Technical University of Denmark *Page 1 of 23 pages.*

Written examination: 22. june 2017

Course name and number: **Introduction to Mathematical Statistics (02403)**

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 7 exercises. To answer the questions you need to fill in the prepared 30-question multiple choice form (on 6 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	I.4	I.5	II.1	III.1	III.2	III.3	III.4
Question	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Answer										

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

Exercise	VI.3	VI.4	VI.5	VI.6	VI.7	VI.8	VI.9	VI.10	VII.1	VII.2
Question	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Answer										

The questionnaire contains 23 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

In a double-blinded placebo-controlled clinical trial, 100 subjects with high long-term blood sugar were given a new type of medicine (termed 'active'). An additional 100 subjects were given placebo. The subjects received the medicine over 26 weeks and their long-term blood sugar was determined at the start of the trial (week 0) and after the last dose was given (week 26). The following table indicates average and standard deviation for long-term blood sugar, which is measured as the concentration of the substance HbA1c [%] in the blood.

Medicine	Week	Average	Standard dev.
Active	0	8.5	1.1
Placebo	0	8.6	1.2
Active	26	7.2	1.4
Placebo	26	8.1	1.5

Question I.1 (1)

A 95% confidence interval for the average long-term blood sugar after 26 weeks in the group receiving the new drug (active) is:

1 ☐ $7.2 \pm t_{0.95} \cdot \frac{1.4}{\sqrt{100}}$, where the t-distribution with 99 degrees of freedom is used

2 ☐ $(8.5 - 7.2) \pm t_{0.975} \cdot \frac{1.4}{10}$, where the t-distribution with 99 degrees of freedom is used

3 ☐ $7.2 \pm t_{0.975} \cdot \frac{1.4}{\sqrt{99}}$, where the t-distribution with 99 degrees of freedom is used

4 ☐ $7.2 \pm t_{0.975} \cdot 0.14$, where the t-distribution with 99 degrees of freedom is used

5 ☐ $(8.1 - 7.2) \pm t_{0.975} \cdot \frac{1.4}{10}$, where the t-distribution with 99 degrees of freedom is used

Continues on page 3

Question I.2 (2)

When the difference in the starting level is taken into account, the p-value for the test, of whether the effect (decrease in HbA1c) over 26 weeks of the new drug (active) differs from the placebo, can be found with the following R code (here `active_week26`, `active_week0`, `placebo_week26`, and `placebo_week0`, indicates vectors with the HbA1c levels for the individual subjects in the active group in week 26 and week 0, and in the placebo group at week 26 and 0, all sorted by subject ID):

1 ☐ `t.test(aktiv_week26, placebo_week26)`

2 ☐ `t.test(aktiv_week26, placebo_week26, paired=TRUE)`

3 ☐
```
t.test(aktiv_week26 - mean(c(placebo_week0, aktiv_week0)),
placebo_week26 - mean(c(placebo_week0, aktiv_week0)))
```

4 ☐ `t.test(aktiv_week26 - aktiv_week0, placebo_week26 - placebo_week0, paired=TRUE)`

5 ☐ `t.test(aktiv_week26 - aktiv_week0, placebo_week26 - placebo_week0)`

Question I.3 (3)

In order to calculate how many subjects are needed in a new study involving a new group of subjects with medium-high long-term blood sugar, there is a need to determine a confidence interval for the standard deviation of blood sugar reduction among subjects who received the new medicine. Assume that the standard deviation of blood sugar reduction in the sample consisting of subjects who received the new drug is calculated at 0.9.

What is the 95% confidence interval for the standard deviation under these assumptions?

