Technical University of Denmark *Page 1 of 29 pages.*

Written examination: 14. August 2016

Course name and number: **Introduction to Statistics (02323, 02402 og 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 11 exercises. To answer the questions you need to fill in the prepared 30-question multiple choice form (on three seperate pages) in CampusNet

5 points are given for a correct answer and -1 point is given for a wrong answer. ONLY the following 5 answer options are valid: 1, 2, 3, 4 or 5. If a question is left blank or another answer is given, then it does not count (i.e. "0 points"). Hence, if more than one answer option is given to a single question, which in fact is technically possible in the online system, it will not count (i.e. "0 points"). The number of points corresponding to specific marks or needed to pass the examination is ultimately determined during censoring.

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

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

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

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

The questionnaire contains 29 pages.

Continues on page 2

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

Exercise I

A power company has developed an app that can help their consumers to analyze and reduce their electricity consumption. It must now be tested whether users of the app have reduced their electricity consumption after they have installed it. The electricity consumption is measured in kWh. Let X denote the difference in electricity consumption between the month before they installed the app (X_{before}) and electricity consumption the month after they installed the app (X_{after}), such that

$$X = X_{\text{after}} - X_{\text{before}}$$

The difference is registered for 40 randomly selected users who have installed the app at different times of the year. The sample average and sample standard deviation are calculated to

$$\bar{x} = -22.6$$
$$s_X = 45.5$$

Question I.1 (1)

Calculate a 95% confidence interval for difference in mean μ_X in electricity consumption from before to after the app was installed:

1 ☐ $-11.3 \pm 2.02 \cdot \frac{45.5}{39} = [-13.7, -8.94]$

2 ☐ $-22.6 \pm 2.02 \cdot \frac{45.5}{39} = [-25.0, -20.2]$

3 ☐ $-22.6 \pm 2.02 \cdot \frac{45.5}{6.32} = [-37.1, -8.06]$

4 ☐ $-22.6 \pm 2.02 \cdot \frac{2070}{6.32} = [-684, 639]$

5 ☐ $-22.6 \pm 2.02 \cdot \frac{2070}{39} = [-130, 84.6]$

Question I.2 (2)

Is there a significant decrease in electricity consumption from from the month before to the month after the installation of the app at 5% significance level (both the conclusion and reasoning (p-value) must be correct)?

1 ☐ Yes, a significant decrease can be detected, since the p-value for the obvious two-sided test is 0.027

2 ☐ No, a significant decrease cannot be detected, since the p-value for the obvious two-sided test is 0.027

3 ☐ Yes, a significant decrease can be detected, since the p-value for the obvious two-sided test is 0.0032

4 ☐ No, a significant decrease cannot be detected, since the p-value for the obvious two-sided test is 0.0032

5 ☐ No, a significant decrease cannot be detected, since the p-value for the obvious two-sided test is 0.21

Question I.3 (3)

It has been found that some consumers who install the app don't start to use the app right away after installation. Therefore, the onboarding (the process the user must go through the first time the app is opened after installation) has been redesigned to get users started faster. If the probability that a user doesn't get started right away is set to $p = 0.20$ and 100 new users are registered, and X denotes the number of those who doesn't get started right away, then find the one of the following R expressions, which calculates the probability of getting less than 10 new users who doesn't get started away, i.e. $P(X < 10)$?

1 ☐ `phyper(q=1, m=20, n=80, k=10)`

2 ☐ `pbinom(q=9, size=100, prob=0.2)`

3 ☐ `dbinom(x=10, size=100, prob=0.2)`

4 ☐ `1 - pbinom(q=10, size=100, prob=0.2)`

5 ☐ `dbinom(x=10, size=100, prob=0.2)`

Continues on page 4

Exercise II

In a series of experiments it has been investigated how the compressive strength of concrete depends on the composition of the concrete. The registered explanatory variables are the amount of cement, water and sand measured in kg/m^3. The compressive strength is measured in MPa. A summary of the data is given in the following table:

	Cement	Water	Fine	Strength
Min.	200.0	146.0	594.0	12.25
1st Qu.	289.0	185.0	754.0	22.49
Median	339.0	186.0	781.0	31.35
Mean	344.9	188.5	776.4	31.83
3rd Qu.	393.0	192.0	809.0	37.42
Max.	540.0	228.0	945.0	74.99

Question II.1 (4)

First a simple linear regression model with the amount of cement as an explanatory variable is fitted and the following output from R is obtained (where a few numbers are replaced by letters):

```
## Call:
## lm(formula = Strength ~ Cement, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.5512  -0.6858   0.6280   1.4791  20.8302
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.736438   3.389030     A   1.96e-05 ***
## Cement        0.137930   0.009567     B   < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.05 on 59 degrees of freedom
## Multiple R-squared:  0.7789,Adjusted R-squared:  0.7752
## F-statistic: 207.9 on 1 and 59 DF,  p-value: < 2.2e-16
```

The relevant test statistic for testing if the amount of cement has a significant effect on the compressive strength is found to?

1 ☐ $0.009567/0.137930 \approx 0.06936$

2 ☐ $1.96e^{-5}$

3 ☐ $-15.736438/3.389030 \approx -4.643$

4 ☐ $0.137930/0.009567 \approx 14.42$

5 ☐ 6.05

Question II.2 (5)

The following model has been fitted

$$Strength_i = \beta_0 + \beta_1 \cdot Cement_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

and it is wanted to use the model fit for predicting the compressive strength given the amount of cement.

