

Written examination: 28. May 2016

Course name and number: **Introduction to Statistics (02323, 02402 and 02593)**

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 12 exercises. To answer the questions you need to fill in the prepared 30-question multiple choice form (on three separate 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"). 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	IV.1	V.1	V.2	V.3	VI.1	VI.2
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer										

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

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

Remember to provide your **study number**. The questionnaire contains 27 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

In an airport the security check screens exactly 10000 passengers each day. Based on data from a long period it was found that 8 out of 10000 passengers bring sharp objects in their carry on luggage. Let X be a random variable denoting the number of passengers with sharp objects on a day (based on exactly 10000 checks). X is assumed to follow a Binomial distribution.

Question I.1 (1)

What is the expected number of passengers with sharp objects on a day, and what is the variance of X ?

- 1 $E[X] = 0.0008$ and $V[X] = 10000 \cdot 0.8 \cdot 0.2 = 1600$
- 2 $E[X] = 0.0008 \cdot 10000 = 8$ and $V[X] = 10000 \cdot 0.0008 \cdot 0.0002 = 0.0016$
- 3 $E[X] = 0.0008 \cdot 10000 = 8$ and $V[X] = 10000 \cdot 0.0008 \cdot 0.9992 = 7.994$
- 4 $E[X] = 0.0008 \cdot 10000 = 8$ and $V[X] = 10000 \cdot 0.0008 = 8$
- 5 $E[X] = 0.8 \cdot 10 = 8$ and $V[X] = 10 \cdot 0.8 \cdot 0.2 = 1.6$

Question I.2 (2)

What is the probability of finding more than 10 passengers with sharp objects on a given day?

- 1 `qbinom(0.9, 10000, 0.0008)`
- 2 `1-dbinom(9990, 10000, 0.0008)`
- 3 `dbinom(10, 10000, 0.0008)`
- 4 `1-pbinom(9990, 10000, 0.0008)`
- 5 `1-pbinom(10, 10000, 0.0008)`

Continues on page 3

Exercise II

A pharmaceutical company made a study in which 300 persons were randomly divided into 3 treatment groups of 100 patients each. One group was assigned to a placebo treatment, one group received the company's own product, and the last group got a competitor's product. For each patient the weight change over a period of time was measured and the final data set consists of 300 observations of weight changes. The focus is on comparing the average weight change in each group.

Question II.1 (3)

What kind of statistical analysis is most suitable for this?

- 1 Multiple linear regression analysis
- 2 Test for independence in a $r \times c$ frequency table (Contingency table)
- 3 Paired t-test
- 4 Two-way analysis of variance
- 5 Oneway analysis of variance

Continues on page 4

Exercise III

A random variable X follows a uniform distribution on the interval $[0; 1]$.

Question III.1 (4)

The expected value and the variance of $(X + 2) \cdot 4$ is

1 $\mu = \frac{5}{2}$ and $\sigma^2 = 4^2$

2 $\mu = 10$ and $\sigma^2 = 4^2$

3 $\mu = 8$ and $\sigma^2 = 4^2$

4 $\mu = 8$ and $\sigma^2 = \frac{1}{3}$

5 $\mu = 10$ and $\sigma^2 = \frac{4}{3}$

Continues on page 5

Exercise IV

A drone manufacturer is focusing on the feasible flight time between recharges. The flight time depends among other things on the weight of the drone. The drone basically consists of a battery (B), a skeleton (S) and four engines with propellers (M_1, \dots, M_4). It is assumed that the weights of the individual parts are independent and in the following all weights are in grams. The weights of the three types of components are given by the following Normal distributions: Battery: $B \sim N(100, 10^2)$, skeleton: $S \sim N(40, 5^2)$ and engines with propellers: $M_i \sim N(15, 2^2)$, $i = 1, \dots, 4$. (Each distribution is given on the usual form: $N(\mu, \sigma^2)$)

Question IV.1 (5)

The expected value and variance for the weight of the assembled drones are found to be

- 1 $\mu = 200$ and $\sigma^2 = 189$
- 2 $\mu = 200$ and $\sigma^2 = 141$
- 3 $\mu = 155$ and $\sigma^2 = 189$
- 4 $\mu = 155$ and $\sigma^2 = 129$
- 5 $\mu = 170$ and $\sigma^2 = 141$

Continues on page 6

Exercise V

There is a recommendation to eat 600 grams of fruit and vegetables each day. Regularly surveys of Danish dietary habits are made to see if the recommendation is met.

The results of the daily intake of fruits and vegetables (in grams) for the last four of this kind of dietary studies (conducted in the years 1995, 2000-2002, 2003-2004 and 2005-2008) can be summarized by the following output from R.

