

Written examination: 28. May 2017

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 11 exercises. To answer the questions you need to fill in the prepared 30-question multiple choice form (on 6 separate pages) in CampusNet. **There is one and only one correct answer to each question.**

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"). Also, if more answers are 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 answer of the exercises should be given by filling in and submitting via the exam module in CampusNet. The table sheet here is ONLY to be used as an "emergency" alternative.

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

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

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

In case of "emergency": Remember to provide your **study number**. The questionnaire contains 43 pages. Please check that your questionnaire contains them all.

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

The island of Moen is the first official Dark Sky area in the Nordic area. A man went there to watch stars and shooting stars. It was a clear night and he had read that he should expect to see 2 shooting stars per minute.

It can be assumed that the intensity of shooting stars is constant and that they arrive independently of each other.

Question I.1 (1)

Which distribution gives the best description of the waiting time in minutes between two shooting stars?

- 1* An exponential distribution with $\lambda = 2$
- 2 A Poisson distribution with $\lambda = 2$
- 3 A Normal distribution with $\mu = 1$ and $\sigma = \sqrt{2}$
- 4 A uniform distribution with $\alpha = 0$ and $\beta = 2$
- 5 A Poisson distribution with $\lambda = 1/2$

————— FACIT-BEGIN —————

As described in Section 2.5.4, waiting times between independent arrivals are best described by an exponential distribution. As seen from Definition 2.27, λ is the average intensity (the number of events per time period), so $\lambda = 2$.

————— FACIT-END —————

Question I.2 (2)

What is the probability that there is no shooting stars in a given 4 minutes interval?

- 1 0.8824969
- 2 0.1353353
- 3 0.9996645

4* 3.3546263×10^{-4}

5 There must have been at least one shooting star in a given 4 minutes interval.

————— FACIT-BEGIN —————

The number of independent events in a given time interval is Poisson distributed. The parameter in the Poisson distribution is the expected number of events which is 2 per minute times 4 minutes = 8 (See equation 2-33 about scaling). So the probability of observing zero can be found as `ppois(0,8)` or `pexp(4,2, lower.tail=FALSE)`.

————— FACIT-END —————

Continues on page 4

Exercise II

When inspecting cars the combustion is tested by measuring the exhaust gasses. Assume that a given car has a true exhaust of particles of 0.12 g/km where the EURO2 norm limit is 0.08 g/km. The measurement has a standard deviation of 0.02 g/km. It is assumed that the measurements are normally distributed around the true exhaust. A car is tested by a single measurement.

Question II.1 (3)

What is the probability that the car is tested as having too high exhaust of particles?

- 1 0.6113513
- 2 0.0227501
- 3 0.3886487
- 4* 0.9772499
- 5 0.6113513

————— FACIT-BEGIN —————

The distribution of the measurements for the car in question is: $X \sim N(0.12, 0.02^2)$ So the answer is:

$$P(X > 0.08)$$

Or in R:

```
1-pnorm(0.08, m=0.12, sd= 0.02)
## [1] 0.9772499
```

————— FACIT-END —————

Continues on page 5

Exercise III

In a purification experiment, the so-called yield was measured after dosing a certain amount of enzyme. The response variable was the yield percentage in relation to the theoretical highest obtainable level (X). Data from 10 different test samples from the experiment were:

x_i
74.7
74.2
74.1
69.6
75.4
76.3
76.7
75.6
72.0
74.3
$\bar{x} = 74.29$
$s = 2.115$

Question III.1 (4)

What is the 80% percentile for these data using the definition from the book?

- 1 74.10
- 2 74.50
- 3* 75.95
- 4 74.29
- 5 75.60

————— FACIT-BEGIN —————

See Definition 1.7. Either we can use R (remember `type=2`, to ensure that the book's definition is used)

```
quantile(x, 0.80, type=2)
```

```
## 80%  
## 75.95
```

Or we do it manually and take the mean of the 8th and 9th observation

```
## Or:  
sort(x)  
  
## [1] 69.6 72.0 74.1 74.2 74.3 74.7 75.4 75.6 76.3 76.7  
  
mean(c(75.6,76.3 ))  
  
## [1] 75.95
```

————— FACIT-END —————

Continues on page 7

Question III.2 (5)

Assuming that $X \sim N(\mu, \sigma^2)$ and applying the usual estimated parameters ($\mu = \bar{x}$ and $\sigma = s$), what is the only statement that can be correct:

- 1 More than 99% of the population is within [72.18, 76.40] (In R: `mean(x) + c(-1, 1) * sd(x)`)
- 2 More than 99% of the population is within [70.06, 78.52] (In R: `mean(x) + c(-1, 1) * 2 *sd(x)`)
- 3 Around 95% of the population is within [72.18, 76.40] (In R: `mean(x) + c(-1, 1) * sd(x)`)
- 4* More than 99% of the population is within [67.95, 80.63] (In R: `mean(x) + c(-1, 1) * 3 *sd(x)`)
- 5 Less than 95% of the population is within [67.95, 80.63] (In R: `mean(x) + c(-1, 1) * 3 *sd(x)`)

————— FACIT-BEGIN —————

In a normal distribution approximately 95% of the population are within the mean plus/minus 2 standard deviations. More than 99% are within plus/minus 3 standard deviations. This means that only answer number 4 can be correct.

```
mean(x)
## [1] 74.29

sd(x)
## [1] 2.114737

qnorm(.995)
## [1] 2.575829

1-2*(1-pnorm(3))
## [1] 0.9973002

mean(x) + c(-1,1)*3*sd(x)
## [1] 67.94579 80.63421

mean(x) + c(-1,1)*2*sd(x)
## [1] 70.06053 78.51947

mean(x) + c(-1,1)*1*sd(x)
## [1] 72.17526 76.40474
```

————— FACIT-END —————

Question III.3 (6)

What is the 95% confidence interval for the mean?