The prediction has the lowest variance when the amount of cement is:

1 ☐ 339.0 kg/m^3

2 ☐ 344.9 kg/m^3

3 ☐ 540.0 kg/m^3

4 ☐ 0.0 kg/m^3

5 ☐ Cannot be answered based on the provided information

Question II.3 (6)

An equivalent analysis for the dependence between the amount of water and the compressive strength is carried out. The following R output is obtained:

```
## Call:
## lm(formula = Strength ~ Water, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.361  -8.593  -1.161   5.239  28.568
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 116.1345    23.6644   4.908 7.62e-06 ***
## Water        -0.4474     0.1253  -3.570 0.000718 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 11.67 on 59 degrees of freedom
## Multiple R-squared:  0.1776,Adjusted R-squared:  0.1637
## F-statistic: 12.74 on 1 and 59 DF,  p-value: 0.0007184
```

If the residuals meet the usual assumptions, then the conclusion of the analysis is:

1 ☐ Water doesn't have a significant effect on the compressive strength. More water results in stronger concrete

2 ☐ Water doesn't have a significant effect on the compressive strength. Less water results in stronger concrete

3 ☐ Water has a significant effect on the compressive strength. More water results in stronger concrete

4 ☐ Water doesn't have a significant effect on the compressive strength. Therefore it cannot be determined how the amount of water affects the compressive strength

5 ☐ Water has a significant effect on the compressive strength. Less water results in stronger concrete

Question II.4 (7)

It is chosen to estimate a multiple linear regression model using the square root of the compressive strength as response. The other three variables are used as explanatory variables:

```
## Call:
## lm(formula = sqrt(Strength) ~ Cement + Water + Fine, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31854 -0.12044  0.06208  0.18913  0.73313
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.0623090  1.3125406   3.857 0.000295 ***
## Cement       0.0120327  0.0007024  17.131  < 2e-16 ***
## Water       -0.0244132  0.0037730  -6.470 2.41e-08 ***
## Fine         0.0012022  0.0009038   1.330 0.188774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3293 on 57 degrees of freedom
## Multiple R-squared:  0.9069,Adjusted R-squared:  0.902
## F-statistic: 185.1 on 3 and 57 DF,  p-value: < 2.2e-16
```

Which of the following conclusions from the above output is most correct:

1 ☐ A model which only has "Cement" as explanatory variable should be fitted

2 ☐ The variable "Fine" should be removed and the reduced model should be fitted

3 ☐ The variable "(Intercept)" should be removed and the reduced model should be fitted

4 ☐ The variable "Water" should be removed and the reduced model should be fitted

5 ☐ A model which only has "Fine" as explanatory variable should be fitted

Question II.5 (8)

Based on the presented R-outputs in the exercise the number of observations used in the analyses is found to:

1 ☐ 61

2 ☐ 60

3 ☐ 59

4 ☐ 2

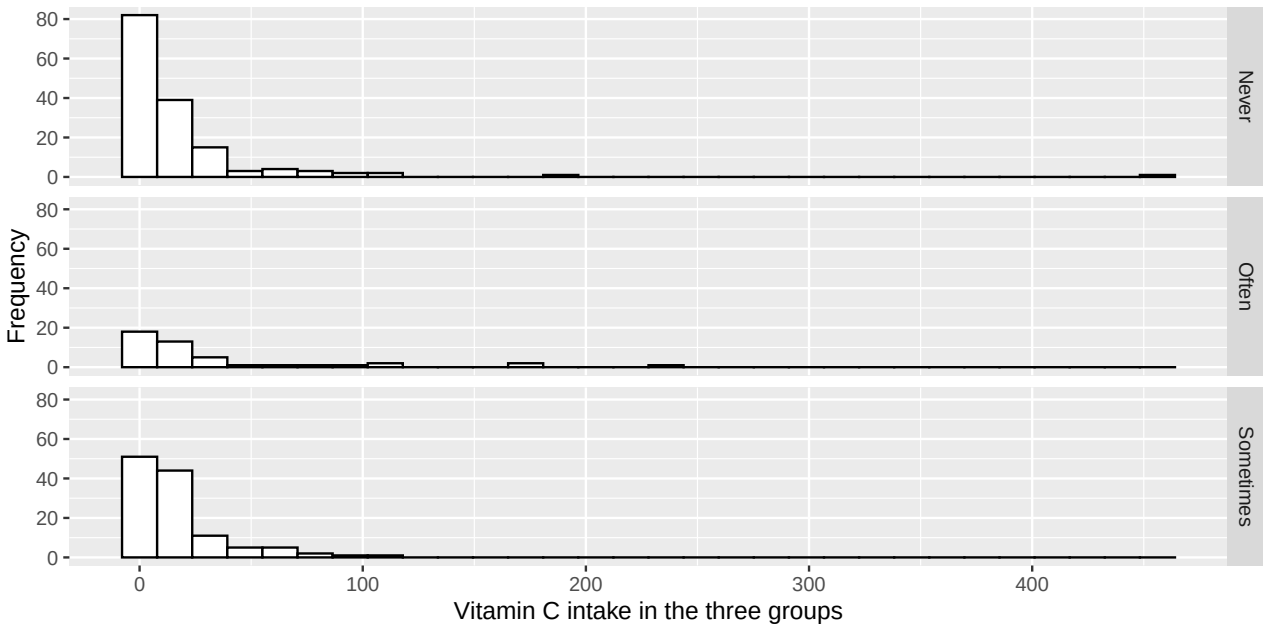
5 ☐ 1

Continues on page 8

Exercise III

A study has been conducted of the relation between intake of vitamin C and frequency of exercise. Below is a table of descriptive statistics for the intake of vitamin C in the three exercise groups as well as histograms of the observations from the experiments:

	Exercise		
	Never	Sometimes	Often
Mean	20.11	16.71	33.26
Median	7.23	9.27	11.27
SD	44.05	20.55	52.42
n	152	120	45



We want to study whether there is a dependence between the intake of vitamin C and the frequency of exercise. To do this a one-way analysis of variance is conducted. Where `z` is the vitamin C intake and `Group` is a factor with three levels (Never exercise, Sometimes exercise, Often exercise).