Survey	n	median	mean	var	std
1995	1564	259.82	290.887	28861.55	169.887
2000-2002	3043	386.057	433.817	62029.21	169.887
2003-2004	1310	404.936	453.279	74159.29	272.322
2005-2008	1983	429.132	479.285	77166.51	277.789

Survey	2.5%	5.0%	Q1	Q3	95.0%	97.5%
1995	66.102	87.062	171.209	374.303	606.609	686.361
2000-2002	98.613	129.574	257.224	555.168	928.673	1055.419
2003-2004	83.48	127.528	256.286	583.723	974.246	1180.891
2005-2008	105.348	141.81	279.359	617.371	991.09	1189.367

In all the questions in this exercise one can assume that the data from each of the four studies are normally distributed.

Question V.1 (6)

The question is no longer part of the curriculum.

Continues on page 7

Question V.2 (7)

A new dietary study is planned on the basis of the observed variation in dietary survey 2005-2008. What should the sample size be if the 90% confidence interval for the mean intake of fruits and vegetables is aimed to have a width of 20 grams?

1 $n \approx 77166.51 / \left(\frac{20}{1.96}\right)^2 = 741.1$

2 $n \approx \left(\frac{479.285 \cdot 1.6449}{10}\right)^2 = 6215.4$

3 $n \approx \frac{77166.51}{1.96 \cdot 20} = 1968.5$

4 $n \approx \left(\frac{1.6449 \cdot 277.789}{10}\right)^2 = 2087.9$

5 $n \approx \left(\frac{1.6449 \cdot \sqrt{1983}}{1.6456}\right)^2 = 1981.3$

Question V.3 (8)

Determine the 95% confidence interval for the mean intake of fruit and vegetables in the 2003-2004 survey.

1 $404.936 \pm 1.9618 \cdot \sqrt{\frac{74159.29}{1310}} = [390.176; 419.697]$

2 $[83.48; 1180.891]$

3 $453.279 \pm 1.9618 \cdot 7.524 = [438.518; 468.040]$

4 $453.279 \pm 1.6460 \cdot \frac{272.322}{\sqrt{1310}} = [440.894; 465.664]$

5 $453.279 \pm 1.96 \cdot 272.322 = [-80.472; 987.030]$

Continues on page 8

Exercise VI

A major company took a random sample of 20 employees and determined their daily intake of fruits and vegetables, and registered the following observations of the daily intake (in grams):

740.59	262.28
667.96	730.55
809.33	324.19
1138.12	421.93
489.42	561.23
352.78	552.96
1309.66	130.96
259.86	440.82
896.01	955.03
481.00	257.80

In all the questions in this exercise one can assume that the data is normally distributed.

Summary from R gives the following results for the intake of fruits and vegetables:

n	median	mean	variance	Std. dev.		
20	521.1898	589.1245	98996.08	314.6364		
	2.5%	5.0%	Q1	Q3	95.0%	97.5%
	191.2095	251.4552	345.635	757.777	1146.697	1228.178

Question VI.1 (9)

Determine the 90% confidence interval for the variance σ^2 for the daily intake of fruits and vegetables of employees in the company

1 $\left[\frac{20 \cdot 314.636}{32.852}; \frac{20 \cdot 314.636}{8.907} \right] = [191.548; 706.492]$

2 $98996.08 \pm 30.144 \cdot \frac{314.636}{\sqrt{20}} = [96875.31; 101116.90] = [311.248^2; 317.989^2]$

3 $98996.08 \pm 1.7959 \cdot \frac{314.636^2}{\sqrt{20}} = [59241.79; 138750.40] = [243.396^2; 372.492^2]$

4 $[314.636^2 - 10.117 \cdot 314.636; 314.636^2 + 30.144 \cdot 314.636] = [309.537^2; 329.364^2]$

5 $\left[\frac{19 \cdot 314.636^2}{30.144}; \frac{19 \cdot 314.636^2}{10.117} \right] = [249.796^2; 431.181^2]$

Continues on page 9

Question VI.2 (10)

Actually, the above data consist of 2 random samples, where the left column indicate the intakes of 10 men and the right column intakes for 10 women. One wants to investigate whether there are differences in men's and women's mean intake of fruits and vegetables.

The following R code is executed (not all necessarily sensible):

```
m <- c(740.59, 667.96, 809.33, 1138.12, 489.42, 352.78,
       1309.66, 259.86, 896.01, 481.00)
f <- c(262.28, 730.55, 324.19, 421.93, 561.23, 552.96,
       130.96, 440.82, 955.03, 257.80)
t.test(m, f, paired = TRUE)

##
## Paired t-test
##
## data: m and f
## t = 1.7378, df = 9, p-value = 0.1163
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -75.65101 577.04701
## sample estimates:
## mean of the differences
##                250.698

mean(f) - mean(m)

## [1] -250.698

t.test(m, f)

##
## Welch Two Sample t-test
##
## data: m and f
## t = 1.9001, df = 16.481, p-value = 0.07506
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -28.33599 529.73199
## sample estimates:
## mean of x mean of y
##    714.473    463.775
```

Continues on page 10

```
t.test(m, mu = median(f))

##
## One Sample t-test
##
## data: m
## t = 2.6577, df = 9, p-value = 0.02614
## alternative hypothesis: true mean is not equal to 431.375
## 95 percent confidence interval:
## 473.5091 955.4369
## sample estimates:
## mean of x
## 714.473

t.test(f)

##
## One Sample t-test
##
## data: f
## t = 5.957, df = 9, p-value = 0.0002135
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 287.6581 639.8919
## sample estimates:
## mean of x
## 463.775
```

Continues on page 11

What is the conclusion of the test of the hypothesis (at the level $\alpha = 0.05$)

$$H_0 : \mu_f = \mu_m$$

$$H_1 : \mu_f \neq \mu_m$$

corresponding to the examination of whether there are differences in men's and women's mean intake of fruits and vegetables.