1 $2.262 \pm 74.29 \frac{2.115}{\sqrt{10}} = [-47.42, 51.95]$

2 $74.29 \pm 1.812 \frac{2.115}{\sqrt{9}} = [73.01, 75.57]$

3 $74.29 \pm 1.96 \frac{2.115}{\sqrt{9}} = [72.91, 75.67]$

4* $74.29 \pm 2.262 \frac{2.115}{\sqrt{10}} = [72.78, 75.80]$

5 $74.29 \pm 1.96 \frac{2.115^2}{\sqrt{10}} = [71.52, 77.06]$

————— FACIT-BEGIN —————

As seen from Method 3.9, a one-sample confidence interval has the form

$$\bar{x} \pm t_{0.975} \frac{s}{\sqrt{n}}$$

using $n - 1$ degrees of freedom for the t-distribution. And as

```
qt(0.975, 9)
```

```
## [1] 2.262157
```

it becomes

$$74.29 \pm 2.262 \frac{2.115}{\sqrt{10}}$$

which becomes:

```
74.29 + c(-1, 1)* 2.262*2.115/sqrt(10)
```

```
## [1] 72.77713 75.80287
```

Or completely by R:

```
x <- c(74.7, 74.2, 74.1, 69.6, 75.4, 76.3, 76.7, 75.6, 72.0, 74.3)
t.test(x)
```



```
##  
## One Sample t-test  
##  
## data: x  
## t = 111.09, df = 9, p-value = 1.97e-15  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 72.77721 75.80279  
## sample estimates:  
## mean of x  
## 74.29
```

————— FACIT-END —————

Continues on page 10

Question III.4 (7)

If you make a 95% confidence interval for the standard deviation, which quantiles should then be used?

- 1 `qnorm(0.025)` and `qnorm(0.975)`
- 2* `qchisq(0.025, 9)` and `qchisq(0.975, 9)`
- 3 `qf(0.025, 9, 9)` and `qf(0.975, 9, 9)`
- 4 `qt(0.025, 9)` and `qt(0.975, 9)`
- 5 `qunif(0.025)` and `qunif(0.975)`

————— FACIT-BEGIN —————

As seen from Method 3.19, the CI for a variance is based on the χ^2 -distribution, so 2 is the only possible answer.

————— FACIT-END —————

Question III.5 (8)

The p -value for the hypothesis test of $H_0 : \mu = 70$ is:

- 1* `2*(1-pt(4.29/(2.115/sqrt(10))), 9)`
- 2 `2*(1-pnorm(70/(2.115/sqrt(10))))`
- 3 `(1-pt(-4.29/(2.115/sqrt(9))), 10)`
- 4 `1-pnorm(-4.29/(2.115/sqrt(10)))`
- 5 `1-qt(2.115/4.29, 9)`

————— FACIT-BEGIN —————

As described in Method 3.23, the observed t-test statistic is:

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{74.29 - 70}{2.115/\sqrt{10}} = \frac{4.29}{2.115/\sqrt{10}}$$

And the p -value then is:

$$2 * P(t > |t_{obs}|) = 2 * P(t > \frac{4.29}{2.115/\sqrt{10}})$$

as the t_{obs} is positive in this case. So answer 1 has the proper R-call.

————— FACIT-END —————

Question III.6 (9)

In a new experiment which is in the planning phase, a 95% confidence interval for the mean with an expected width of around 1 is wanted. Assume that the standard deviation is 2.115. How large a sample does it approximately require to achieve this desired precision?

1 5

2 1230

3 4

4 100

5* 69

————— FACIT-BEGIN —————

Use Method 3.63, with $\alpha = 0.05$ and $ME = 0.5$ (half of the confidence interval):

$$n = \left(\frac{z_{1-\alpha/2} \cdot \sigma}{ME} \right)^2 = \left(\frac{1.96 \cdot 2.115}{0.5} \right)^2 = 68.74 \approx 69$$

————— FACIT-END —————

Continues on page 12

Exercise IV

In a purification experiment, two different doses of an enzyme have been investigated called d_1 and d_2 , with the purpose to investigate a possible effect on the yield. The response variable was the yield percentage in relation to the theoretical highest obtainable level. Data from 19 different test samples from the experiment were:

Dose d_1	Dose d_2
74.7	79.6
74.2	77.5
74.1	82.5
69.6	76.7
75.4	78.2
76.3	76.7
76.7	76.6
75.6	78.1
72.0	79.2
74.3	
$\bar{x}_1 = 74.29$	$\bar{x}_2 = 78.34$
$s_1 = 2.115$	$s_2 = 1.898$

The following was run in R:

```
x1 <- c(74.7, 74.2, 74.1, 69.6, 75.4, 76.3, 76.7, 75.6, 72.0, 74.3)
x2 <- c(79.6, 77.5, 82.5, 76.7, 78.2, 76.7, 76.6, 78.1, 79.2)
t.test(x2, x1)

##
## Welch Two Sample t-test
##
## data: x2 and x1
## t = 4.4041, df = 17, p-value = 0.0003878
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 2.112129 5.996760
## sample estimates:
## mean of x mean of y
## 78.34444 74.29000
```

Continues on page 13

Question IV.1 (10)

What is the 99% confidence interval for the difference in means between dose d_2 and dose d_1 ?

- 1 [2.45, 5.66]
- 2* [1.39, 6.72]
- 3 [1.97, 6.14]
- 4 [2.11, 6.00]
- 5 [74.29, 78.34]

————— FACIT-BEGIN —————

From Method 3.47, we see that the 99% two-sample confidence interval has the form

$$\bar{x}_2 - \bar{x}_1 \pm t_{0.995} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

From the R output we can read the degrees of freedom to be 17, rather than calculating it. We use this to find the t -quantile

```
qt(0.995, 17)
```

```
## [1] 2.898231
```

the interval becomes

$$4.054 \pm 2.898 \sqrt{\frac{2.115^2}{10} + \frac{1.898^2}{9}}$$

which becomes:

```
4.054 + c(-1, 1)* 2.898*sqrt(2.115^2/10 + 1.898^2/9)
```

```
## [1] 1.385967 6.722033
```

Or completely by R:

```
t.test(x2, x1, conf.level = 0.99)
```

```
##
```

```
## Welch Two Sample t-test
```

```
##
```

```
## data:  x2 and x1
## t = 4.4041, df = 17, p-value = 0.0003878
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  1.386305 6.722584
## sample estimates:
## mean of x mean of y
##  78.34444  74.29000
```

————— FACIT-END —————

Question IV.2 (11)

The conclusion of the usual t -test (based on $\alpha = 0.05$) for this situation is (both conclusion and argument must be correct):

- 1 The two variances are significantly different as the p -value is small
- 2 The two means are significantly different as the p -value is large
- 3 The two means are approximately equal as the p -value is large
- 4 The two means are approximately equal as the p -value is small
- 5* The two means are significantly different as the p -value is small

————— FACIT-BEGIN —————

From the R output we see that it is a Welsch two-sample test, hence we are comparing whether the two mean values are different or identical. Since the p -value is 0.0004 it is much smaller than $\alpha = 0.05$. This means that we reject the null hypothesis ($H_0 : \mu_a = \mu_b$). Hence, they are significantly different.