Two analyses have been conducted by running the following R-code (the reading of data is not shown):

```
anova(lm(z ~ Group, data = dat))

## Analysis of Variance Table
##
## Response: z
##            Df Sum Sq Mean Sq F value Pr(>F)
## Group       2   9060  4529.8   3.064 0.0481 *
```

```
## Residuals 314 464227  1478.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(log(z) ~ Group, data = dat))
```

```
## Analysis of Variance Table
##
## Response: log(z)
##            Df Sum Sq Mean Sq F value Pr(>F)
## Group       2   2.94  1.4693   0.918 0.4004
## Residuals 314 502.53  1.6004
```

Question III.1 (9)

Which of the following statements is correct?

1 ☐ The analysis of `z` is the most correct since it is seen from the table that the intake of vitamin C is higher in the group who exercise often

2 ☐ The analysis of `z` is the most correct since we want to determine the effect on vitamin C and not the effect on the logarithm of vitamin C

3 ☐ The analysis of `log(z)` is the most correct since here the effect of Group was not statistically significant

4 ☐ The analysis of `log(z)` is the most correct since the histograms indicate that the assumption of normality is not be valid for `z`

5 ☐ The analysis of `z` is the most correct since here the effect of Group was statistically significant

Continues on page 10

Exercise IV

In a study of high blood pressure it is investigated whether two different diets influences the blood pressure. The study was conducted with 150 persons which were divided in 2 groups of 75 people. Group I is a control group receiving normal diet, while group II received healthy dietary supplements. The results are shown in the following table; a certain level of the blood pressure (or above) is defined as 'high' (not necessarily too high):

Group	Normal blod pressure	High blod pressure	Total
I	55	20	75
II	57	18	75
Total	112	38	150

Question IV.1 (10)

A 95% confidence interval for the proportion of persons on normal diet (Group I) with high blood pressure is:

1 ☐ $0.49 \pm 1.96\sqrt{\frac{0.49107\cdot(1-0.49107)}{112}}$ giving 0.49 ± 0.09

2 ☐ $0.49 \pm 1.645\sqrt{\frac{0.49107\cdot(1-0.49107)}{55}}$ giving 0.49 ± 0.11

3 ☐ $0.27 \pm 1.645\sqrt{\frac{0.26667\cdot(1-0.26667)}{112}}$ giving 0.27 ± 0.07

4 ☐ $0.27 \pm 1.96\sqrt{\frac{0.49107\cdot(1-0.49107)}{75}}$ giving 0.27 ± 0.09

5 ☐ $0.27 \pm 1.96\sqrt{\frac{0.26667\cdot(1-0.26667)}{75}}$ giving 0.27 ± 0.10

Question IV.2 (11)

Which of the following computed values is a usual test statistic for a test of the hypothesis that the proportions of people with high blood pressure in Groups I and II are the same?

1 ☐ $1.96^2 = 3.84$

2 ☐ $\frac{1}{56} + \frac{1}{19} = 0.07$

3 ☐ $\frac{1}{56} + \frac{1}{56} + \frac{1}{19} + \frac{1}{19} = 0.14$

4 ☐ $\frac{55-57}{\sqrt{150}} = 0.16$

5 ☐ $\frac{(112-38)^2}{150} = 36.5$

Question IV.3 (12)

Suppose you had compared the proportions between 3 diet groups, that is, had a data table in the following form (e.g. x_{32} denotes the number of people in Group III, who had high blood pressure):

Group	Normal blod pressure	High blod pressure
I	x_{11}	x_{12}
II	x_{21}	x_{22}
III	x_{31}	x_{32}

Which distribution would be used in the usual hypothesis test for a comparison of the three proportions (i.e. the proportions of persons in each of the three groups with high blood pressure)?

1 ☐ A χ^2-distribution with 3 degrees of freedom

2 ☐ An F-distribution with 2 and 3 degrees of freedom

3 ☐ An F-distribution with 2 and 5 degrees of freedom

4 ☐ A χ^2-distribution with 2 degrees of freedom

5 ☐ A χ^2-distribution with 1 degree of freedom

Question IV.4 (13)

This question is based on the text and the table given in the previous question.

What is a pre-planned 95% confidence interval for the difference between the proportions with normal blood pressure in Group I and Group III?