- 1 Yes, there is a significant difference between men's and women's intake of fruits and vegetables per day, given that the relevant p -value is 0.1163
- 2 It is apparent that there is a significant difference between men's and women's intake of fruits and vegetables per day, with $\hat{\mu}_f - \hat{\mu}_m = 463.775 - 714.473 = -250.698$. It appears that men eat more fruits and vegetables per day than women
- 3 There is no significant difference between men's and women's intake of fruits and vegetables per day, as the relevant p -value is 0.07506
- 4 Yes, there is a significant difference between men's and women's intake of fruits and vegetables per day given that the relevant p -value is 0.02614
- 5 No, there is no significant difference in the intake of fruit and vegetables per day for men and women since the relevant p -value is 0.0002135

Question VI.3 (11)

What is the upper quartile (75% percentile) for the 10 intake for women, based on the textbook definition of this?

- 1 561.23
- 2 262.28
- 3 431.375
- 4 709.97
- 5 246.19

Continues on page 12

Exercise VII

A study investigated dioxin emissions from a Danish incineration plant. Parts of the measured variables are shown in the table below. The 3 variables are: Dioxin measured in “parts per million”, load of the plant measured as relative deviation from a reference, and the content of water in the emitted gas (measured in %). As seen in the table, there are in total 23 measurements. Average and empirical standard deviation (“sample standard deviation”) are listed at the bottom of the table.

	Dioxin (<i>ppm</i>)	Load	H ₂ O (%)
	DIOX	NEFF	H2O
1	984.10	0.2560	13.78
2	662.00	0.3520	14.59
3	270.90	-0.0200	12.55
⋮	⋮	⋮	⋮
21	112.70	0.0490	13.84
22	94.20	0.1350	14.18
23	323.20	0.2820	12.56
\bar{x}	329.16	-0.0266	12.589
<i>s</i>	254.95	0.2105	1.980

The primary interest in the study is related to the question: can dioxin emissions be influenced by adjusting the load. For this purpose the following R code is executed (the data input is, however, omitted)

```
fit1 <- lm(DIOX ~ NEFF)
summary(fit1)

##
## Call:
## lm(formula = DIOX ~ NEFF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -348.41 -116.61  -22.98  101.19  496.16
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    347.8      44.7    7.781 0.000000128 ***
## NEFF           702.2     215.3    3.262  0.00373 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 212.6 on 21 degrees of freedom
## Multiple R-squared:  0.3362, Adjusted R-squared:  0.3046
## F-statistic: 10.64 on 1 and 21 DF, p-value: 0.00373
```

Continues on page 13

Hence, the following model is examined

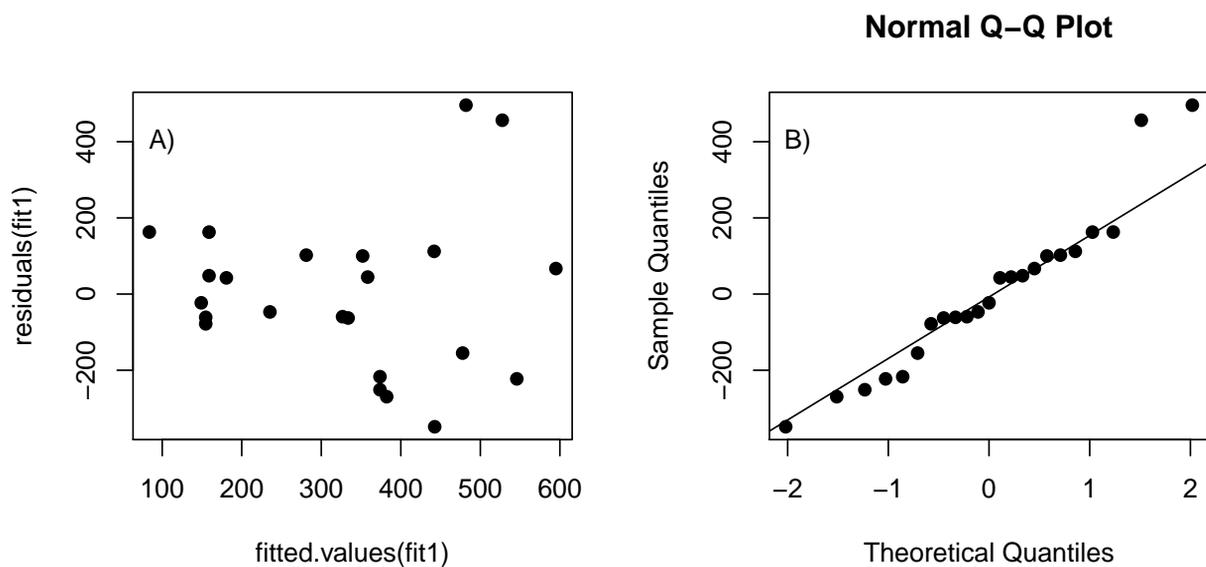
$$\text{DIOX}_i = \beta_0 + \beta_1 \text{NEFF}_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