————— FACIT-END —————

Continues on page 15

Question IV.3 (12)

A new study with 2 doses is planned. It is assumed that the standard deviation within each group is 2, and that a t -test on level $\alpha = 0.05$ should be carried out. The following things are run in R:

```
power.t.test(n=30, delta=2, sd=2, sig.level=0.05)

##
##      Two-sample t test power calculation
##
##              n = 30
##             delta = 2
##              sd = 2
##      sig.level = 0.05
##             power = 0.9677083
##      alternative = two.sided
##
## NOTE: n is number in *each* group

power.t.test(power=0.80, delta=1, sd=2, sig.level=0.05)

##
##      Two-sample t test power calculation
##
##              n = 63.76576
##             delta = 1
##              sd = 2
##      sig.level = 0.05
##             power = 0.8
##      alternative = two.sided
##
## NOTE: n is number in *each* group
```

Only one of the following statements is true. Which?

- 1 The risk that a study with $n = 30$ in each group does not find a significant difference between the means, if the real difference is 1, is around 97%
- 2 The chance that a study with $n = 30$ in each group finds a significant difference between the means, if the real difference is 1, is around 97%
- 3* The chance that a study with $n = 64$ in each group finds a significant difference between the means, if the real difference is 1, is around 80%
- 4 The risk that a study with $n = 64$ in each group does not find a significant difference between the means, if the real difference is 1, is around 80%
- 5 The risk that a study with $n = 30$ in each group does not find a significant difference between the means, if the real difference is 2, is around 97%

————— FACIT-BEGIN —————

Let us go through the answers:

- 1 A sample size of 30 is expressed in the first of the two tests, however in this test the difference delta that is specified is 2 (and not 1 like the question). Furthermore, power is expressing the probability of rejecting the null hypothesis (NOT the risk of no rejection).
- 2 This time the explanation of power is correct, however the wrong delta is still used for the first test.
- 3 This is the correct answer, since the delta, n and explanation of power is correct.
- 4 Like answer 1, this interpretation of power is incorrect
- 5 This refers to test one, and with the correct delta this time. However, the interpretation of what power is, is still incorrect.

See Section 3.3.2 for more on power.

————— FACIT-END —————

Continues on page 17

Exercise V

In Danish power plants materials are being burned to generate electricity and in this combustion CO₂ is emitted. It's a gas which enhances the Greenhouse effect and therefore contributes to warming up the atmosphere. This has many negative consequences, and it is of interest to reduce these emissions. This is done by introducing more wind and solar energy production into the system.

Each day, CO₂ emissions are calculated (in grams of CO₂ equivalent gas) per kWh electricity produced in Denmark based on data from ENTSO-E about the production.

Column 1 of the table below shows the date. In Column 2, average values of CO₂ emissions are given for the 15 days with the highest wind energy production in the period from December 1, 2016 to April 1, 2017. Column 3 in the table shows electricity generation with coal, this column is not used in the first two questions.

t	co2intensity (gCO ₂ eq/kWh)	coal (MW)
2016-12-02	230	1016
2016-12-25	205	817
2016-12-26	203	746
2017-01-01	212	948
2017-01-05	292	1448
2017-01-12	260	1398
2017-02-08	317	1409
2017-02-12	321	1578
2017-02-21	235	1102
2017-02-22	268	1325
2017-02-23	233	1187
2017-03-01	253	1195
2017-03-02	260	1093
2017-03-16	212	976
2017-03-22	250	1095

The data is read into R in a data.table X and the following is run:

```
## Put observations in x
x <- X$co2intensity
## Number of simulated samples
k <- 100000
n <- length(x)

## Simulation
simsamples <- replicate(k, sample(x, replace = TRUE))
## Calculate the mean of each simulated sample
simmeans <- apply(simsamples, 2, mean)
```

Continues on page 18

```

## Quantiles of the differences gives the CI
quantile(simmeans, c(0.005, 0.995))

##      0.5%      99.5%
## 227.2667 275.2000

quantile(simmeans, c(0.01, 0.99))

##       1%       99%
## 229.3333 272.8000

quantile(simmeans, c(0.025, 0.975))

##      2.5%      97.5%
## 232.3333 269.2667

quantile(simmeans, c(0.05, 0.95))

##       5%       95%
## 235.0000 265.9333

quantile(simmeans, c(0.1, 0.9))

##      10%      90%
## 238.0667 262.4000

```

Question V.1 (13)

A 99% bootstrapped confidence interval for the mean of the CO2-intensity is wanted, without assumptions about the distribution. What is the correct interval?

- 1* [227, 275]
- 2 [229, 273]
- 3 [232, 269]
- 4 [235, 266]
- 5 [238, 262]

————— FACIT-BEGIN —————

We need to identify the interval based on the correct quantiles. It is the 99% confidence interval, so $\alpha = 0.01$ and therefore we need to take the $q_{\alpha/2} = q_{0.005}$ to $q_{1-\alpha/2} = q_{0.995}$ interval, which can be read from the output to be [227, 275].

————— FACIT-END —————

Continues on page 20

Question V.2 (14)

Which one of the following statements is not correct? Each statement is about a null hypothesis for the mean level of CO2 intensity μ_{CO2} at high wind energy production, and the conclusion is drawn based on the results of the R code? (Note again: there are 4 true and 1 false statements - you must find the false statement!)

