Written examination: 16 December 2018

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

Duration: 4 hours

Aids and facilities allowed: All

The questions were answered by

____________________ ____________________ __________

(student number) (signature) (table number)

This exam consists of 30 questions of the "multiple choice" type, which are divided between 14 exercises. To answer the questions, you need to fill in the "multiple choice" form (6 separate pages) on CampusNet with the numbers of the answers that you believe to be correct.

5 points are given for a correct "multiple choice" answer, and −1 point is given for a wrong answer. ONLY the following 5 answer options are valid: 1, 2, 3, 4, or 5. If a question is left blank or an invalid answer is entered, 0 points are given for the question. Furthermore, if more than one answer option is selected for a single question, which is in fact technically possible in the online system, 0 points are given for the question. The number of points needed to obtain a specific mark or to pass the exam is ultimately determined during censoring.

The final answers should be given by filling in and submitting the form online via CampusNet. The table provided here is ONLY an emergency alternative. Remember to provide your student number if you do hand in on paper.

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

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

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

The exam paper contains 21 pages. Continue on page 2

Multiple choice questions: *Note that in each question, one and only one of the answer options is correct. Furthermore, not all the suggested answers are necessarily meaningful. Always remember to round your own result to the number of decimals given in the answer options before you choose your answer.*

Exercise I

In the analysis of a single sample, 10 measurements are assumed to be independent and sampled from a normal distribution with mean μ and variance σ^2. The sample mean is $\bar{x} = 0.57$, while the sample standard deviation is $s = 0.32$.

Question I.1 (1)

Which of the following is a standard 99% confidence interval for the theoretical standard deviation σ?

1 ☐ $[0.20, 0.73]$

2 ☐ $0.57 \pm 1.96 \cdot 0.32$

3 ☐ $[0.22, 0.58]$

4 ☐ $[0.05, 0.34]$

5 ☐ $[0.03, 0.53]$

Exercise II

We would like to determine the median of X_1/X_2, when X_1 and X_2 are independent stochastic variables, which are both normal distributed with mean 1 and variance 1. The distribution of the ratio is not trivial; therefore we resort to simulation to determine an estimate and a confidence interval for the median of the distribution of X_1/X_2.

Question II.1 (2)

First, 10000 medians are simulated, each being the median of 10000 ratios. We store these in R in the vector `medians`:

```
ratio <- replicate(10000, rnorm(10000, mean = 1)/rnorm(10000, mean = 1))
medians <- apply(ratio, 2, median)
```

Subsequently, the sample mean and a series of percentiles are calculated for these 10000 medians:

```r
mean(medians)

## [1] 0.6193

quantile(medians, c(0.005, 0.025, 0.05, 0.5, 0.95, 0.975, 0.995), type = 2)

##   0.5%   2.5%     5%    50%    95%  97.5%  99.5%
## 0.5873 0.5949 0.5989 0.6193 0.6402 0.6443 0.6515
```

Which of the following choices yields an estimate for the median of X_1/X_2 and a 95% confidence interval for this median?

1 ☐ Estimate: 1.
95% confidence interval: $[1 - 1.96 \cdot 0.6193, 1 + 1.96 \cdot 0.6193]$.

2 ☐ Estimate: 0.6193.
95% confidence interval: $[0.5949, 0.6443]$.

3 ☐ Estimate: 1.
95% confidence interval: $[1 - 0.5949, 1 + 0.6443]$.

4 ☐ Estimate: 0.6193.
95% confidence interval: $[0.5873, 0.6515]$.

5 ☐ Estimate: 0.6193.
95% confidence interval: $[0.6193 - 0.5949, 0.6193 + 0.5949]$.

Exercise III

A normal distributed population has mean $\mu = 100$ and standard deviation $\sigma = 15$.

Question III.1 (3)

In a random draw, what is the probability of obtaining an observation below 90?

1 ☐ 0.252

2 ☐ 0.482

3 ☐ 0.518

4 ☐ 0.631

5 ☐ 0.748

Question III.2 (4)

If a random sample of $n = 10$ independent observations is drawn from the population, what is the probability that the sample mean is below 90?

1 ☐ 0.000783

2 ☐ 0.0175

3 ☐ 0.146

4 ☐ 0.252

5 ☐ 0.482

Question III.3 (5)

Suppose that a random sample of n independent observations is repeatedly drawn from the population, and that the sample variance S^2 is calculated in each repetition. What holds true for S^2?