1 ☐ $\left[\sqrt{\frac{(100-1)0.81}{128.42}},\ \sqrt{\frac{(100-1)0.81}{73.36}}\right]$

2 ☐ $\left[\sqrt{\frac{(200-1)0.81}{128.42}},\ \sqrt{\frac{(200-1)0.81}{73.36}}\right]$

3 ☐ $\left[\frac{(100-1)0.9^2}{128.42},\ \frac{(100-1)0.9^2}{73.36}\right]$

4 ☐ $\left[\sqrt{\frac{(100-1)0.81}{123.23}},\ \sqrt{\frac{(100-1)0.81}{77.05}}\right]$

5 ☐ $\left[\sqrt{\frac{(200-1)0.9^2}{239.96}},\ \sqrt{\frac{(200-1)0.9^2}{161.83}}\right]$

Continues on page 4

Question I.4 (4)

Based on the above analysis, the newly developed medicine is now required to be compared with a competitor's medicine in a so-called active-control study, also over 26 weeks. It is expected that subjects, who are randomised to the group receiving the competitor's medicine on average will experience a decrease in the blood sugar by 0.9 units (HbA1c [%]) over 26 weeks. While subjects who are randomised to the group receiving the newly developed medicine will experience the same lowering of blood sugar as in the previous placebo controlled study. It is also assumed that the standard deviation is 1.2.

It was desired to have 95% power to show a significant difference, under the specified assumptions, between the newly developed medicine and the competitors medicine with a 5% level of significance. Based on the following R-code, how many subjects should in total be recruitment for the study?

```
power.t.test(delta=-0.4, sd=1.2, power = 0.95, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 234.8696
##           delta = 0.4
##              sd = 1.2
##       sig.level = 0.05
##           power = 0.95
##     alternative = two.sided
##
## NOTE: n is number in *each* group

power.t.test(delta=0.9, sd=1.2, power = 0.95, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 47.18603
##           delta = 0.9
##              sd = 1.2
##       sig.level = 0.05
##           power = 0.95
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Continues on page 5

```r
power.t.test(delta=-0.4, sd=1.2, power = 0.95, sig.level = 0.05, type="paired")
```

```
##
##      Paired t test power calculation
##
##               n = 118.8917
##           delta = 0.4
##              sd = 1.2
##       sig.level = 0.05
##           power = 0.95
##     alternative = two.sided
##
## NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

1 ☐ 119

2 ☐ 470

3 ☐ 96

4 ☐ 48

5 ☐ 235

Continues on page 6

Question I.5 (5)

Following the active control study, a 95% confidence interval for the estimated treatment effect measured as the long-term blood sugar reduction (HbA1c [%]) after 26 weeks of treatment for the newly developed medicine relative to the competitor's medicine (that is, for the difference between treatment groups) is estimated to $[0.33; 0.82]$.

Judge which of the following statements is a reasonable conclusion on the study:

1 ☐ The newly developed medicine results in a significant reduction in blood sugar after 26 weeks, which with 95% confidence is between 0.33% and 0.82% HbA1c.

2 ☐ There is a significant difference between the two treatments (p-value < 0.05) and for a potential subject there is a maximum of 5% risk that the newly developed medicine will not reduce blood sugar more than the competitor's medicine.

3 ☐ Any subject, can with 95% confidence, expect that the newly developed medicine reduce their long-term blood sugar (HbA1c [%]) by 0.33% - 0.82% more than if they had taken the competitor's medicine.

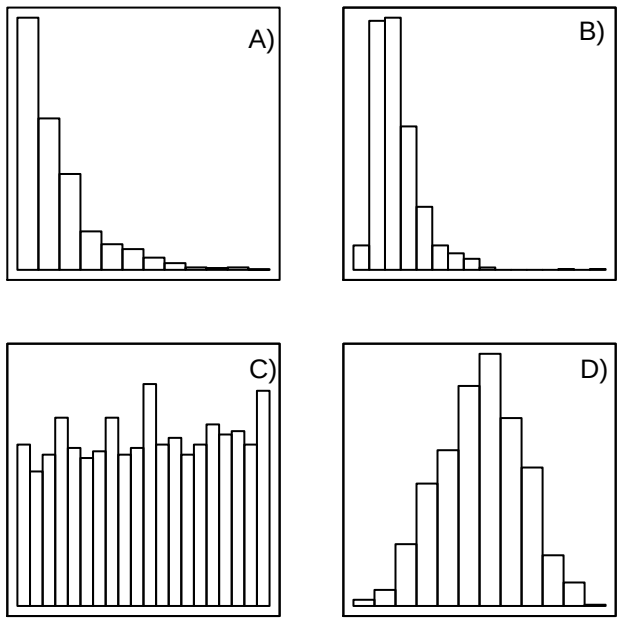
4 ☐ There is a significant difference between the two treatments (p-value < 0.05) and with 95% confidence it can be expected that the newly developed medicine on average will lower the blood sugar (HbA1c [%]) by 0.33% - 0.82% more than the competitor's medicine.