- 1 The null hypothesis $H_0 : \mu_{CO2} = 200$ would have been rejected on a 10% significance level
- 2 The null hypothesis $H_0 : \mu_{CO2} = 220$ would have been rejected on a 5% significance level
- 3 The null hypothesis $H_0 : \mu_{CO2} = 230$ would have been rejected on a 5% significance level
- 4 The null hypothesis $H_0 : \mu_{CO2} = 270$ would have been rejected on a 5% significance level
- 5* The null hypothesis $H_0 : \mu_{CO2} = 270$ would have been rejected on a 1% significance level

————— FACIT-BEGIN —————

The null hypothesis for a value of the mean is rejected if the value is outside the confidence interval. When looking at the R output above, the first four answers all fall outside the corresponding confidence interval and thus their conclusions are all correct. In the last statement the value 270 falls inside the 90% confidence interval and thus the null hypothesis would not have been rejected.

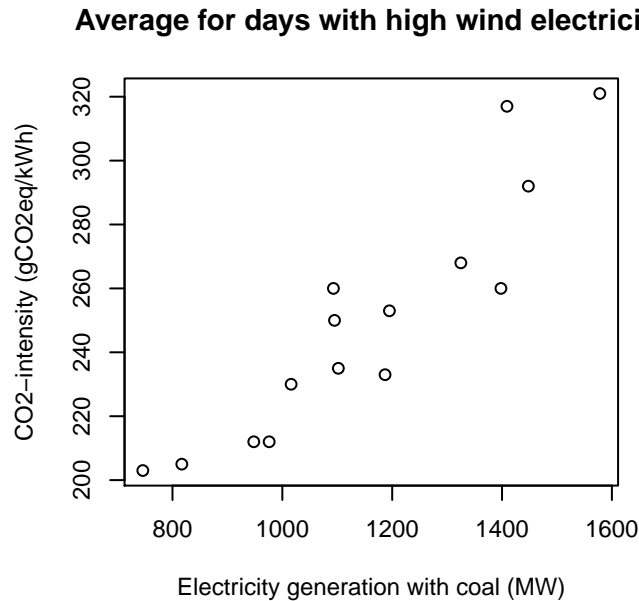
————— FACIT-END —————

Continues on page 21

Question V.3 (15)

In this question, the relationship between the electricity generation with coal (Column 3) and the CO₂-intensity at high wind electricity generation (Column 2) is investigated.

To visualize the relation the following scatter plot with the observations is created:



Based on a consideration of the plot, which of the following statements is the most correct conclusion?

- 1 The correlation is approximately -1.2
- 2 The correlation is approximately 0.1
- 3 The correlation is approximately 0
- 4* The correlation is approximately 0.9
- 5 The correlation is approximately 1.2

————— FACIT-BEGIN —————

The correlation can only be between -1 to 1, and it is certainly not 0, and it is higher than 0.1, so clearly the most correct conclusion is that it is 0.9. In fact it is:

```
cor(X$coal, X$co2intensity)
```

```
## [1] 0.9206539
```

————— FACIT-END —————

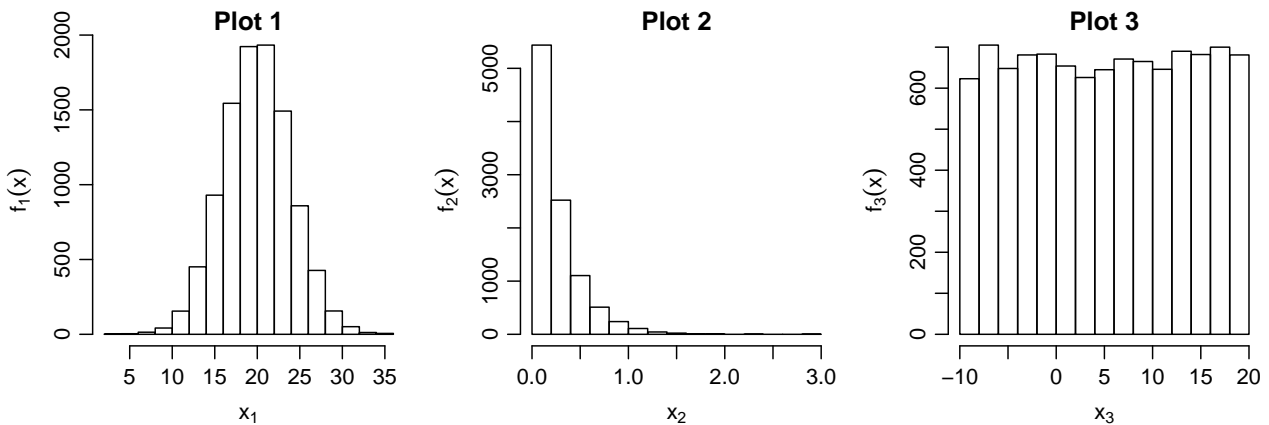
Continues on page 23

Exercise VI

This exercise is about random variables and simulation.

Question VI.1 (16)

The following three histograms are of values from simulations of $n = 10000$ observations from three distributions:



Which of the following distributions are simulated (both the ordering and parameter values must be correct)?

- 1 Plot 1: $X_1 \sim N(10, 4^2)$, Plot 2: $X_2 \sim \text{Exp}(4)$ and Plot 3: $X_3 \sim U(5, 20)$
- 2 Plot 1: $X_1 \sim U(5, 20)$, Plot 2: $X_2 \sim \text{Exp}(1)$ and Plot 3: $X_3 \sim N(20, 4^2)$
- 3 Plot 1: $X_1 \sim U(-10, 20)$, Plot 2: $X_2 \sim \text{Exp}(1)$ and Plot 3: $X_3 \sim N(20, 4^2)$
- 4* Plot 1: $X_1 \sim N(20, 4^2)$, Plot 2: $X_2 \sim \text{Exp}(4)$ and Plot 3: $X_3 \sim U(-10, 20)$
- 5 Plot 1: $X_2 \sim \text{Exp}(4)$, Plot 2: $X_1 \sim N(10, 4^2)$ and Plot 3: $X_3 \sim U(5, 20)$

————— FACIT-BEGIN —————

From the shape of the empirical distribution (the histograms) we identify: the normal distribution in Plot 1 (Example 2.39), the exponential distribution in Plot 2 (Example 2.50), and the uniform distribution in Plot 3 (Figure 2.3). Now we have to identify the correct parameters, which is easiest done in the uniform, which can be seen to go from $\alpha = -10$ to $\beta = 20$.