1 ☐ n^2S^2 is F-distributed with $n-1$ and $n-2$ degrees of freedom.

2 ☐ S^2 is χ^2-distributed with $n-1$ degrees of freedom.

3 ☐ $(n-1)S^2/\sigma^2$ is χ^2-distributed with $n-1$ degrees of freedom.

4 ☐ S^2 is normal distributed with mean μ and variance σ^2/n^2.

5 ☐ S^2 has the same distribution as $(Z-\sigma^2)/n$, where Z is standard normal distributed.

Exercise IV

10 people have had their daily energy intake measured (in kJ). The measurements in the sample are shown in the table below:

Energy intake (kJ):	8230	5470	7515	5260	6390	6180	6515	6805	7515	5640

Question IV.1 (6)

What is the median of the sample?

1 ☐ 6390

2 ☐ 6515

3 ☐ (8230+5260)/2

4 ☐ (6390+6180)/2

5 ☐ (6390+6515)/2

Question IV.2 (7)

The sample mean is $\bar{x} = 6552$, while the sample standard deviation is $s = 975.94$. It is assumed that the daily energy intake may be modelled by a normal distribution, and that the observations are independent and identically distributed. What is the p-value for the t-test that tests the hypothesis that the mean daily energy intake is 7725 kJ?

1 ☐ 0.4

2 ☐ 0.06

3 ☐ 0.04

4 ☐ 0.006

5 ☐ 0.004

Exercise V

A married couple visits the same restaurant several times a month. Typically, they order a glass of red wine with their food. One day, they decide to complain to the owner. They believe that one of the waiters pours less wine into the glass than what they pay for. Consequently, the owner launches an experiment with three of the restaurant's waiters in order to investigate how much they pour into wine glasses, when they pour using a rule of thumb. Each of the three waiters (here anonymized by A, B, and C) were asked to pour red wine into 20 wine glasses, after which the content in each glass was measured. The data were read into R in two variables: `waiter`, indicating which waiter poured the wine, and `wine`, indicating the amount of wine in the glass (in mL).

The following code was run in R to analyze the data:

```
anova(lm(wine ~ waiter))

## Analysis of Variance Table
##
## Response: wine
##            Df Sum Sq Mean Sq F value   Pr(>F)
## waiter      2 1043.4  521.71  6.9594 0.001976 **
```

```
## Residuals 57 4273.0   74.97
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question V.1 (8)

What may be concluded from the R output above, when a significance level of 5% is used (both the reasoning and conclusion must be correct)?

1 ☐ As the observed F-test statistic is larger than the 0.95 quantile of the $F(57, 2)$-distribution, there is a significant difference in the expected amount of wine in glasses poured by the three different waiters.

2 ☐ As the p-value is larger than 5%, there is no significant difference in the expected amount of wine in glasses poured by the three different waiters.

3 ☐ As the sum of squared errors, *SSE*, is more than four times the size of the treatment sum of squares, *SS(Tr)*, there is too much noise in the data for it to be meaningful to perform a one-way analysis of variance.

4 ☐ As the observed F-test statistic is larger than the 0.95 quantile of the $F(2, 57)$-distribution, there is a significant difference in the expected amount of wine in glasses poured by the three different waiters.

5 ☐ As the p-value is less than 5%, there is no significant difference in the expected amount of wine in glasses poured by the three different waiters.

Question V.2 (9)

Among other things, the owner would like to make a comparison between waiter A (the young waiter, whom the couple complained about) and waiter B (an older waiter with many years of experience in the business). On average, waiter A poured 127 mL of wine into each glass, while waiter B poured 135 mL. Compute the t-test statistic for the post hoc pairwise hypothesis test which compares the expected amount of wine in glasses poured by waiter A and waiter B.

1 ☐ $t_{obs} = -0.92$

2 ☐ $t_{obs} = -4.13$

3 ☐ $t_{obs} = -2.92$

4 ☐ $t_{obs} = -1.07$

5 ☐ $t_{obs} = -0.11$

Question V.3 (10)

In addition to the information in the previous question, it is given that, on average, waiter C poured 136 mL into each glass. Compute the Bonferroni corrected LSD ("least significant difference") value used to perform all possible pairwise comparisons between the three waiters, and determine where there are significant differences (both the LSD value and the conclusion must be correct). Use the significance level $\alpha = 5\%$.

1 ☐ $LSD_{0.05/3} = 7$ mL, so there is a significant difference between the expected amount of wine in glasses poured by waiters B and C, but no significant difference between waiters A and B or between waiters A and C.

2 ☐ $LSD_{0.05/3} = 7$ mL, so there is a significant difference between the expected amount of wine in glasses poured by waiters A and B as well as between waiters A and C, but no significant difference between waiters B and C.

3 ☐ $LSD_{0.05/3} = 4$ mL, so there is a significant difference between the expected amount of wine in glasses poured by waiters A and B and between waiters A and C, but no significant difference between waiters B and C.

4 ☐ $LSD_{0.05/3} = 17$ mL, so there is a significant difference between the expected amount of wine in glasses poured by waiters A and B and between waiters A and C, but no significant difference between waiters B and C.

5 ☐ $LSD_{0.05/3} = 4$ mL, so there is a significant difference between the expected amount of wine in glasses poured by waiters B and C, but no significant difference between waiters A and B or between waiters A and C.

Exercise VI

A spring is characterized by its spring constant, k. When a spring is stretched, Hooke's law states that

$$F = -k \cdot x\,,$$

where x is the length (in meters) by which the spring is extended, and F is the applied force (in Newtons). The following six observations were made for a given spring:

	1	2	3	4	5	6
x	0.22	0.24	0.26	0.28	0.30	0.32
F	-0.51	-0.85	-0.89	-1.59	-1.97	-2.06

The observations were read into two vectors in R, `x` (length) and `F` (force), respectively, after which the following model was estimated:

