in the sample is also zero.

III. If the sample has zero variance, then the mean and median is the same in the sample.

Question XVII.1 (29)

Which of the three above statements are correct?

1 ☐ Only I. and II.

2 ☐ Only I. and III.

*3 ☐ Only II. and III.

4 ☐ I., II., and III. are all correct

5 ☐ None of the above

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

Lets try to verify the statements

- I.: If we take a sample from a discrete variable with multiple outcomes (e.g. a dice rool), then we can easily imagine that we could get a sample with e.g. 6 equal values (a Yatzy!), which would then have a sample variance of zero. However, the population would in this case not have a variance of zero. Hence Statement I. is not correct
- II.: If the population variance is zero, then there is only a single possible outcome value (e.g. a dice with 5 marked on each side), and every sample would be with only that value (e.g. we would roll a 5 every time). In this case the sample variance will always be zero. Hence Statement II. is correct.
- III.: If the sample has zero variance, then all the values in the sample are equal. In this case we can see that the sample mean will also be equal to this value, and also the median, since it is the value in the middle when the sample is ordered. As an example: the sample is (5,5,5,5,5,5), then the sample mean is 5, and the value in the middle (median) is 5. Hence Statement III. is correct.

So only Statement II. and III. are correct.

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

Continues on page 40

Exercise XVIII

It is well known that cuckoos lay their eggs in another bird species nests, and thus leaves the task to raise their offspring to the host bird. Furthermore, it is a theory that cuckoos are able to adapt the size of their eggs depending on the size of the host bird.

To investigate this theory an ornithologists has over a period measured the size (length of the egg) of 10 eggs in each of two different host bird nests, here called the host bird A and B, that is, a total 20 eggs are measured. She gets the estimates 2 mm for the standard deviation of the size for both the host bird A and B.

Question XVIII.1 (30)

It now turns out that the observed difference in size, $\bar{x}_A - \bar{x}_B$, of the eggs is 1 mm. What conclusion does one arrive at when you want to test the hypothesis $H_0 : \mu_A = \mu_B$ against $H_1 : \mu_A \neq \mu_B$, using an ordinary t-test and significance level $\alpha = 5\%$?

1 ☐ The difference in the size of the eggs is statistically significant

*2 ☐ The difference in the size of the eggs is statistically non-significant

3 ☐ One can not conclude anything without stating the actual size of the eggs for A and B

4 ☐ One can not conclude anything without the knowledge of the population size

5 ☐ It is not appropriate to use an ordinary t-test for this analysis

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

It is a two-sample test for the difference in mean, so we use Method 3.49. First we calculate the test statistic

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{1}{\sqrt{4/10 + 4/10}} = 1.25,$$

which we use to calculate the p-value

$$p\text{-value} = 2 \cdot P(T > 1.25)$$

where the degrees of freedom in the t-distribution is

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{\left(\frac{4}{10} + \frac{4}{10}\right)^2}{\frac{(4/10)^2}{9} + \frac{(4/10)^2}{9}} = 18.$$

Thus the p-value is

```r
2*(1 - pt(q=1.25, df=18))

## [1] 0.2273077
```

which the only conclusion fitting this is answer 2.

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

The exam is over. Have a good Christmas vacation!