1 ☐ $\frac{x_{11}}{x_{11}+x_{12}} - \frac{x_{31}}{x_{31}+x_{32}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{(x_{11}+x_{12})^3} + \frac{x_{31}\cdot x_{32}}{(x_{31}+x_{32})^3}}$

2 ☐ $\frac{x_{11}}{x_{12}} - \frac{x_{31}}{x_{32}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{(x_{11}+x_{12})^3} + \frac{x_{31}\cdot x_{32}}{(x_{31}+x_{32})^3}}$

3 ☐ $\frac{x_{11}}{x_{12}} - \frac{x_{31}}{x_{32}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{x_{11}+x_{12}} + \frac{x_{31}\cdot x_{32}}{x_{31}+x_{32}}}$

4 ☐ $\frac{x_{11}}{x_{11}+x_{21}+x_{31}} - \frac{x_{31}}{x_{11}+x_{21}+x_{31}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{(x_{11}+x_{21}+x_{31})^2} + \frac{x_{31}\cdot x_{32}}{(x_{11}+x_{21}+x_{31})^2}}$

5 ☐ $\frac{x_{11}}{x_{11}+x_{12}} - \frac{x_{31}}{x_{31}+x_{32}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{(x_{11}+x_{21}+x_{31})^2} + \frac{x_{31}\cdot x_{32}}{(x_{11}+x_{21}+x_{31})^2}}$

Question IV.5 (14)

A new study is planned to explore a completely new diet. The required precision is that the 90% confidence interval for the proportion of people with normal blood pressure achieves a mean width of 0.1. The total cost for handling a single person is 100 kr, and there is assigned in total 25000 kr for this in the budget for the study. Can the requirement be fulfilled within the given budget (note, you know nothing about the value of the proportion)?

1 ☐ Yes, since $100 \cdot \left(\frac{1.96}{0.05}\right)^2 = 153664 < 25000$

2 ☐ Yes, since $100 \cdot 0.3 \cdot \cdot 0.7 \cdot \left(\frac{1.645}{0.05}\right)^2 = 22730.61 < 25000$

3 ☐ No, since $\frac{1}{4} \cdot \left(\frac{1.96}{0.05}\right)^2 \approx 384$

4 ☐ Yes, since $\frac{1}{4} \cdot \left(\frac{1.96}{0.05}\right)^2 \approx 384$

5 ☐ No, since $100 \cdot \frac{1}{4} \cdot \left(\frac{1.645}{0.05}\right)^2 = 27060.25 > 25000$

Continues on page 13

Exercise V

Question V.1 (15)

Let $X_i \sim N(\mu, \sigma^2)$, $i = 1, ..., n$. If $n = 10$ what is then

$$P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} > 2\right)$$

where S^2 and $\overline{X}$ is the empirical variance and average, respectively (random variables)?

1 ☐ $6.8 \cdot 10^{-5}$

2 ☐ 0.038

3 ☐ 0.012

4 ☐ 0.988

5 ☐ 0.962

Continues on page 14

Exercise VI

Question VI.1 (16)

A multiple choice exam have 30 questions and 5 answer options for each question. There is one and only one correct answer for each question. What is the probability of answering correct on to at least 15 of the 30 questions, if you answer completely at random?

1 ☐ 0.000119

2 ☐ 0.00023

3 ☐ 0.835

4 ☐ 0.999

5 ☐ $4.5 \cdot 10^{-9}$

Question VI.2 (17)

Which of the following random variables should not be approximated by a normal distribution?

1 ☐ $\sum_{i=1}^{200} X_i,\ X_i \sim Pois(2)$

2 ☐ $\sum_{i=1}^{5} X_i,\ X_i \sim N(\mu, \sigma^2)$

3 ☐ $\sum_{i=1}^{10} X_i,\ X_i \sim Binom(1000, 0.5)$

4 ☐ $\sum_{i=1}^{100} X_i,\ X_i \sim Exp(1)$

5 ☐ $\sum_{i=1}^{7} X_i,\ X_i \sim Exp(1)$

Continues on page 15

Exercise VII

The lifetime of a particular type of electronic buttons is assumed to follow an exponential distribution with a mean of 5 years.

Question VII.1 (18)

What is the variance of the lifetime of such buttons?

1 ☐ 1 years

2 ☐ 1/5 years

3 ☐ 5 years

4 ☐ 25 years

5 ☐ 1/25 years

Question VII.2 (19)

If 10 of such buttons are installed in different systems (without interaction), what is then the probability that none of these breaks down within the first year?

1 ☐ 0.1353

2 ☐ 0.4350

3 ☐ 0.1074

4 ☐ 0.3758

5 ☐ 0.6241

Continues on page 16

Exercise VIII

An experiment is carried out to examine four different methods (A, B, C, D) for removing impurities in a chemical process. At the same time it is wanted to adjust for using three different reactors in the experiment. The data are shown in the table below:

	Reactor 1	Reactor 2	Reactor 3	Sum
Method A	23.97	29.54	37.91	91.42
Method B	12.67	17.48	20.28	50.43
Method C	25.85	40.09	38.00	103.94
Method D	21.29	23.58	20.19	65.06
Sum	83.78	110.69	116.38	310.85

The sums of squares have been calculated:

	SS
Reactor	151.61
Method	593.40
Residual	100.73
Total variation	845.74

Question VIII.1 (20)

We now want to investigate whether it is reasonable to assume that the four methods clean equally well. Using the above sums of squares the usual test statistic, here denoted by A, as well as the critical value (the level of significance is $\alpha = 0.05$), is found to:

1 ☐ $A = 5.89$ and $A < 8.94$ (i.e. no significant effect of Method)

2 ☐ $A = 11.78$ and $A > 4.76$ (i.e. a significant effect of Method)

3 ☐ $A = 441.79$ and $A > 4.76$ (i.e. a significant effect of Method)

4 ☐ $A = 0.70$ and $A < 8.94$ (i.e. no significant effect of Method)

5 ☐ $A = 3.91$ and $A < 4.76$ (i.e. no significant effect of Method)

Question VIII.2 (21)

A model for two-way analysis of variance will often be formulated for this type of experiment by

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma^2),\ i = (1, 2, 3, 4),\ j = (1, 2, 3)$$

Here y_{ij} is the observed purity for method i reactor j, μ the overall mean, α the method-effect and β the reactor-effect.