Question VII.1 (12)

At significance level $\alpha = 0.05$, what is the conclusion about the effect of the load on dioxin emissions (both conclusions and argument must be correct)?

- 1 There is an effect since $1.3 \cdot 10^{-7} < 0.05$, and $\beta_1 > 0$ because $347.8 > 0$
- 2 There is an effect since $702.2 > 347.2$, and $\beta_1 > 0$ because $3.26 > 0$
- 3 There is an effect since $0.0037 < 0.05$, and $\beta_1 > 0$ because $702.2 > 0$
- 4 There is no evidence of an effect as $3.26 < 7.78$.
- 5 There is no evidence of an effect as $0.0037 > \frac{0.05}{100}$.

In order to investigate whether the conditions for using the model are satisfied, 2 residual plots are shown in the figure below.



Continues on page 14

Question VII.2 (13)

Which assumptions are primarily examined in each of the 2 plots (both assumptions and figure reference must be correct)?

- 1 Variance homogeneity (A) and the normal distribution assumption (B)
- 2 $E(\epsilon) = 0$ (A) and $V(\epsilon) = \sigma^2$ (B)
- 3 Variance-homogeneity (A) and assumption of linearity (B)
- 4 $E(\epsilon) = 0$ (A) and independence (B)
- 5 Independence (A) and variance homogeneity (B)

Regardless of the outcome of the previous question it is decided to make the analysis on log-transformed dioxin data. The result of the analysis conducted in R is shown below (some of the numbers are, however, replaced by letters)

```
> fit2 <- lm(log(DIOX) ~ NEFF)
> summary(fit2)
```

Call:

```
lm(formula = log(DIOX) ~ NEFF)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.29588	-0.44048	0.05093	0.49403	0.94119

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.5927	A	B	< 2e-16 ***
NEFF	1.8416	C	D	E

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6266 on 21 degrees of freedom

Multiple R-squared: 0.2862, Adjusted R-squared: 0.2522

F-statistic: 8.42 on 1 and 21 DF, p-value: 0.00853

Continues on page 15

Question VII.3 (14)

What is D?

1 $D = \frac{0.623^2}{21} = 0.019$

2 $D = 0.623 \cdot \sqrt{\frac{1}{22 \cdot 0.211^2}} = 0.63$

3 $D = \frac{1.84}{c}$

4 $D = \frac{c}{B}$

5 $D = \frac{0.623}{\sqrt{22}} = 0.13$

Question VII.4 (15)

What is the usual 95% confidence interval for the slope in the model for $\log(\text{DIOX})$?

1 $1.84 \pm 1.72 \cdot B$

2 $1.84 \pm 2.08 \cdot 0.2862$

3 $1.84 \pm 1.72 \cdot D$

4 $1.84 \pm 2.08 \cdot 0.623$

5 $1.84 \pm 2.08 \cdot C$

Continues on page 16

It is now decided to investigate whether water vapor should be included in the model. For this purpose, a multiple regression model is formulated

$$\log(\text{DIOX}_i) = \beta_0 + \beta_1 \text{NEFF}_i + \beta_2 \text{H2O}_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

In order to investigate the model, the following R-code has been executed (including the result)

```
fit3 <- lm(log(DIOX) ~ NEFF + H2O)
summary(fit3)