5 ☐ There is 95% probability that the newly developed medicine reduce the blood sugar more than the competitor's medicine.

Continues on page 7

Exercise II

The figure below shows histograms of data for realizations from different distributions.



Question II.1 (6)

Which of the following statements about the origin of data is probably correct?

1 ☐ B) exponentially distributed data, C) normally distributed data

2 ☐ C) exponentially distributed data, D) normally distributed data

3 ☐ A) exponentially distributed data, D) uniformly distributed data

4 ☐ A) log-normally distributed data, B) normally distributed data

5 ☐ A) exponentially distributed data, B) log-normally distributed data

Continues on page 8

Exercise III

The bitterness of wine depends on various factors in the wine making. During the wine making, the grapes are crushed into a mixture of juice and grapes shells and you can, among other things, vary the temperature of the mixture and whether the grape shells are in contact with the juice over a period of time.

In this experiment, 9 evaluators have assessed the bitterness (on a continuous scale) of white wines made under 2 conditions where the contact between shells and juices has been varied (`yes`, `no`). Each assessor has rated 4 wines under each of the 2 conditions. It is assumed that we can ignore the variation between assessors and that all observations are independent.

The following analysis intends to investigate the effect, if any, of contact on the bitterness of the white wines. The following analysis of variance, where some of the numbers have been masked, has been performed:

```
> anova(lm(response ~ contact, data=wine))
Analysis of Variance Table

Response: response
          Df  Sum Sq Mean Sq F value    Pr(>F)
contact    X  3226.7       X    X     0.003479
Residuals 70  X        352.7
```

Question III.1 (7)

The total sum of squared deviations (SST) is:

1 ☐ 24689

2 ☐ 27563

3 ☐ 3579.4

4 ☐ 27915.7

5 ☐ 2874

Continues on page 9

Question III.2 (8)

The test statistics and critical value ($\alpha = 0.05$) for the test of whether there are differences in the contact levels are:

1 ☐ The test statistic is 4.57 and the critical value is $F_{0.95}(2, 70) = 3.13$

2 ☐ The test statistic is 9.149 and the critical value is $F_{0.95}(1, 70) = 3.98$

3 ☐ The test statistic is 0.131 and the critical value is $F_{0.95}(1, 70) = 3.98$

4 ☐ The test statistic is 9.149 and the critical value is $F_{0.975}(1, 70) = 5.25$

5 ☐ The test statistic is 6.38 and the critical value is $F_{0.95}(2, 70) = 3.13$

Question III.3 (9)

With the averages of the two levels of contact calculated to $\mu_{contact:no} = 40.52$ and $\mu_{contact:yes} = 53.91$ the 95% confidence interval (post-hoc) for the difference between the two levels of contact is:

1 ☐ $13.39 \pm 1.99\sqrt{24689(\frac{1}{36} + \frac{1}{36})}$

2 ☐ $13.39 \pm 1.96\sqrt{24689(\frac{1}{36} + \frac{1}{36})}$

3 ☐ $13.39 \pm 2.29\sqrt{352.7(\frac{1}{36} + \frac{1}{36})}$

4 ☐ $13.39 \pm 2.29\sqrt{\frac{27915.4}{70}(\frac{1}{36} + \frac{1}{36})}$

5 ☐ $13.39 \pm 1.99\sqrt{352.7(\frac{1}{36} + \frac{1}{36})}$

Question III.4 (10)

At significance level $\alpha = 0.05$ what is the conclusion of the study (both conclusion and argument must be correct)?