What is the usual estimate of the effect of Method B (i.e. $\hat{\alpha}_2$) in the model?

1 ☐ $\hat{\alpha}_2 = 50.43$

2 ☐ $\hat{\alpha}_2 = -260.42$

3 ☐ $\hat{\alpha}_2 = 16.81$

4 ☐ $\hat{\alpha}_2 = 4.20$

5 ☐ $\hat{\alpha}_2 = -9.09$

Question VIII.3 (22)

Before the experiment was conducted it was decided to compare Method B and Method C.

What is the 95% confidence interval for the comparison of B and C, if this is the only post-hoc comparison to be carried out?

1 ☐ $(50.43 - 103.94)/3 \pm 2.3060\sqrt{\frac{100.73}{12-4}(2/3)} = [-24.52, -11.16]$

2 ☐ $(50.43 - 103.94) \pm 2.4469\sqrt{\frac{100.73}{6}(2/3)} = [-61.70, -45.32]$

3 ☐ $(50.43 - 103.94)/3 \pm 2.4469\sqrt{\frac{100.73}{6}(1/3 + 1/4)} = [-25.49, -10.18]$

4 ☐ $(50.43 - 103.94)/3 \pm 2.4469\sqrt{\frac{100.73}{6}(2/3)} = [-26.02, -9.65]$

5 ☐ $(50.43 - 103.94)/3 \pm 2.3060\sqrt{\frac{100.73}{12-4}(1/3 + 1/4)} = [-24.09, -11.59]$

Question VIII.4 (23)

If it was decided to make all pairwise comparisons between the methods, what is then the Bonferroni corrected "least significant difference (LSD)"?

1 ☐ $3.8630\sqrt{2 \cdot 16.788/3} = 12.92$

2 ☐ $3.4789\sqrt{2 \cdot 12.591/3} = 10.08$

3 ☐ $2.4469\sqrt{2 \cdot 16.788/6} = 5.79$

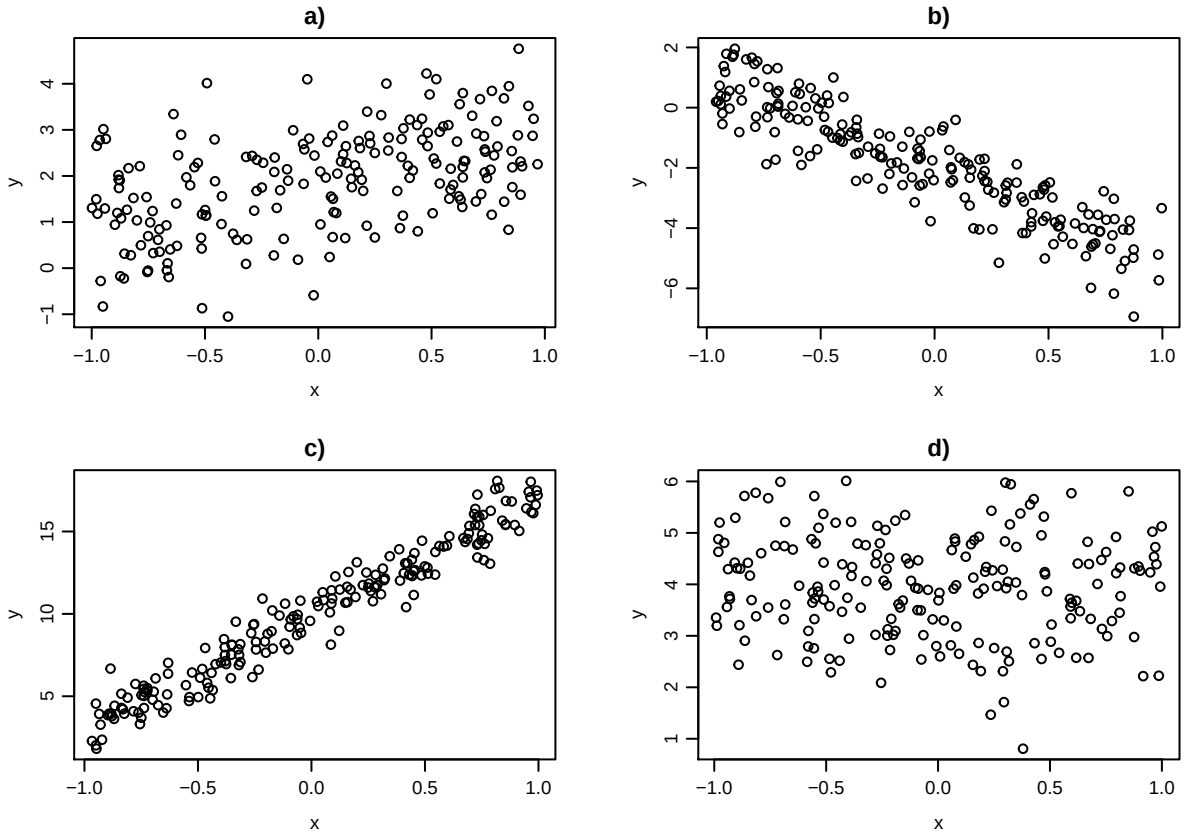
4 ☐ $3.8630\sqrt{2 \cdot 12.591/3} = 11.19$

5 ☐ $2.4469\sqrt{2 \cdot 16.788/3} = 8.19$

Continues on page 18

Exercise IX

Below are four scatter plots of y and x observations:



Question IX.1 (24)

Which four correlation coefficients (in the order a), b), c), d)) fits best with the observations in the figure?