##
## Call:
## lm(formula = log(DIOX) ~ NEFF + H2O)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11709 -0.36741  0.05337  0.36192  0.90410
##
## Coefficients:
##              Estimate Std. Error t value    Pr(>|t|)
## (Intercept)   7.4704     0.8098   9.225 0.0000000121 ***
## NEFF          2.1963     0.5955   3.688   0.00146 **
## H2O          -0.1484     0.0633  -2.345   0.02948 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5687 on 20 degrees of freedom
## Multiple R-squared:  0.4401, Adjusted R-squared:  0.3841
## F-statistic:  7.86 on 2 and 20 DF,  p-value: 0.003028
```

Question VII.5 (16)

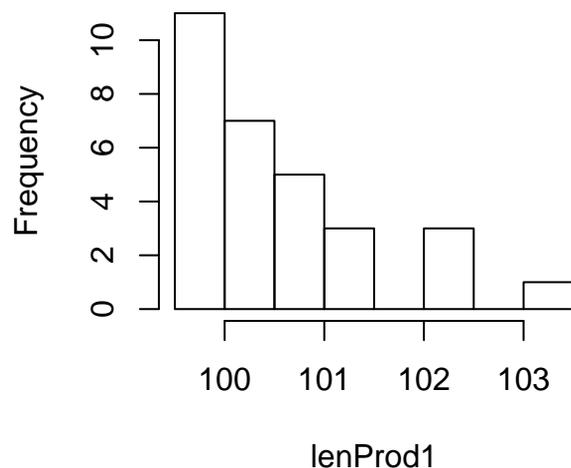
What are the parameter estimates for the model?

- 1 $\hat{\beta}_0 = 7.47, \hat{\beta}_1 = 2.20, \hat{\beta}_2 = -0.148$ og $\hat{\sigma} = 0.569$
- 2 $\hat{\beta}_0 = 9.22, \hat{\beta}_1 = 3.69, \hat{\beta}_2 = -2.35$ og $\hat{\sigma} = 0.4401$
- 3 $\hat{\beta}_0 = 7.47, \hat{\beta}_1 = 2.20, \hat{\beta}_2 = -0.148$ og $\hat{\sigma} = 0.384$
- 4 $\hat{\beta}_0 = 9.22, \hat{\beta}_1 = 3.69, \hat{\beta}_2 = -2.35$ og $\hat{\sigma} = 0.569$
- 5 $\hat{\beta}_0 = 9.22, \hat{\beta}_1 = 3.69, \hat{\beta}_2 = -2.35$ og $\hat{\sigma} = 7.86$

Continues on page 17

Exercise VIII

In a production, it is anticipated that part of the production must be discarded due to a minimum length requirement. It is found that it is economically feasible if not more than 25% of the produced elements are discarded. An experiment is carried out with a particular production method and the length of 50 produced items are observed. The observations are loaded and stored in the vector `lenProd1`. A histogram of the observations is



A confidence interval for the lower quartile (i.e. the 25% quantile) must be calculated without any assumptions of the distribution. The following R code is run:

```
## Simulate 10000 samples
k = 10000
simSamples = replicate(k, sample(lenProd1, replace = TRUE))

simStat = apply(simSamples, 2, quantile, probs=0.25)
quantile(simStat, c(0.005,0.025,0.05,0.95,0.975,0.995))

##      0.5%      2.5%      5%      95%      97.5%      99.5%
## 99.5825 99.7225 99.7600 100.0575 100.1025 100.2300

simStat = apply(simSamples, 2, quantile, probs=0.5)
quantile(simStat, c(0.005,0.025,0.05,0.95,0.975,0.995))

##      0.5%      2.5%      5%      95%      97.5%      99.5%
## 99.9450 99.9850 100.0000 100.5900 100.6550 100.8801

simStat = apply(simSamples, 2, quantile, probs=0.75)
quantile(simStat, c(0.005,0.025,0.05,0.95,0.975,0.995))