1 ☐ Contact has a significant influence on the bitterness, since $0.0035 < 0.05$.

2 ☐ Contact has a significant influence on the bitterness, since $3226.7 < 70 \cdot 352.7$.

3 ☐ Contact has a significant influence on the bitterness, since $3226.7 > 352.7$.

4 ☐ Contact does not have a significant influence on the bitterness, since $0.0035 < 0.05$.

5 ☐ Contact does not have a significant influence on the bitterness, since $3226.7 > 352.7$.

Continues on page 10

Question III.5 (11)

It is now reported that "repetitions" at different contact levels is due to different temperature levels. Which of the following statements about a new analysis of variance table that takes into account the different temperature levels is correct (SSQ is short for sum of squares of Q where Q can be residuals or treatment effect)?

1 ☐ MS for contact will remain unchanged, but the p-value for contact will change.

2 ☐ SSQ and MS and therefore also the p-value for contact will remain unchanged.

3 ☐ SSQ and MS for the residual term will grow but the p-value for contact will remain unchanged.

4 ☐ SSQ and MS for the residual term will remain unchanged.

5 ☐ SSQ and MS for both contact and the residual term will remain unchanged and therefore the p-value for contact will also remain unchanged.

Continues on page 11

Exercise IV

An object is thrown (from the coordinate $(x, y) = (0, 0)$), it is assumed that its path follows the projectile motion without air resistance. The throwing length is given by the expression

$$x_{max} = \frac{v_0^2 \sin(2\theta)}{g}$$

where g is the gravity acceleration, θ is the throwing angle and v_0 is the initial speed. Now suppose that the initial speed and throwing angle are both uncertain, more specifically, that the mean values are μ_v and μ_θ for v_0 and θ, respectively, and the variances are σ_v^2 and σ_θ^2. It is further assumed that throwing angle and initial speed are independent, and that g is known without uncertainty. As a help with the task, it is stated that $\frac{\partial \sin(z)}{\partial z} = \cos(z)$.

Question IV.1 (12)

Which of the following terms is the law of error propagation approximation to the variance of the throwing length?

1 ☐ $\left(\frac{2\mu_v \sin(2\mu_\theta)}{g}\right)^2 \sigma_v^2 + \left(\frac{2\mu_v^2 \cos(2\mu_\theta)}{g}\right)^2 \sigma_\theta^2$

2 ☐ $\left(\frac{2\sigma_v^2 \sin(2\sigma_\theta^2)}{g}\right)^2 \mu_v^2 + \left(\frac{2\sigma_v^2 \cos(2\sigma_\theta^2)}{g}\right)^2 \mu_\theta^2$

3 ☐ $\frac{4\mu_v \cos(2\mu_\theta)}{g}(\sigma_v^2 + \sigma_\theta^2)$

4 ☐ $\left(\frac{4\mu_v \cos(2\mu_\theta)}{g}\right)^2 (\sigma_v^2 + \sigma_\theta^2)$

5 ☐ $\frac{2\sigma_v^2 \sin(2\sigma_\theta^2)}{g}$

For specific distribution assumptions, it is possible to construct simulation-based distributions and thus intervals, assuming that the angle (measured in radians) follows a uniform distribution in the range $[\frac{\pi}{8}; \frac{3\pi}{8}]$, and that the initial speed follows a normal distribution with mean 20 and standard deviation $\sigma_v = 2$, and further assumed that $g = 9.81$. As a help to find the interval, the following R-code have been executed (all of which does not necessarily make sense):