```r
model1 <- lm(F ~ x)
```

The output from `summary(model1)` is shown below, where some numbers are replaced by letters:

```
##
## Call:
## lm(formula = F ~ x)
##
## Residuals:
##        1        2        3        4        5        6
## -0.04484 -0.04146  0.25365 -0.10667 -0.15758  0.09690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2433     0.5483     A       C    **
## x           -16.8663     2.0148     B       D    **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1686 on 4 degrees of freedom
## Multiple R-squared:  0.946,Adjusted R-squared:  0.9325
## F-statistic: 70.08 on 1 and 4 DF,  p-value: 0.001114
```

Question VI.1 (11)

How may the statistical model corresponding to `model1` be described?

1 ☐ $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where Y_i represents the length by which the spring is extended when the force x_i is applied, and $\varepsilon_1, \ldots, \varepsilon_6$ are assumed to be independent and identically $N(0, \sigma^2)$-distributed.

2 ☐ $Y_i = \beta_1 x_i + \varepsilon_i$, where Y_i represents the force used to extend the spring by the length x_i, and $\varepsilon_1, \ldots, \varepsilon_6$ are assumed to be independent and identically $N(0, 1)$-distributed.

3 ☐ $Y_i = \beta_1 x_i + \varepsilon_i$, where Y_i represents the length by which the spring is extended when the force x_i is applied, and $\varepsilon_1, \ldots, \varepsilon_6$ are assumed to be independent and identically $N(0, 1)$-distributed.

4 ☐ $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where Y_i represents the length by which the spring is extended when the force x_i is applied, and $\varepsilon_1, \ldots, \varepsilon_6$ are assumed to be independent and identically $N(0, 1)$-distributed.

5 ☐ $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where Y_i represents the force used to extend the spring by the length x_i, and $\varepsilon_1, \ldots, \varepsilon_6$ are assumed to be independent and identically $N(0, \sigma^2)$-distributed.

Question VI.2 (12)

Based on the estimated slope in `model1`, give an estimate of the spring constant, k:

1 ☐ 0.5483

2 ☐ 3.2433

3 ☐ 2.0148

4 ☐ 16.8663

5 ☐ 5.2004

Question VI.3 (13)

It is of interest to test whether the model's intercept is significantly different from zero. Give the relevant test statistic:

1 ☐ -8.371

2 ☐ 5.915

3 ☐ 0.004

4 ☐ 0.548

5 ☐ 0.169

Question VI.4 (14)

What is the distribution of the test statistic used to test whether the model's slope can be assumed to be zero?

1 ☐ A t-distribution with 6 degrees of freedom.

2 ☐ A standard normal distribution.

3 ☐ An F-distribution with 6 degrees of freedom.