1 ☐ 0.9, -0.5, 0.65, 0

2 ☐ 0.5, -0.9, 0.97, 0

3 ☐ 0.5, -0.9, 0.65, 0

4 ☐ 0.5, 0.97, 0, -0.9

5 ☐ 0.97, -0.9, 0, 0.5

Question IX.2 (25)

From inspection of plot `c)` in the figure, which estimates of the parameters in the usual simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$ fits best to those observations in plot `c)`?

1 ☐ $\hat{\beta}_0 = 10$, $\hat{\beta}_1 = 5$ and $\hat{\sigma} = 10$

2 ☐ $\hat{\beta}_0 = 16$, $\hat{\beta}_1 = -5$ and $\hat{\sigma} = 1$

3 ☐ $\hat{\beta}_0 = 10$, $\hat{\beta}_1 = 7$ and $\hat{\sigma} = 10$

4 ☐ $\hat{\beta}_0 = -10$, $\hat{\beta}_1 = -7$ and $\hat{\sigma} = 5$

5 ☐ $\hat{\beta}_0 = 10$, $\hat{\beta}_1 = 7$ and $\hat{\sigma} = 1$

Continues on page 21

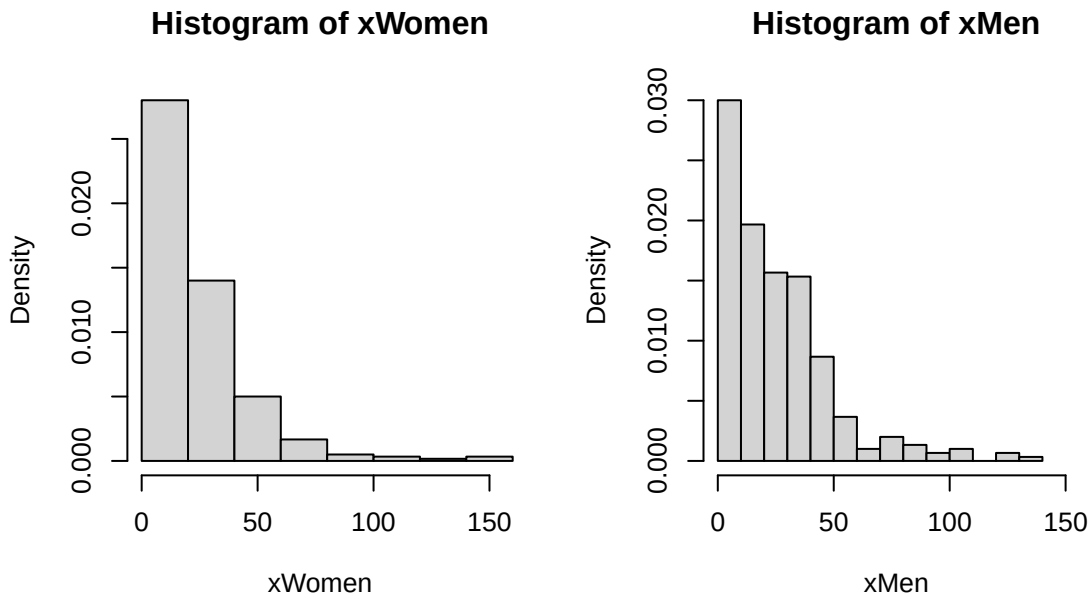
Exercise X

DTU Food monitors and provides analyses of Danish dietary habits. They have done so since 1985 at the request of the Ministry of Environment and Food of Denmark, and the results are obtained on the basis of surveys. The results are used as input for adjustments to the official dietary recommendations and the basis for a political priority of prevention efforts.

In the surveys the respondents were asked how much fish they eat in grams of fish per day. A random sample of 300 female respondents is taken from the dietary survey from 2005 to 2008, and equivalently also a random sample of 300 male respondents from the same dietary survey.

The observations are read into a vector for women `xWomen` and a vector for men `xMen`. The empirical density (density histogram) for the observations for each gender is plotted:

```
par(mfrow=c(1,2), cex=0.8)
hist(xWomen, prob=TRUE, xlim=range(xWomen,xMen))
hist(xMen, prob=TRUE, xlim=range(xWomen,xMen))
```



It is found that the assumption of normal distribution is not met. Therefore no assumption of distribution should be made in the analysis.