————— FACIT-END —————

Continues on page 24

Question VI.2 (17)

Let the random variables $X_i \sim N(2, 4^2)$ for $i = 1, \dots, 20$ be i.i.d. and define the following random variables as function of these

$$\begin{aligned}\bar{X} &= \frac{1}{20} \sum_{i=1}^{20} X_i, \\ S &= \sqrt{\frac{1}{19} \sum_{i=1}^{20} (X_i - \bar{X})^2}, \\ Y &= \frac{\bar{X} - 2}{S}.\end{aligned}$$

Which distribution does the random variable Y follow?

- 1 Y follows the normal distribution $N(0, 1^2)$
- 2 Y follows the normal distribution $N(0, 4^2)$
- 3 Y follows the χ^2 -distribution with 20 degrees of freedom
- 4 Y follows t -distribution with 20 degrees of freedom
- 5* Y follows t -distribution with 19 degrees of freedom

————— FACIT-BEGIN —————

Comparing to Theorem 3.3 and Definition 1.11, we identify \bar{X} as the sample mean and S as the sample standard deviation, all as a random variables (capitalized letters). Therefore Y is the S -standardized sample mean, which we know, as stated in Theorem 3.5, follows a t -distribution with $n - 1$ degrees of freedom. From the sample mean formula we can see that there are $n = 20$ observations in the sample, so the t -distribution has 19 degrees of freedom.

————— FACIT-END —————

Continues on page 25

Exercise VII

A recreational runner wants to measure the effect of his training. For this purpose, he has measured values of average pulse (beats per minute), weeks in the training program and speed (km/h), for a particular stretch he runs frequently.

The recreational runner has decided to measure the effect by examining whether the average speed increases over time (weeks).

Data reading in R is:

```
week <- c(1, 1, 1, 3, 3, 4, 5, 5, 5, 5, 6, 6, 7, 7, 8, 9, 9, 10, 12, 12, 13, 13,
          15, 15, 15, 16, 16)

pulse <- c(137.6, 140.1, 143.0, 148.6, 135.6, 139.0, 155.8, 135.0, 149.0, 133.0,
           135.3, 139.8, 137.2, 137.9, 136.8, 134.6, 152.3, 131.9, 137.2, 160.3,
           130.9, 130.9, 131.8, 131.4, 135.6, 138.6, 136.3)

speed <- c(10.01, 10.02, 10.39, 11.86, 9.65, 10.40, 12.60, 9.80, 11.52, 9.59,
           10.26, 10.42, 10.05, 10.48, 10.03, 10.29, 12.22, 10.27, 10.80, 13.79,
           10.40, 9.49, 10.09, 10.34, 11.18, 11.33, 11.34)
```

Initially, parameters in the following model are estimated

$$\text{speed}_i = \beta_0 + \beta_1 \cdot \text{week}_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

Further it is assumed that the ϵ_i 's are independent.

The result of the estimation in R is:

```
summary(lm(speed~week))

##
## Call:
## lm(formula = speed ~ week)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.3909 -0.6058 -0.3407  0.2741  2.9492
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.36053    0.38514  26.901  <2e-16 ***
## week         0.04003    0.04046   0.989   0.332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.008 on 25 degrees of freedom
## Multiple R-squared:  0.03766, Adjusted R-squared:  -0.0008306
## F-statistic: 0.9784 on 1 and 25 DF,  p-value: 0.3321
```

Continues on page 26

Question VII.1 (18)

At the significance level $\alpha = 0.05$ what is the conclusion in relation to increased speed (both argument and conclusion must be correct)?

- 1 There is no significant effect of the training, since $0.0377 < 0.05$
- 2 There is a significant effect of the training, since $26.9 > t_{0.975}$ (where the degrees of freedom for the t -distribution is 25)
- 3* There is no significant effect of the training, since $0.332 > 0.05$
- 4 There is a significant effect of the training, since $0.04 < 0.05$
- 5 There is a significant effect of the training, since $0.989 > 0.95$

————— FACIT-BEGIN —————

The conclusion is that there is no significant effect. This can be based on the p-value for the hypothesis $\beta_1 = 0$ against the two-sided alternative (the p-value is $0.332 > 0.05$). Hence the possibilities are 1 or 3. In 1 the R^2 value is compared with the significance level (which does not make any sense). In 3 the p-value is correctly compared with the significance level. Hence the correct answer is 3. See Method 5.14 for details.

————— FACIT-END —————

Question VII.2 (19)

With the model above what is the 95% confidence interval for β_0 ?

- 1 $10.36 \pm 1.008 \cdot 1.96$
- 2 $10.36 \pm \frac{0.385}{\sqrt{25}} \cdot 2.06$
- 3 $0.040 \pm 0.0405 \cdot 1.96$
- 4 $0.040 \pm \frac{0.989}{25} \cdot 1.96$
- 5* $10.36 \pm 0.385 \cdot 2.06$

————— FACIT-BEGIN —————

Method 5.15 describes how to find confidence intervals for a parameter. The estimate of β_0 is 10.36 and the standard error is given directly as 0.385, further we need the 0.975 quantile of the t -distribution with 25 degrees of freedom, which is can be calculated in R:

```
qt(0.975, 25)
```

```
## [1] 2.059539
```

Hence the correct answer is answer no. 5.

————— FACIT-END —————

Continues on page 28

The recreational runner now decides to investigate a model for the relationship between pulse and velocity (ignoring weeks).

The result of the estimation in R is:

```
summary(lm(speed ~ pulse))

##
## Call:
## lm(formula = speed ~ pulse)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78371 -0.40486 -0.00015  0.35978  0.96661
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.06134     1.82173  -2.778   0.0102 *
## pulse        0.11324     0.01308   8.659 5.39e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5139 on 25 degrees of freedom
## Multiple R-squared:  0.7499, Adjusted R-squared:  0.7399
## F-statistic: 74.98 on 1 and 25 DF,  p-value: 5.388e-09
```

Based on the above model, the recreational runner wants an uncertainty interval for a new run. He uses as the assumption that he can hold an average pulse of 160 beats per minute.