4 ☐ A normal distribution with mean zero and standard deviation 0.1686.

5 ☐ A t-distribution with 4 degrees of freedom.

Question VI.5 (15)

In a simple linear regression like the above, the estimators of the intercept and slope parameters are often correlated. When is this correlation zero?

1 ☐ When the standard deviation of the dependent variable is 1.

2 ☐ When the slope is estimated as zero.

3 ☐ When the average of the explanatory variable is zero.

4 ☐ When the standard deviation of the explanatory variable is 1.

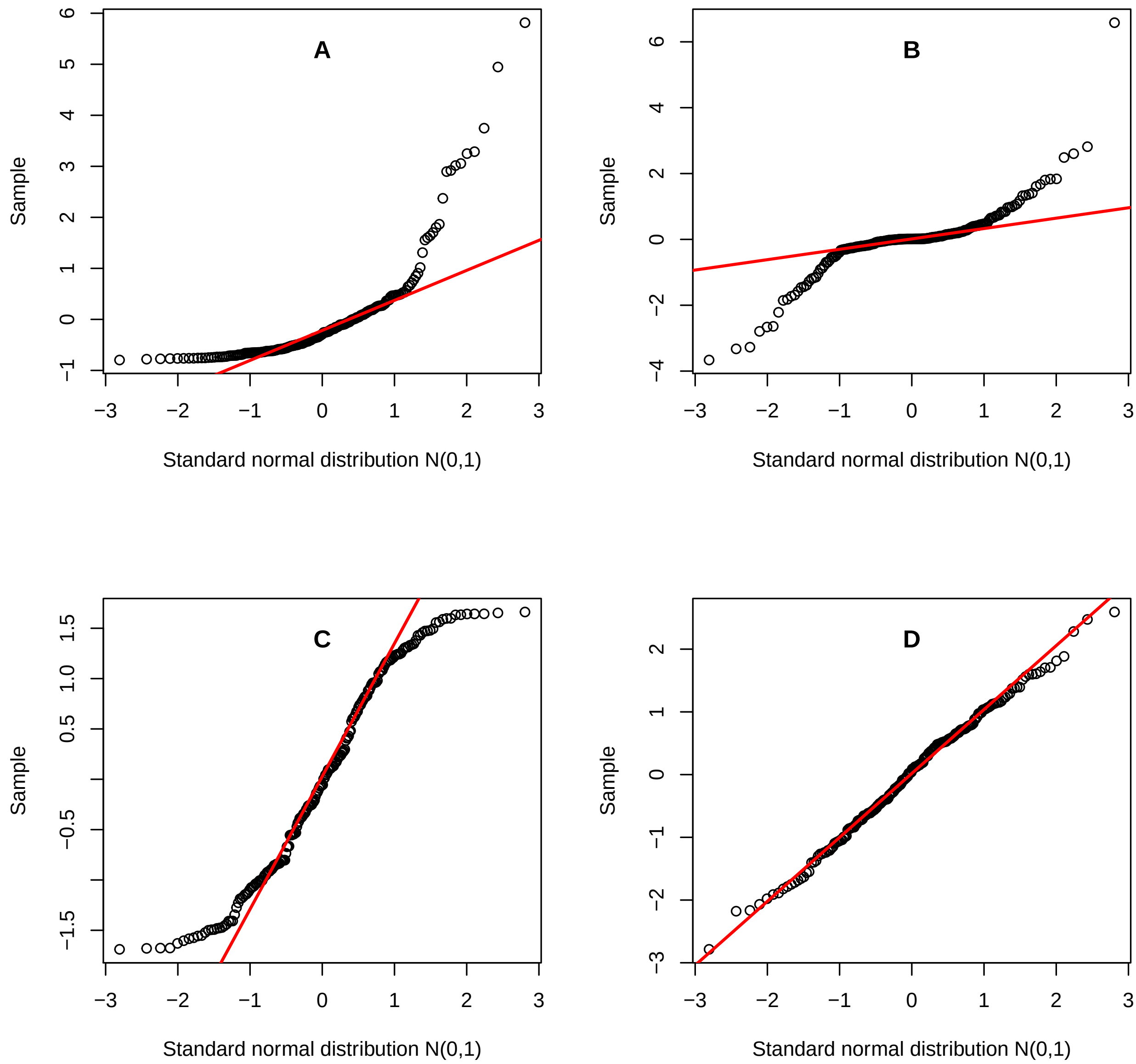
5 ☐ When the average of the dependent variable is zero.

Exercise VII

In order to investigate whether data from a single sample is log-normal distributed, one could compare the data to a normal distribution using a qq-plot. If the data is log-normal distributed there will (typically) be fewer small values and more large values in the data, compared to a normal distribution with the same mean and variance as the sample.

Question VII.1 (16)

Below, four qq-plots are shown in which four different samples with mean 0 and variance 1 are each compared to a standard normal distribution. Let $z_{0.25}$ and $z_{0.75}$ denote the first and third quartile of the standard normal distribution, respectively, while $q_{0.25}$ and $q_{0.75}$ denote the first and third quartile of the sample. The red line is drawn through the points $(z_{0.25}, q_{0.25})$ and $(z_{0.75}, q_{0.75})$. Which sample fulfills the above description of log-normal distributed data?



1 ☐ A

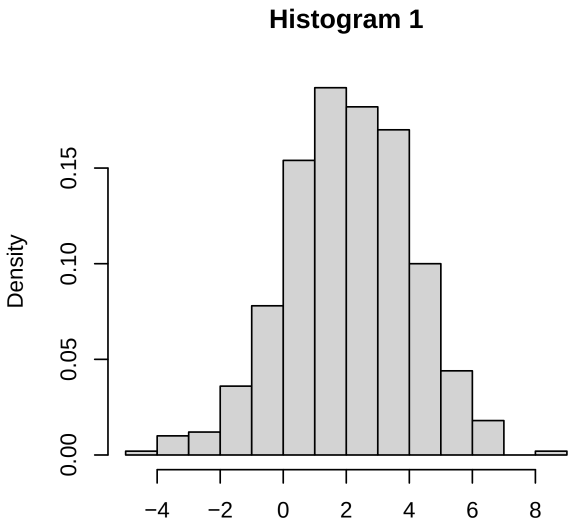
2 ☐ B

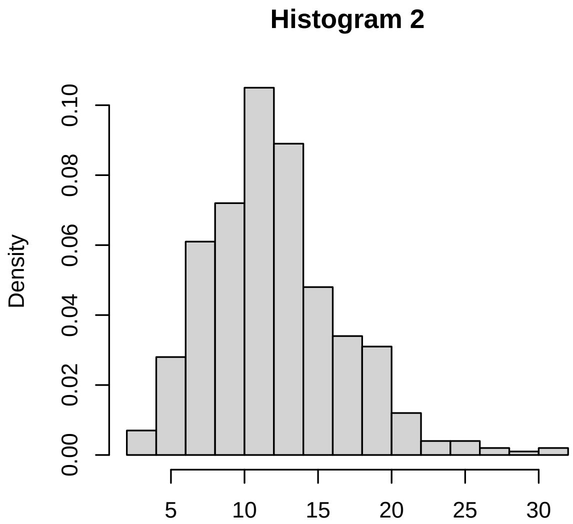
3 ☐ C

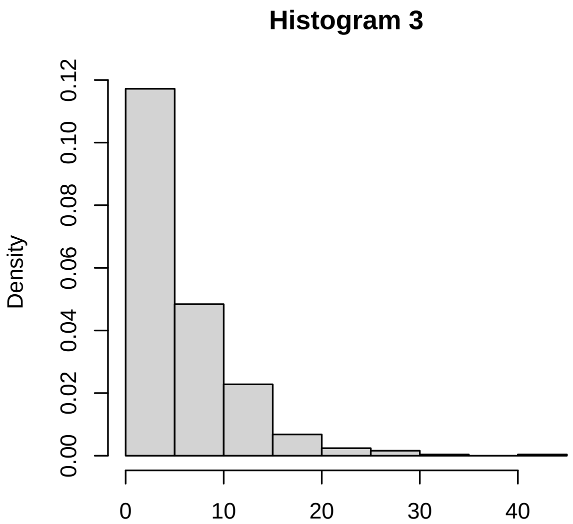
4 ☐ D

5 ☐ None of the above.

Exercise VIII







Question VIII.1 (17)

Which distributions are simulated above? ($N(\mu, \sigma^2)$ refers to the normal distribution with mean μ and variance σ^2, χ^2_a to the χ^2 distribution with a degrees of freedom, and $Exp(\beta)$ to the exponential distribution with rate β).

1 ☐ 1: $N(0, 4)$, 2: χ^2_{10}, 3: $Exp(1/5)$

2 ☐ 1: χ^2_4, 2: $N(2, 4)$, 3: χ^2_1.

3 ☐ 1: $N(2, 4)$, 2: χ^2_{12}, 3: $Exp(1/5)$

4 ☐ 1: $N(2,4)$, 2: $Exp(5)$, 3: χ^2_1

5 ☐ 1: $N(2,4)$, 2: χ^2_1, 3: $Exp(1/5)$

Exercise IX

Two groups of rats are put on a diet while growing up, and their weight gain between day 28 and day 84 is recorded. 10 rats are put on a diet with a high protein content, while 7 rats are put on a diet with a low protein content. The collected data (weight gain in grams) is shown in the table below, with the total weight gain in each group given in the last row:

	High protein content	Low protein content
	134	70
	146	118