##      0.5%      2.5%      5%      95%      97.5%      99.5%
## 100.3000 100.4625 100.5125 101.4300 101.9550 102.1450
```

Continues on page 18

Note that the option `probs` is "passed on" to the `quantile` function, such that for each of the three calls to `apply` a different quantile is calculated by the `quantile` function.

Question VIII.1 (17)

What is the 95% confidence interval for the lower quartile (i.e. the 25% quantile) for the length?

- 1 [99.72, 100.10]
- 2 [100.00, 100.59]
- 3 [99.59, 100.23]
- 4 [100.46, 101.96]
- 5 [100.49, 101.43]

Question VIII.2 (18)

In the following Q denotes a quartile, such that Q_1 is the lower quartile, Q_2 is the median and Q_3 is the upper quartile. In which of the following two-sided tests would the null hypothesis have been rejected on significance level $\alpha = 0.01$ under the assumptions and simulation results presented above?

- 1 $H_0 : Q_1 = 100$ vs. $H_1 : Q_1 \neq 100$
- 2 $H_0 : Q_2 = 100$ vs. $H_1 : Q_2 \neq 100$
- 3 $H_0 : Q_2 = 101$ vs. $H_1 : Q_2 \neq 101$
- 4 $H_0 : Q_3 = 101$ vs. $H_1 : Q_3 \neq 101$
- 5 $H_0 : Q_3 = 102$ vs. $H_1 : Q_3 \neq 102$

Continues on page 19

Exercise IX

A new wind turbine is to be build on a site and some investigations of the wind conditions on the site have been carried out. The outcome is that the average hourly wind speed on the site can be represented with the probability density function plotted below in Figure 1:

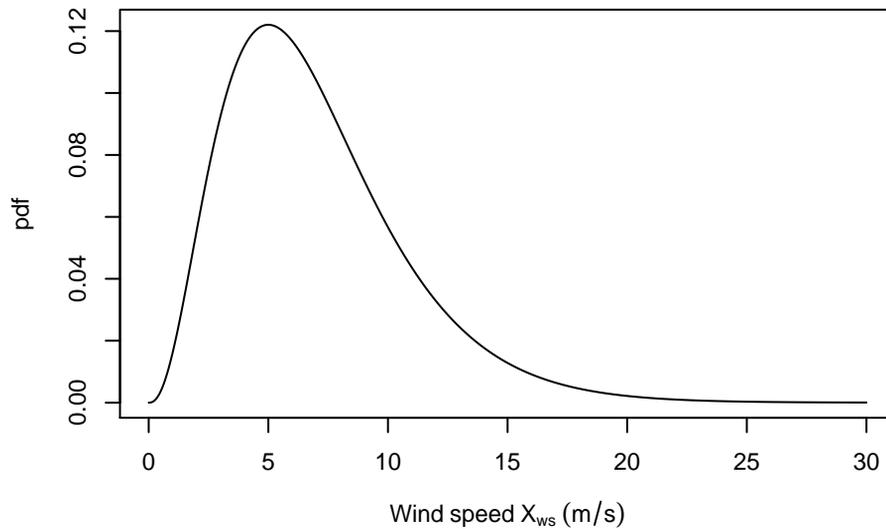


Figure 1: *Probability density function (pdf) for the wind speed X_{ws} .*

In order to investigate the power production of a wind turbine build on the site a function called the 'power curve' for the wind turbine is used (it is the power output as a function of the wind speed, it has nothing to do with the power of a statistical test). The power curve used is plotted below in Figure 2:

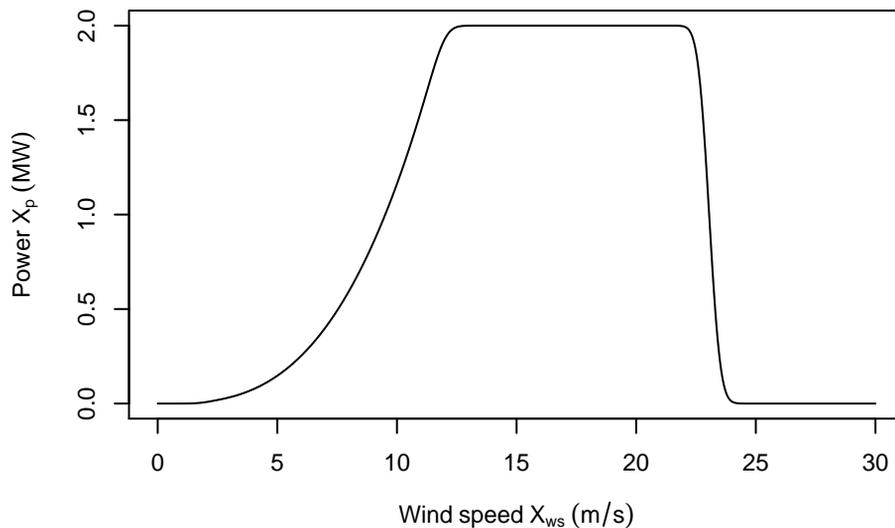


Figure 2: *Power curve, i.e. the function between the wind speed X_{ws} and the output power X_p .*

It can be seen that if the wind speed is 5 m/s the output power will be around 0.15 MW and at 15 m/s the output power will be 2 MW. This function can be applied directly on average hourly values of the wind speed and gives then hourly average values of output power.

Continues on page 20

Let X_{ws} be the average hourly wind speed in m/s and the power output in MW

$$X_p = f_{\text{powercurve}}(X_{ws})$$

where $f_{\text{powercurve}}()$ is the power curve function.

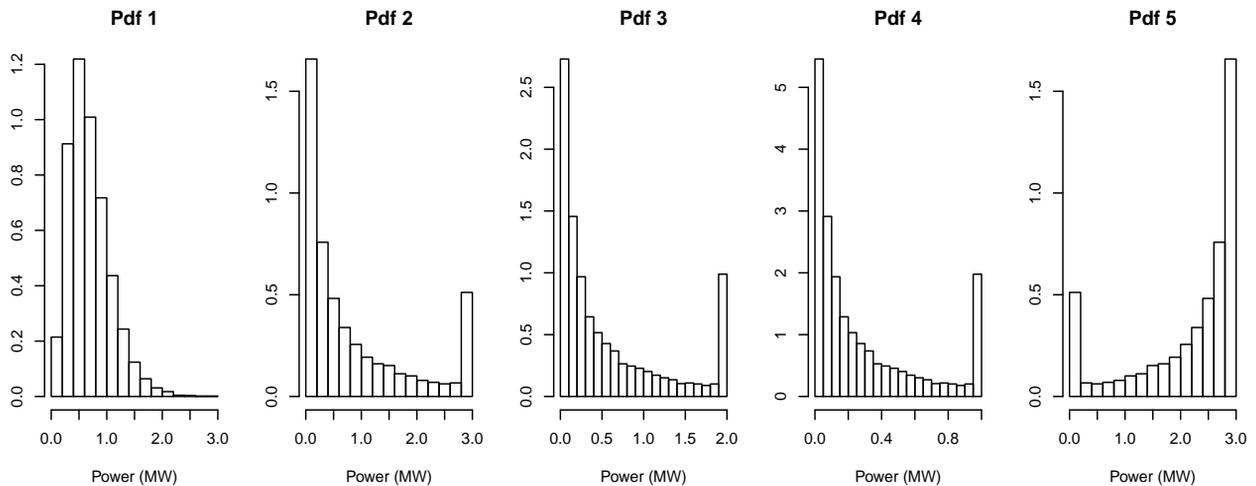
Question IX.1 (19)

From the plot of the pdf in Figure 1 conclude which of the following statements is not correct (Note: you must mark the FALSE statement - four of the statements are correct!):

- 1 $P(X_{ws} > 12) \approx 0.10$
- 2 $P(X_{ws} < 5) \approx 0.34$
- 3 $P(X_{ws} > 10) \approx 0.19$
- 4 $P(X_{ws} > 0) \approx 1$
- 5 $P(X_{ws} < 15) \approx 0.04$

Question IX.2 (20)

The probability density function of the hourly output power X_p is found by simulation. Which of the following pdfs can be the probability density function of the hourly output power at the site, i.e. the pdf of X_p ?



Continues on page 21

1 Pdf 1

2 Pdf 2

3 Pdf 3

4 Pdf 4

5 Pdf 5

Question IX.3 (21)

In wind power forecasting it is very important to include the forecast uncertainty. It can be described by the variance of the power output forecast σ_p^2 . In the range from 5 to 10 m/s the power curve is constructed by the following relation

$$f_{\text{powercurve}}(X_{\text{ws}}) = aX_{\text{ws}}^3 \quad \text{for } 5 < X_{\text{ws}} < 10 \quad (1)$$