As an aid to the task he has calculated the following number:

```
length(week)

## [1] 27

c(mean(week), var(week))

## [1] 8.222222 23.871795

c(mean(pulse), var(pulse))

## [1] 139.09259 59.38225

c(mean(speed), var(speed))

## [1] 10.689630 1.015396
```

Continues on page 29

Question VII.3 (20)

What is the 95% prediction interval for the speed?

1 $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51 \cdot \sqrt{\frac{1}{27} + \frac{(139.1-160)^2}{27 \cdot 59.38}}$

2 $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51 \cdot \sqrt{1 + \frac{1}{27} + \frac{(139.1-160)^2}{26 \cdot 1.02}}$

3* $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51 \cdot \sqrt{1 + \frac{1}{27} + \frac{(139.1-160)^2}{26 \cdot 59.38}}$

4 $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51^2 \cdot \sqrt{\frac{1}{27} + \frac{(139.1-160)^2}{59.38}}$

5 $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51 \cdot \sqrt{\frac{1}{27} + \frac{(139.1-160)^2}{59.38^2}}$

————— FACIT-BEGIN —————

The general formula for the prediction interval can be seen from Method 5.18 to be

$$\hat{\beta}_0 + \hat{\beta}_1 x_{new} + t_{0.975} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$

where the quantile of the t -distribution is based on $n - 1$ degrees of freedom. In our concrete case we have $\hat{\beta}_0 = -5.06$, $\hat{\beta}_1 = 0.113$, $\hat{\sigma} = 0.51$. These numbers are taken from the summary of `lm`. With $n = 27$ we have $t_{0.975} = 2.06$. x_{new} is equal to 160 in our case, \bar{x} is the average of the measured pulses, i.e. 160. $S_{xx} = \sum_{i=1}^{27} (x_i - \bar{x})^2$ (See Theorem 5.4), which can be calculated by $S_{xx} = 26s_x^2 = 26 \cdot 59.38$, hence the correct answer is no. 3.

————— FACIT-END —————

The recreational runner now decides to estimate a multiple regression model that contains both weeks and pulse:

$$\text{speed}_i = \beta_0 + \beta_1 \cdot \text{week}_i + \beta_2 \cdot \text{pulse}_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

The result from R is:

```
summary(lm(speed ~ week + pulse))  
  
##  
## Call:  
## lm(formula = speed ~ week + pulse)  
##
```

```

## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59154 -0.13508 -0.00055  0.15562  0.43438
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.033531   0.945936  -8.493 1.08e-08 ***
## week         0.094014   0.010414   9.028 3.48e-09 ***
## pulse        0.129052   0.006603  19.545 3.02e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2501 on 24 degrees of freedom
## Multiple R-squared:  0.9431, Adjusted R-squared:  0.9384
## F-statistic: 199 on 2 and 24 DF, p-value: 1.148e-15

```

Continues on page 31

Question VII.4 (21)

At level $\alpha = 0.05$ which of the following statements is correct (both conclusion and argument must be correct)?

- 1 Neither the effect of weeks nor the effect of pulse is significant as $0.094 > 0.05$ and $0.129 > 0.05$
- 2 Both the effect of weeks and the effect of pulse is significant as $0.01 < 0.05$ and $0.0066 < 0.05$
- 3 Since $0.094 < 0.129$ the effect of weeks is significant, while the effect of pulse is not significant
- 4* Both the effect of weeks and the effect of pulse is significant since $3.5 \cdot 10^{-9} < 0.05$ and $3.0 \cdot 10^{-16} < 0.05$
- 5 Neither the effect of weeks nor the effect of pulse is significant since $3.5 \cdot 10^{-9} < 0.05$ and $3.0 \cdot 10^{-16} < 0.05$

————— FACIT-BEGIN —————

Let's go through the answers:

- 1 The number 0.094 and 0.129 are parameter estimates, and it does not make sense to compare those with the significance level hence answer no 1 is wrong.
- 2 In answer 2 standard errors are compared with significance levels, this is also incorrect.
- 3 In answer 3 the magnitude of the parameters are compared and this does not say anything about significance
- 4 In answer 4 the p-values of the parameters are correctly compared with the significance level, also the conclusion is correct (p-values < significance level), hence this is the correct answer
- 5 In answer 5 the p-values are correctly compared with the significance level, however the conclusion is wrong.

————— FACIT-END —————

Question VII.5 (22)

Which statement about the interpretation of the model is correct?

- 1 When weeks increase by 0.094 the pulse increase by 0.129
- 2* For a given pulse the expected speed increase with 0.094km/h per week.
- 3 The model is meaningless since $-8.03 < 0$ and the speed must be positive
- 4 Since the degree of explanation is about 0.25, the model have explained about $3/4$ of the variation
- 5 Since all parameters are significant, it can be seen that all model assumptions are fulfilled

————— FACIT-BEGIN —————

Again we need to go through all the statements.

- 1 When weeks increase with 1 the expected speed increase by 0.094 (for a given pulse). The two explanatory variables do not say anything about how they affect each other, hence this answer is wrong.
- 2 It can be seen from the output that the parameter for week is 0.094, hence this means that the expected speed increases with 0.094 every week, hence this answer is correct.
- 3 The speed cannot be negative but the prediction of -8.03 is only obtained when week=0 and pulse = 0, and a prediction at zero pulse is of course meaningless, but not the model, hence answer no 3 is wrong.
- 4 About $1/4$ of the variation is explained when the degree of explanation is 0.25, hence answer no 4 is wrong.
- 5 Significance of parameters does not tell anything about the model assumptions (we need diagnostics plots for this), hence answer no 5 is wrong

————— FACIT-END —————

Continues on page 33

Exercise VIII

A person has twice evaluated the sharpness (**Sharpness**) for each of 12 different setups (**Treat**) of images on computer screens, ie. 24 observations of sharpness in total split on 12 setups. The scale is a continuous scale from 0 to 15, in practice, done by marking the value on a line.