	104	101
	119	85
	124	107
	161	132
	107	94
	83	
	113	
	129	
Total	1220	707

Using the numbers in the table, the sample variances in the two groups are calculated to be $s_H^2 = 495$ and $s_L^2 = 425$, where H and L indicate the groups with high and low protein content, respectively. It is further given that the usual test, for whether the expected weight gain is the same for rats on high and low protein diets, has 13.7 degrees of freedom.

Question IX.1 (18)

Which of the following choices is correct (both statements need to be correct)?

1 ☐ Rats in the low protein diet group gain more weight than rats in the high protein diet group. However, the difference is not statistically significant at the significance level $\alpha = 0.05$.

2 ☐ Rats in the high protein diet group gain more weight than rats in the low protein diet group. The difference is statistically significant at the significance level $\alpha = 0.05$.

3 ☐ Rats in the high protein diet group gain more weight than rats in the low protein diet group. The difference is statistically significant at the significance level $\alpha = 0.01$.

4 ☐ Rats in the low protein diet group gain more weight than rats in the high protein diet group. The difference is statistically significant at the significance level $\alpha = 0.05$.

5 ☐ Rats in the high protein diet group gain more weight than rats in the low protein diet group. However, the difference is not statistically significant at the significance level $\alpha = 0.05$.

Exercise X

Statistics Denmark provides data related to Denmark at `www.statistikbanken.dk`, among it data on traffic accidents. The following count data is taken from there:

Year	2010				2017			
Type	All		Alcohol		All		Alcohol	
Zone	City	Rural	City	Rural	City	Rural	City	Rural
Single-vehicle accidents	240	491	107	178	174	340	55	96
Others	1779	988	161	84	1456	819	106	48

Values under "all" count all accidents (including drunk-driving accidents) while numbers under "alcohol" include only drunk-driving accidents.

Question X.1 (19)

Give a 99% confidence interval for the total proportion of drunk driving accidents in 2010, where you use the relevant normal distribution approximation.

1 ☐ $0.848 \pm 2.58\sqrt{\frac{0.848}{3498}}$

2 ☐ $0.152 \pm 2.58\sqrt{\frac{0.848}{3498}}$

3 ☐ $0.848 \pm 1.96\sqrt{\frac{0.152 \cdot 0.848}{3498}}$

4 ☐ $0.848 \pm 2.58\sqrt{\frac{0.152}{3498}}$

5 ☐ $0.152 \pm 2.58\sqrt{\frac{0.152 \cdot 0.848}{3498}}$

Question X.2 (20)

Assume that the proportion of drunk-driving accidents in the "single-vehicle accidents" category is representative of the total proportion of drunk-driving. (Thus, data from the "others" category should *not* be used in this question).