Further, it is known that the variance of the wind speed forecast in the same range is σ_{ws}^2 .

Which of the following expressions calculates an approximation to the variance of the output power forecast σ_p^2 for a wind speed X_{ws} in the range from 5 to 10 m/s?

1 $\sigma_p^2 = X_{\text{ws}}^3 \sigma_{\text{ws}}^2$

2 $\sigma_p^2 = 9a^2 X_{\text{ws}}^4 \sigma_{\text{ws}}^2$

3 $\sigma_p^2 = \int_5^{10} \sigma_{\text{ws}}^2 3ax^2 dx$

4 $\sigma_p^2 = \int_5^{10} \sigma_{\text{ws}}^2 x^3 dx$

5 $\sigma_p^2 = a^2 \sigma_{\text{ws}}^2$

Continues on page 22

Exercise X

A supermarket chain would like to track the trend in sales of organic meat. Therefore, they have for four years conducted a survey among their customers, asking whether the customers bought organic meat. The distribution of the answers is seen in the table below.

	2011	2012	2013	2014
Bought organic meat	68	72	81	90
Bought non-organic meat	432	428	419	410

Question X.1 (22)

The supermarket chain wants to test the hypothesis that the proportion buying organic meat is the same each year.

$$H_0 : p_1 = p_2 = p_3 = p_4$$

Here p_1 is the proportion that buys organic meat in 2011, p_2 is the proportion that buys organic meat in 2012 etc.

What is the expected number of organic meat purchases in 2014, under the hypothesis of equal proportions each year?

- 1 144.69
- 2 250.00
- 3 77.75
- 4 43.48
- 5 422.25

Continues on page 23

Question X.2 (23)

A χ^2 distributed test statistic is used in order to test the hypothesis

$$H_0 : p_1 = p_2 = p_3 = p_4$$

What is the contribution $q_{No,2011}$ to the test statistic χ_{obs}^2 from the respondents, who answer that they bought non-organic meat in 2011?

- 1 $q_{No,2011} = 0.2251$
- 2 $q_{No,2011} = 1.2227$
- 3 $q_{No,2011} = 9.75$
- 4 $q_{No,2011} = 0.0231$
- 5 $q_{No,2011} = 0.2201$

Question X.3 (24)

The supermarket chain has now conducted the test of the hypothesis

$$H_0 : p_1 = p_2 = p_3 = p_4$$

to track the trend in sales of organic meat.

In this case the relevant test statistic becomes 4.3977.

Which of the following R commands calculates the p-value for the hypothesis test?

- 1 `pchisq(4.3977, 3)`
- 2 `2*(1-pnorm(4.3977))`
- 3 `1-pchisq(4.3977, 3)`
- 4 `2*(1-pchisq(4.3977, 4))`
- 5 `1-pchisq(4.3977, 6)`

Continues on page 24

Exercise XI

Studies have shown that teenage girls have a lower life satisfaction than boys. Therefore, a team of first-year students decided to study life satisfaction among their peers. The results of their study were as follows.

	High life satisfaction	Lower life satisfaction
Men	68	208
Women	18	74

Question XI.1 (25)

What is the correct 95% confidence interval for the estimate of the difference between the proportion of high life satisfaction for men and women?

1 $(0.2464 - 0.1957) \pm 1.64 \cdot \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}} = (-0.029; 0.131)$

2 $(0.2464 - 0.1957) \pm 1.96 \cdot \sqrt{\left(\frac{0.2464(1-0.2464)}{276}\right)^2 + \left(\frac{0.1957(1-0.1957)}{92}\right)^2} = (0.047; 0.054)$

3 $\frac{0.2464}{0.1957} \pm 1.96 \cdot \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}} = (1.16; 1.35)$

4 $(0.2464 - 0.1957) \pm 3.84 \cdot \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}} = (-0.137; 0.238)$

5 $(0.2464 - 0.1957) \pm 1.96 \cdot \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}} = (-0.045; 0.146)$

Continues on page 25

Question XI.2 (26)