The result of the usual analysis of variance of these data gave the following R output, however, some of the values are replaced by the letters A, ..., F:

Analysis of Variance Table

```
Response: Sharpness
          Df Sum Sq Mean Sq F value Pr(>F)
Treat      A   93.7      C      E      F
Residuals B   51.9      D
```

Question VIII.1 (23)

What are the values of A and B?

- 1* A=11 and B=12
- 2 A=1 and B=22
- 3 A=93.7/51.9 and B= 51.9/22
- 4 A=12 and B=24
- 5 A=11 and B=23

————— FACIT-BEGIN —————

We need to find the degrees of freedom of a one-way ANOVA as described in Theorem 8.6. As the number of treatments is $k = 12$ and the number of total observations is $n = 24$. Hence $A = k - 1 = 11$ and $B = n - k = 24 - 12 = 12$.

————— FACIT-END —————

Continues on page 34

Question VIII.2 (24)

Similarly, another person evaluated the sharpness (**Sharpness**) a number of times for each of different setups (**Treat**) of images on computer screens. The result of the usual analysis of variance of these data gave the following R output, however, some of the values are replaced by the letters G,...,J:

Analysis of Variance Table

Response: Sharpness

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treat	7	111.5	G	I	J
Residuals	32	88.4	H		

What is the value of J?

- 1 1.261
- 2 5.766
- 3* 0.0002
- 4 0.2188
- 5 0.3002

————— FACIT-BEGIN —————

Since J is the p-value of a one-way ANOVA, we first need to find F_{obs} . As stated in 8.6 this can be found as:

$$F_{obs} = \frac{MS(Trt)}{MSE} = \frac{SS(Trt)/DF(Trt)}{MSE/DFE} = \frac{111.5/7}{88.4/32} = 5.766$$

And we can then find the p-value as

$$P(F > 5.766) = 1 - P(F \leq 5.766)$$

were F follows an F -distribution with with degrees of freedom (7, 32):

```
(Fobs <- (111.5/7)/(88.4/32))  
## [1] 5.765999  
1-pf(Fobs, 7, 32)  
## [1] 0.00022284
```

————— FACIT-END —————

Continues on page 35

Exercise IX

Eight experts have each assessed the sharpness (**Sharpness**) for each of 12 different setups (**Treat**) of images on computer screens, ie. 96 observations of sharpness in total split on 12 setups. The scale is a continuous scale from 0 to 15, in practice, done by marking the value on a line:

	Setup 1	Setup 2	Setup 3	Setup 4	Setup 5	Setup 6	Setup 7	Setup 8	Setup 9	Setup 10	Setup 11	Setup 12
Person 1	9.30	4.70	6.60	8.80	5.90	7.20	7.60	5.50	8.10	8.20	6.40	7.40
Person 2	10.20	7.00	8.80	10.70	9.80	7.00	9.20	9.60	8.00	11.80	8.90	10.20
Person 3	11.50	9.50	8.00	12.90	10.00	8.20	11.50	6.40	8.60	11.20	7.70	11.00
Person 4	11.90	6.60	8.20	12.70	5.40	9.00	4.90	8.10	10.10	12.90	8.20	8.70
Person 5	10.70	4.20	5.40	11.40	8.30	7.10	6.80	3.80	9.60	8.60	3.80	10.80
Person 6	10.90	9.10	7.10	11.40	8.60	5.90	8.50	10.50	6.40	11.70	9.50	7.10
Person 7	8.50	5.00	6.30	10.80	6.80	4.60	4.70	8.80	6.70	10.00	5.50	7.50
Person 8	12.60	8.90	10.70	13.50	11.40	8.90	9.50	8.60	7.40	13.50	8.70	9.80

The result of a usual twoway analysis of variance of these data gave the following R-output:

```
## Analysis of Variance Table
##
## Response: Sharpness
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Person      7  122.42  17.4881   8.4596 1.212e-07 ***
## Setup     11  224.28  20.3894   9.8630 6.864e-11 ***
## Residuals 77  159.18   2.0673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question IX.1 (25)

What are the conclusions of the usual hypothesis tests for such an analysis? (Both conclusions and arguments must be correct)

- 1 There is a difference between the mean sharpness for persons, but not for setups
- 2 There is a difference between the mean sharpness for setups, but not for persons
- 3 There is a difference between the variances for persons, but not for setups
- 4* There is a difference between the mean sharpness for both setups and persons
- 5 There is a difference between the variances for setups, but not for persons

————— FACIT-BEGIN —————

We look at the p-values. Since both the p-values (1.212e-07 and 6.864e-11 respectively) are below $\alpha = 0.05$, we can reject the null-hypotheses of no difference and conclude that both setup and person make a difference.

————— FACIT-END —————

Continues on page 38

Question IX.2 (26)

Which probability distribution has been used to find the p -value provided for **Setup** in the output?

- 1 The z -distribution (= standard normal distribution)
- 2 The t -distribution with 159 degrees of freedom
- 3 The χ^2 -distribution with 159 degrees of freedom
- 4 The F -distribution with degrees of freedom 7 and 11
- 5* The F -distribution with degrees of freedom 11 and 77

————— FACIT-BEGIN —————

See Theorem 8.22. Since we are trying to find the degrees of freedom associated to the **setup** they are $df1 = l - 1 = 12 - 1 = 11$ and $df2 = (8 - 1)(12 - 1) = 77$.

————— FACIT-END —————

Continues on page 39

Exercise X

In a questionnaire survey under an Introduction to Statistics lecture the participants were asked about different topics. This assignment will cover the analysis of the answers to one of the questions. There were 32 respondents in total.

The question asked was: "Are you worried that we don't do enough to stop climate change?". To this the students answered the following:

Answer	Count
Yes	27
No	5

Question X.1 (27)

Using the "Plus 2" correction when calculating the usual 95% confidence interval for the proportion of students who are worried about the climate (answering yes), one gets:

$$1 \quad \square \quad 0.844 \pm 2.04 \cdot \frac{0.844}{36} = [0.796, 0.892]$$

$$2 \quad \square \quad 0.806 \pm 1.69 \cdot \frac{0.806}{36} = [0.768, 0.844]$$

$$3^* \quad \square \quad 0.806 \pm 1.96 \sqrt{\frac{0.806 \cdot 0.194}{36}} = [0.677, 0.935]$$

$$4 \quad \square \quad 0.844 \pm 1.69 \sqrt{\frac{0.844 \cdot 0.156}{32}} = [0.736, 0.952]$$

$$5 \quad \square \quad 0.844 \pm 1.96 \sqrt{\frac{0.844 \cdot 0.156}{36}} = [0.725, 0.963]$$

————— FACIT-BEGIN —————

The usual formula for confidence interval shown in Method 7.3 requires a large sample size (i.e. high n), however this is not a large sample, so we can use the "plus 2" correction, see Remark 7.7. This is carried out by adding 2 to the count (x) and adding 4 to the number of observations (n), and thereafter using the formulas. So

$$\hat{p}_{\text{plus2}} = \frac{x + 2}{n + 2} = \frac{29}{36}$$

and then $z_{1-\alpha/2}$ is found in R by `qnorm(0.975)`, and finally

$$\begin{aligned} \hat{p}_{\text{plus2}} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}_{\text{plus2}}(1 - \hat{p}_{\text{plus2}})}{n + 4}} &= 0.806 \pm 1.96 \sqrt{\frac{0.806 \cdot 0.194}{36}} \\ &= [0.677, 0.935]. \end{aligned}$$

————— FACIT-END —————

Continues on page 40

Exercise XI

In a questionnaire survey 114 respondents were asked about their traffic preferences. Two questions were asked, which had the following three identical answer options: “Car”, “Bike” and “Train or bus”.

The two questions were: “If you have 10 km to DTU from your home, what kind of transportation would you prefer during the summer (they take about equal time)?”, and: “if you have 10 km to DTU from your home, which kind of transportation would you prefer during the winter (they take about equal time)?”.

The following distribution of answers were observed:

		Winter		
		Car	Bike	Train or bus
Summer	Car	27	2	4
	Bike	20	22	11
	Train or bus	13	3	12

The following is run in R:

```
## The data table
tbl <- matrix(c(27, 20, 13, 2, 22, 3, 4, 11, 12), nrow = 3)
rownames(tbl) <- c("Car", "Bike", "Trainorbus")
colnames(tbl) <- c("Car", "Bike", "Trainorbus")
tbl

##           Car Bike Trainorbus
## Car       27   2           4
## Bike      20  22          11
## Trainorbus 13   3          12

## Row sums (distribution for summer)
margin.table(tbl, 1)

##           Car      Bike Trainorbus
##           33      53           28

## Column sums (distribution for winter)
margin.table(tbl, 2)

##           Car      Bike Trainorbus
##           60      27           27

## Chi2-test
chisq.test(tbl, correct = FALSE)

##
## Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 27.608, df = 4, p-value = 1.498e-05
```

Continues on page 41

Question XI.1 (28)

What is the expected count of preferences for: bike in the summer and car in the winter, under the null hypothesis: there is independence between traffic preferences in summer and winter in the surveyed population?

1 $e_{11} = 114 \cdot \frac{33}{114} \cdot \frac{60}{114} = 17.37$

2 $e_{23} = 114 \cdot \frac{53}{114} \cdot \frac{27}{114} = 12.55$

3 $e_{12} = 114 \cdot \frac{33}{114} \cdot \frac{27}{114} = 7.816$

4 $e_{33} = 114 \cdot \frac{28}{114} \cdot \frac{27}{114} = 6.877$

5* $e_{21} = 114 \cdot \frac{53}{114} \cdot \frac{60}{114} = 27.89$

————— FACIT-BEGIN —————

Under the null hypothesis there is independence between the preferences, which essentially means that the distribution of the counts in each row is the same as the distribution of counts in the row sums, and the same for the columns. So it means that the distribution in one variable is not changed as a function of the other variable. This is the null hypothesis in Theorem 7.24 and it is tested using the χ^2 -test.

We take the two total proportions (in the row and in the column), which is our best estimates if H_0 is true, and multiply with the total number of observations

$$e_{21} = 114 \cdot \frac{53}{114} \cdot \frac{60}{114} = 27.89.$$

————— FACIT-END —————

Question XI.2 (29)

What is the conclusion at significance level 1% of the test for independence of traffic preferences in summer and winter (both conclusion and argument must be correct)?

1 No significant dependence between traffic preferences is found since the p -value > 0.01

2* A significant dependence between traffic preferences is found since the p -value < 0.01

3 No significant dependence between traffic preferences is found since the p -value < 0.01

4 A significant dependence between traffic preferences is found since the p -value > 0.01

5 The question cannot be answered with the given information

————— FACIT-BEGIN —————

The p -value of the χ^2 -test for independence is calculated in the last line in the R code and we simply read off the p -value to $1.5 \cdot 10^{-5}$, hence much smaller than $\alpha = 0.01$. Thus a significant dependence between traffic preferences is found since p -value < 0.01

————— FACIT-END —————

Continues on page 43

Question XI.3 (30)

There are 60 out of 114 who prefer to drive car in the winter. Would the following null hypothesis

$$H_0 : p_{\text{car,winter}} = 50\%,$$

be rejected at the 5% significance level with the usual test (both conclusion and the p -value must be correct)?

- 1 Yes, since the p -value is $0.024 < 0.05$
- 2* No, since the p -value is $0.57 > 0.05$
- 3 No, since the p -value is $0.40 > 0.05$
- 4 Yes, since the p -value is $0.089 > 0.05$
- 5 No, since the p -value is $0.21 > 0.05$

————— FACIT-BEGIN —————

We use Method 7.11, since we are dealing with a single proportion. This is easiest calculated using R by:

```
prop.test(60, 114, p=0.5, correct=FALSE)

##
## 1-sample proportions test without continuity correction
##
## data: 60 out of 114, null probability 0.5
## X-squared = 0.31579, df = 1, p-value = 0.5741
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
## 0.4353037 0.6156122
## sample estimates:
##          p
## 0.5263158
```

————— FACIT-END —————

THE EXAM IS FINISHED. ENJOY THE SUMMER!