Question X.1 (26)

To determine the 95% confidence interval for the mean intake of fish per day for women in the population on the basis of the sample from the 2005-2008 dietary survey, the following R-code is run (of which everything might not necessarily be meaningful):

```r
## Number of simulated samples
k <- 10000

simsamples <- replicate(k, sample(xWomen, replace=TRUE))
simprms <- apply(simsamples, 2, mean)
quantile(simprms, c(0.025,0.975))

##  2.5% 97.5%
##  19.0  24.3

simsamples <- replicate(k, rnorm(length(xWomen),
                                 mean=log(mean(xWomen)), sd=sd(xWomen)))
simprms <- apply(simsamples, 2, mean)
quantile(simprms, c(0.025,0.975))

##  2.5% 97.5%
## 0.348 5.777

simsamples <- replicate(k, sample(xWomen, replace=TRUE))
simprms <- apply(simsamples, 2, median)
quantile(simprms, c(0.005,0.995))

##  0.5% 99.5%
##  11.1  20.6

simsamples <- replicate(k, sample(xMen, replace=TRUE))
simprms <- apply(simsamples, 2, median)
quantile(simprms, c(0.005,0.995))

##  0.5% 99.5%
##  17.0  25.2

simsamples <- replicate(k, runif(100,0,40))
simprms <- apply(simsamples, 2, mean)
quantile(simprms, c(0.025,0.975))

##  2.5% 97.5%
##  17.7  22.3
```

Based on the above outputs, what is then a correct 95% confidence interval for the mean intake of fish per day for women in the population?

1 ☐ $[19.0, 24.3]$

2 ☐ $[0.348, 5.777]$

3 ☐ $[11.1, 20.6]$

4 ☐ $[17.0, 25.2]$

5 ☐ $[17.7, 22.3]$

Question X.2 (27)

Note, that an assumption of normal distribution of the intake of fish is not met and no transformations are found to be reasonable to meet such an assumption.

In the following, we want to investigate whether men and women eat significant different amounts of fish per day, corresponding to the following hypothesis:

$$H_0 : q_{0.5,\text{male}} = q_{0.5,\text{female}}$$
$$H_1 : q_{0.5,\text{male}} \neq q_{0.5,\text{female}}$$

where $q_{0.5,gender}$ denotes the 50% quantile for the specified gender.

In order to test the hypothesis the following R-code has been run (note that everything not necessarily is meaningful):

```
t.test(xMen-xWomen)

##
##  One Sample t-test
##
## data:  xMen - xWomen
## t = 2.0508, df = 299, p-value = 0.04115
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.1553667 7.5300188
## sample estimates:
## mean of x
##  3.842693

t.test(xMen, xWomen)

##
##  Welch Two Sample t-test
##
## data:  xMen and xWomen
## t = 1.9908, df = 597.95, p-value = 0.04695
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.05193859 7.63344693
## sample estimates:
## mean of x mean of y
##  25.37784  21.53515
```

```
simprms <- replicate(k, median(sample(xMen, replace=TRUE))-
                     median(sample(xWomen, replace=TRUE)))
quantile(simprms, c(0.025, 0.975))

##   2.5%  97.5%
## 0.0354 9.8422

simprms <- replicate(k,
                     mean(rnorm(length(xMen),mean(xMen),sd(xMen))) -
                     mean(rnorm(length(xWomen),mean(xWomen),sd(xWomen))))
quantile(simprms, c(0.025, 0.975))

##    2.5%   97.5%
## -0.0544  7.5434

median(xMen) > median(xWomen)

## [1] TRUE
```

At a significance level $\alpha = 0.05$, what is then the conclusion of the hypothesis test (both the conclusion and argumentation must be correct)?

1 ☐ Since p-value $= 0.04115 < 0.05$ it can be rejected that $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

2 ☐ Since $0 \in [-0.0544, 9.842]$ it cannot be rejected that $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

3 ☐ Since $0 \notin [0.0354, 7.714]$ it can be rejected that $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

4 ☐ Since $0 \in [-0.0544, 7.543]$ it cannot be rejected that $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

5 ☐ Since $\hat{q}_{0.5,\text{male}} > \hat{q}_{0.5,\text{female}}$ it can be rejected that $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

Question X.3 (28)

Among the 10 official dietary guidelines from the Food Authority is one which states that a person should eat plenty of fish and preferably 350 g per week. This applies to both men and women.

In the following we will examine whether the sample from the dietary survey 2005-2008 provides evidence that dietary recommendation is met. Since dietary survey measured the daily intake of fish, it corresponds (a little simplified) to the following hypothesis:

$$H_0 : \mu = 50$$
$$H_1 : \mu \neq 50$$

Data is not normally distributed, mainly because of the many 0-observations, i.e. respondents who do not eat fish. To test the hypothesis the following R-code has been run (note that all of it may not be meaningful):

```
## Number of simulated samples
k <- 10000

simsamples <- replicate(k, sample(xMen, replace=TRUE))
simprms <- apply(simsamples, 2, mean)
quantile(simprms, c(0.005, 0.025, 0.975, 0.995))

##  0.5%  2.5% 97.5% 99.5%
##  22.0  22.8  28.1  29.0

simsamples <- replicate(k, sample(xWomen, replace=TRUE))
simprms <- apply(simsamples, 2, sd) - 50
quantile(simprms, c(0.005, 0.025, 0.975, 0.995))

##  0.5%  2.5% 97.5% 99.5%
## -31.7 -30.5 -22.0 -20.7

simsamples <- replicate(k, sample(c(xMen,xWomen), replace=TRUE))
simprms <- apply(simsamples, 2, mean) - 50
quantile(simprms, c(0.005, 0.025, 0.975, 0.995))

##  0.5%  2.5% 97.5% 99.5%
## -29.0 -28.4 -24.6 -23.9

simsamples <- replicate(k, sample(c(xMen,xWomen), replace=TRUE))
simprms <- apply(simsamples, 2, quantile, probs=0.90)
quantile(simprms, c(0.005, 0.025, 0.975, 0.995))

##  0.5%  2.5% 97.5% 99.5%
##  44.7  45.8  54.5  57.4
```

Using a significance level of $\alpha = 0.05$ what is the conclusion on the hypothesis that the dietary recommendation regarding intake of fish is met (both the conclusion and argumentation must be correct)?