Now we want to test the hypothesis that the proportion of high life satisfaction is the same for men and women. I.e. we testing the hypothesis (at significance level $\alpha = 0.05$).

$$H_0 : p_1 = p_2$$

$$H_A : p_1 \neq p_2$$

Here p_1 is the proportion of high life satisfaction amongst men and p_2 is the proportion of high life satisfaction amongst women.

What is the conclusion to this test? (Both the conclusion and the argumentation must be correct).

- 1 H_0 is rejected, since the test statistic $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{0.2337(1-0.2337)(\frac{1}{276}+\frac{1}{92})}} = 2.298$ leads to a p-value of 0.02
- 2 H_0 is accepted, since the test statistic $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{0.2337(1-0.2337)(\frac{1}{276}+\frac{1}{92})}} = 0.995$ leads to a p-value of 0.32
- 3 H_0 is accepted, since the test statistic $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}}} = 1.038$ leads to a p-value of 0.15
- 4 H_0 is accepted, since the test statistic $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{0.2337(1-0.2337)(\frac{1}{276}+\frac{1}{92})}} = 2.298$ leads to a p-value of 0.02
- 5 H_0 is rejected, since the test statistic $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}}} = 21.3$ leads to a p-value < 0.0001

Continues on page 26

Exercise XII

18 test persons evaluated the bass quality of 3 different headphones, so that all 18 persons evaluated all 3 headphones such that the data consist of 54 observations of bass quality on a scale between 0 and 150. The average of the three headphone bass qualities were:

Headphone	Average
1	53.5
2	55.5
3	97.1

Question XII.1 (27)

How is the $SS(Tr)$ calculated in the 2-way analysis of variance, which compares the mean bass quality for the three headphones? (Where "Tr" now refers to the 3 headphones)

- 1 $18 \cdot (53.5 - 68.7)^2 + 18 \cdot (55.5 - 68.7)^2 + 18 \cdot (97.1 - 68.7)^2$
- 2 $(53.5 - 68.7)^2 + (55.5 - 68.7)^2 + (97.1 - 68.7)^2$
- 3 $\frac{(53.5-68.7)^2}{53.5} + \frac{(55.5-68.7)^2}{55.5} + \frac{(97.1-68.7)^2}{55.5}$
- 4 $\frac{(53.5-68.7)}{53.5} + \frac{(55.5-68.7)}{55.5} + \frac{(97.1-68.7)}{55.5}$
- 5 $3 \cdot (53.5 - 68.7)^2 + 3 \cdot (55.5 - 68.7)^2 + 3 \cdot (97.1 - 68.7)^2$

Question XII.2 (28)

If, in line with the above, we let "persons" constitute "blocks", we are given that $SS(BI) = 6003.5$ and that $SSE = 7160.3$ in the 2-way analysis of variance. What will the F-test statistic for the hypothesis that the 18 persons have the same mean value be?

- 1 $F_{obs} = \frac{18 \cdot 6003.5}{210.6/3}$
- 2 $F_{obs} = \frac{3 \cdot 6003.5}{7160.3/17}$
- 3 $F_{obs} = \frac{6003.5/17}{7160.3/34}$
- 4 $F_{obs} = \frac{(6003.5-210.6)^2}{7160.3}$
- 5 $F_{obs} = \frac{(6003.5/18-210.6)}{\sqrt{(210.6)}}$

Continues on page 27

Question XII.3 (29)

The hypothesis of no difference in mean bass quality of the three headphones is by the usual test evaluated by which sampling distribution?

- 1 z -distribution (= standard normal distribution)
- 2 t -distribution with 53 degrees of freedom
- 3 χ^2 -distribution with 53 degrees of freedom
- 4 F -distribution with 2 and 34 degrees of freedom
- 5 F -distribution with 3 and 51 degrees of freedom

Question XII.4 (30)

What will the 95% confidence interval be for the mean difference between headphone 2 and 1? (It can be assumed that this is a "pre-planned" comparison)

- 1 $2 \pm 2 \cdot 210.6$
- 2 $2 \pm 2.03 \cdot \sqrt{210.6}$
- 3 $2 \pm 1.96 \cdot \frac{210.6}{54}$
- 4 2 ± 1.96
- 5 $2 \pm 2.03 \cdot \sqrt{2 \cdot 210.6 \frac{1}{18}}$

THE EXAM IS FINISHED. ENJOY THE SUMMER!