Then, using the numbers from the table above and the wording from Table 3.1 of the book, what may be concluded about the difference in drunk driving between the years 2010 and 2017?

1 ☐ There is very strong evidence of a decrease in the proportion of drunk-driving accidents.

2 ☐ There is weak evidence of a decrease in the proportion of drunk-driving accidents.

3 ☐ There is little or no evidence of a difference in the proportion of drunk-driving accidents.

4 ☐ There is weak evidence of an increase in the proportion of drunk-driving accidents.

5 ☐ There is very strong evidence of an increase in the proportion of drunk-driving accidents.

Question X.3 (21)

From the same source, there is also data available on the speed limits for the road stretches where the accidents occurred. The following data, describing the number of rural zone accidents at different speed limits in the years 2010 and 2017, were extracted:

	2010	2017
0 to 50 km/h	54	58
50 to 100 km/h	1280	966
100 to 130 km/h	144	135

What is the result of the usual test for no change in the distribution of accidents in the speed limit intervals between the two years (both your conclusion and reasoning must be correct)? Use the significance level $\alpha = 1\%$.

1 ☐ No significant difference is found in the distribution of speed limits between the two years, as the p-value is larger than the significance level.

2 ☐ A significant difference is found in the distribution of speed limits between the two years, as the p-value is larger than the significance level.

3 ☐ A significant difference is found in the distribution of speed limits between the two years, as the p-value is smaller than the significance level.

4 ☐ No significant difference is found in the distribution of speed limits between the two years, as the p-value is smaller than the significance level.

5 ☐ None of the above statements are true.

Question X.4 (22)

In connection with the usual test for whether the distribution of speed limits is the same in the two years, the following question is asked: What is the estimated proportion of accidents on roads with speed limits from 50 to 100 km/h in 2017 under the null hypothesis?

1 ☐ $(58 + 966 + 135)/(54 + 1280 + 144 + 58 + 966 + 135) = 0.440$

2 ☐ $(966)/(54 + 1280 + 144 + 58 + 966 + 135) = 0.366$

3 ☐ $(966)/(58 + 966 + 135) = 0.833$

4 ☐ $(1280 + 966)/(54 + 1280 + 144 + 58 + 966 + 135) = 0.852$

5 ☐ $(54 + 58 + 144 + 135)/(54 + 1280 + 144 + 58 + 966 + 135) = 0.148$

Exercise XI

Below is a sample of 20 independent observations, read into R in the vector `x`:

```
x <-
  c(13, 12, 9, 7, 12, 15, 12, 10, 6, 13, 7, 13, 19, 12, 6, 4, 15, 16, 11, 18)
```

The data do not originate from a known distribution, but we are interested in the population mean and the uncertainty of its estimate.

Question XI.1 (23)

What is the sample mean $\bar{x}$ and variance s^2 (both quantities must be correct)?

1 ☐ $\bar{x} = 11.2$ and $s^2 = 16.7$.

2 ☐ $\bar{x} = 11.5$ and $s^2 = 16.7$.

3 ☐ $\bar{x} = 11.2$ and $s^2 = 4.1$.

4 ☐ $\bar{x} = 11.5$ and $s^2 = 4.1$.

5 ☐ $\bar{x} = 11.5$ and $s^2 = 16.7^2$.

Question XI.2 (24)

Now, we perform a resampling of `x` to get an idea of the uncertainty of the sample mean. We draw 200 resamples with replacement from the 20 observations in `x`, each with sample size 20. Subsequently, the mean of each of the 200 resamples is calculated. The R code for this operation is:

```
apply(replicate(200, sample(x, replace = TRUE)), 2, mean)
```

Below, the 10 largest and 10 smallest sample means of the 200 resamples are shown.

smallest	9.00	9.65	9.65	9.80	9.90	9.95	10.00	10.00	10.00	10.05
largest	12.95	12.95	12.95	13.00	13.05	13.10	13.10	13.10	13.15	13.40

Using the results above and the book's definition of percentiles ("`type = 2`" in R), which of the following is a 95% bootstrap confidence interval for the population mean?

1 ☐ $[10.05, 12.95]$

2 ☐ $[9.00, 13.40]$

3 ☐ $[9.80, 13.10]$

4 ☐ $[9.65, 13.10]$

5 ☐ $[9.925, 13.075]$

Exercise XII

During the preparation for a small festival, the toilet facilities are taken under consideration. Mobile toilets need to be ordered such that the capacity is sufficient, but not too high, since will lead to more cleaning and higher costs.

It is assumed that, on average, 150 guests need to use the toilets every hour, and that their arrival follows a Poisson distribution. In addition, it is assumed that each toilet can serve 20 guests per hour.

Question XII.1 (25)

Suppose that 10 toilets are ordered. What is then the probability that, in a randomly selected hour, the number of guests who arrive at the toilets exceeds the capacity?

1 ☐ 0.0042%

2 ☐ 2.3%

3 ☐ 11%

4 ☐ 24%

5 ☐ 99%

Question XII.2 (26)

A group of DTU students have decided to help small festivals optimize their logistical conditions. Among other things, the students have collected data on the use of toilets at small festivals. An examination of these data shows that a better model can be made to represent the number of guests who need to use the facilities in a randomly selected hour. This number can be modelled by an exponential distribution with mean $\frac{\text{“number of guests”}}{10}$, where “number of guests” is the total number of guests at the festival. In this question, this new model must be used.

A festival with 1500 guests is now considered. How many toilets should, at least, be ordered to ensure that the probability that not everyone can use the facilities is less than 2% in a randomly selected hour (given as a call in R)? It is still assumed that each toilet can serve 20 guests per hour.

1 ☐ `ppois(20, lambda = 15) * 20`

2 ☐ `qpois(0.98, lambda = 1500/10) / 20`

3 ☐ `qexp(0.98, rate = 10/15)`

4 ☐ `qexp(0.98, rate = 10/1500) / 20`

5 ☐ `qexp(0.98, rate = 10/1500) * 20`

Exercise XIII

Below, there's a small sample with 5 independent observations:

Observations:	11.8071067	-1.7913888	-9.1872410	-4.4860901	-0.2324924

Question XIII.1 (27)

Which of the following answer options is the only one that can possibly be correct?

1 ☐ It is impossible that the observations were sampled from a normal distribution with mean 0 and variance 10^2.

2 ☐ It is possible that the observations were sampled from a uniform distribution with parameters -9 and 12.

3 ☐ It is possible that the observations were sampled from a t-distribution with 1 degree of freedom.

4 ☐ It is possible that the observations were sampled from an F-distribution with 1 and 2 degrees of freedom.

5 ☐ It is possible that the observations were sampled from an exponential distribution with rate 1.

Exercise XIV

As the share of wind power in the energy production in Denmark and the rest of Europe increases, accurate predictions become more crucial. The prediction of wind energy naturally depends on the weather forecast, but also the estimation method and model structure have an influence. The table below shows a sample of a data set giving the average weekly predicted wind production for a wind farm (measured as a percentage of installed power) for different prediction models (m1, ..., m5).

Week	m1	m2	m3	m4	m5
1	0.6039	0.6232	0.6083	0.5751	0.6232
2	0.5143	0.5301	0.5049	0.4644	0.4850
3	0.5551	0.5603	0.5415	0.5091	0.5219
4	0.5396	0.5393	0.5766	0.4697	0.5245
⋮	⋮	⋮	⋮	⋮	⋮

Below, the data is loaded into R. The vector `prediction` contains the average weekly predictions, `model` indicates which of the five prediction models was used, and `week` indicates the week number.

One wants to investigate whether the 5 different models (m1, ..., m5) can be assumed to give the same expected predictions (on a weekly basis) or whether there is a significant difference.

To investigate the hypothesis, the following model has been formulated

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where α_i describes the effect of model i, and β_j describes the effect of week j.

Question XIV.1 (28)

Under the usual assumptions, which of the following statements is correct?

1 ☐ $\alpha_i + \beta_j = 0$ for all combinations of i and j.

2 ☐ ϵ_{ij} are independent and normal distributed with mean 0 and a variance which depends on α_i and β_j.

3 ☐ Y_{ij} are independent and normal distributed with the same variance for all combinations of i and j.

4 ☐ $\sum_i \alpha_i = \sum_j \beta_j = \mu$.

5 ☐ Y_{ij} are independent and identically distributed for all combinations of i and j.

Question XIV.2 (29)

To investigate the hypothesis that there is no difference in the expected predictions, the following R-code was run. Note that some of the results have been removed, and some numbers have been replaced by letters.