1 ☐ Since $0 \notin [-28.4, -24.6]$ the null hypothesis is rejected, hence there is not a significant difference (i.e. $\mu \neq 50$) hence the recommendation is not met

2 ☐ Since $-50 \notin [-30.5, -22.0]$ the null hypothesis is rejected, hence there is not a significant difference (i.e. $\mu \neq 50$) hence the recommendation is not met

3 ☐ Since $50 \in [45.8, 54.5]$ it cannot be rejected that $\mu \neq 50$, hence the recommendation is met

4 ☐ Since $50 \notin [22.0, 29.0]$ the null hypothesis is rejected, hence there is not a significant difference (i.e. $\mu \neq 50$) hence the recommendation is not met

5 ☐ Since $44.7 - 50 = -5.3 < 0$ the null hypothesis is rejected, hence there is not a significant difference (i.e. $\mu \neq 50$) hence the recommendation is not met

Continues on page 27

Exercise XI

A new scanner for measuring the mass of the muscles in the body has been developed. It is much easier and faster to use compared to the otherwise available scanners. It is tested in an experiment to find out if it gets the similar results as the normally used scanner. For the experiment 20 randomly selected women aged 20 to 40 years have been scanned with both scanners.

The measured muscle mass in kg are read into R, such that the order of the women is the same in each vector:

```
## Sample from the new scanner
x1 <- c(37.6, 31.3, 22.9, 27.1, 41.8, 23.3, 24.5, 24.6, 32.1, 23.8, 33.9, 37.7,
        22.5, 38.6, 31.8, 21.0, 32.2, 17.1, 32.6, 15.5)
## Sample from the old scanner
x2 <- c(35.9, 28.7, 27.9, 29.8, 46.8, 24.2, 28.0, 23.7, 35.2, 26.4, 36.0, 40.9,
        24.8, 42.1, 32.5, 23.7, 36.7, 19.2, 37.7, 16.3)
```

and the following R code is run:

```
(mean(x1)-mean(x2)) + c(-1,1) * qt(0.975, df=38) * sd(x2-x1)/sqrt(40)

## [1] -2.9228 -1.5372

t.test(x1, x2)

##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -0.916, df = 37.8, p-value = 0.37
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.1591  2.6991
## sample estimates:
## mean of x mean of y
##    28.595    30.825

t.test(x1, x2, paired=TRUE)

##
##  Paired t-test
##
## data:  x1 and x2
## t = -4.61, df = 19, p-value = 0.00019
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -3.2429 -1.2171
## sample estimates:
## mean of the differences
##                   -2.23

t.test(x2-mean(x1))

##
##  One Sample t-test
##
## data:  x2 - mean(x1)
## t = 1.25, df = 19, p-value = 0.23
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.5043  5.9643
## sample estimates:
## mean of x
##      2.23

t.test(x1-mean(x2))

##
##  One Sample t-test
##
## data:  x1 - mean(x2)
## t = -1.35, df = 19, p-value = 0.19
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -5.6965  1.2365
## sample estimates:
## mean of x
##     -2.23
```

Question XI.1 (29)

What is the correct 95% confidence interval for the mean difference in measurement of muscle mass between the old and the new scanner (here rounded to three significant digits)?

1 ☐ [-2.92, -1.54]

2 ☐ [-7.16, 2.70]

3 ☐ [-3.24, -1.22]

4 ☐ [-1.50, 5.96]

5 ☐ [-5.70, 1.24]

Question XI.2 (30)

The accuracy of a scan depends on how much the person to be scanned is moving. In the new scanner the person is not fastened, so it is of interest to find out how large the variation of the measurements is. To investigate this the 20 women were scanned two times with the new scanner each and the difference in muscle mass (X_Δ) for each woman between the 2 scans were measured to:

i	1	2	3	4	5	6	7	8	9	10
$x_{\Delta,i}$	-0.62	1.12	0.24	2.07	-2.91	2.02	0.36	0.43	-1.77	-0.18
i	11	12	13	14	15	16	17	18	19	20
$x_{\Delta,i}$	1.02	0.85	0.43	1.39	2.82	-4.03	2.84	2.8	1.36	-0.07

The sample mean and standard deviation were calculated to

$$\bar{x}_\Delta = 0.509$$
$$s_{x_\Delta} = 1.82$$

Which of the following is a correct 99% confidence interval for the standard deviation of the measured muscle mass of the new scanner?

1 ☐ $0.509 \pm 2.86 \cdot \frac{1.35}{\sqrt{20}} = [-0.35, 1.37]$

2 ☐ $\left[\frac{20 \cdot 1.82^2}{38.6}, \frac{20 \cdot 1.82^2}{6.84}\right] = [1.72, 9.69]$

3 ☐ $0.509 \pm 2.09 \cdot \frac{1.35}{\sqrt{20}} = [-0.12, 1.14]$

4 ☐ $0.509 \pm 2.86 \cdot \frac{1.82}{\sqrt{19}} = [-0.69, 1.70]$

5 ☐ $\left[\sqrt{\frac{19 \cdot 1.82^2}{38.6}}, \sqrt{\frac{19 \cdot 1.82^2}{6.84}}\right] = [1.28, 3.03]$

THE EXAM IS FINISHED. Enjoy the late summer!