```
anova(lm(prediction ~ model+factor(week),data=dat))

## Analysis of Variance Table
##
## Response: pred
##               Df  Sum Sq   Mean Sq F value    Pr(>F)
## model          4 0.01056 0.0026391  7.3754 5.389e-05 ***
## factor(week) 17 0.33946 0.0199684 55.8051 < 2.2e-16 ***
## Residuals    68 0.02433 0.0003578
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How many weeks are there in the data set? (Due to rounding in the R-output, the relevant calculation will result in a decimal number which must be rounded correctly to the nearest integer).

1 ☐ 18

2 ☐ 68

3 ☐ 56

4 ☐ 17

5 ☐ 55

Question XIV.3 (30)

Consider again the R-output from the previous question, and use the significance level $\alpha = 0.01$. Is there a significant difference between the five models $m_1, \ldots, m_5$, when the statistical model takes into account the difference between weeks (both the conclusion and argument must be correct)?

1 ☐ No, as $0.0106 > 0.01$.

2 ☐ Yes, as $5.389 \cdot 10^{-5} < 0.01$.

3 ☐ No, as $0.020 > 0.01$.

4 ☐ Yes, as $0.0026 < 0.01$.

5 ☐ Yes, as $0.024 > 0.01$.

The exam paper is finished. Have a great Christmas vacation!