Continues on page 12

```r
g <- 9.81
v0 <- 20
sigma.v <- 2
k <- 100000
theta <- runif(k, pi/8, pi * 3/8)
v <- rnorm(k, mean=v0, sd=sigma.v)
xm <- v^2 * sin(2 * theta) / g
quantile(xm, prob = c(0.005,0.01,0.025,0.975,0.99,0.995))

##     0.5%       1%     2.5%    97.5%      99%    99.5%
## 19.15289 20.43880 22.53632 54.52955 58.23360 60.63147

mean(xm) + c(-1, 1) * qnorm(0.995) * sd(xm)

## [1] 15.86690 58.26297

mean(xm) + c(-1, 1) * qnorm(0.99)  * sd(xm)

## [1] 17.92003 56.20984

mean(xm) + c(-1, 1) * qnorm(0.975) * sd(xm)

## [1] 20.93523 53.19465
```

Question IV.2 (13)

Which of the following intervals is the simulation-based interval (x_1, x_2) such that $P(X_{max} < x_1) = P(X_{max} > x_2)$ and $P(x_1 < X_{max} < x_2) = 0.99$?

1 ☐ $(19.15; 60.63)$

2 ☐ $(20.44; 58.23)$

3 ☐ $(17.92; 56.21)$

4 ☐ $(15.87; 58.26)$

5 ☐ $(20.94; 53.19)$

Continues on page 13

Exercise V

This exercise is about making a model for predicting CO2 emissions from electricity generation. In power plants materials are burned to generate electricity and in this combustion CO2 is emitted. It is desirable to predict the CO2 level in electricity generation, so that electricity consumption can be moved to periods with the lowest CO2 level and thus minimise emissions.

Every day, CO2 emissions per kWh electricity produced are calculated in Denmark based on data from ENTSO-E (European Network of Transmission System Operators for Electricity).

Data set for determining a good model for predicting CO2 consists of average hourly values of the following variables, the last three being 24 Hour forecasts:

Variable	Description	Range	Unit
`co2intensity`	CO2 intensity	$[113, 566]$	gCO2eq/kW
`windspeed`	Windspeed	$[1.7, 11.4]$	m/s
`importDE`	Power-import from Germany	$[-2300, 1845]$	kW
`generation`	Generated power	$[920, 3910]$	kW

Data is read into R in a data.table `X`.

A linear regression model with an intercept and the three explanatory variables is fitted

$$Y_{\text{co2intensity},i} = \beta_0 + \beta_{\text{w}} X_{\text{windspeed},i} + \beta_{\text{im}} X_{\text{importDE},i} + \beta_{\text{g}} X_{\text{generation},i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

and the result is:

```
##
## Call:
## lm(formula = "co2intensity ~ windspeed + importDE + generation",
##     data = X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.232 -26.026  -5.613  19.376 153.485
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 263.146473   8.542834   30.80   <2e-16 ***
## windspeed   -17.202953   0.761038  -22.61   <2e-16 ***
## importDE     -0.031120   0.001192  -26.11   <2e-16 ***
## generation    0.076791   0.002926   26.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 39.73 on 721 degrees of freedom
## Multiple R-squared:  0.7855,Adjusted R-squared:  0.7846
## F-statistic:   880 on 3 and 721 DF,  p-value: < 2.2e-16
```

In the following it is assumed that the assumption about independent errors is fulfilled.

Question V.1 (14)

Based on the usual model selection procedure (using significance level $\alpha = 0.05$), should this model be reduced (both conclusion and argument must be correct)?

1 ☐ Yes, since the explained variation is higher than the chosen significance level

2 ☐ No, since the explained variation is higher than the chosen significance level

3 ☐ Yes, since $\hat{\sigma}/df = 39.73/721 = 0.0551$ is higher than the chosen significance level

4 ☐ No, since $\hat{\sigma}/df = 39.73/721 = 0.0551$ is higher than the chosen significance level

5 ☐ No, since all the coefficients in the model are significantly different from zero on the chosen significance level

Question V.2 (15)

Which of the following statements is not correct (the false statement should be identified)?

1 ☐ The model predicts that the CO2-intensity is 417 gCO2eq/kW in conditions with no wind, no power import from Germany and a power generation at 2000 kW per hour

2 ☐ The standard deviation of the errors is estimated to 39.73

3 ☐ The model has explained 79% of the variation

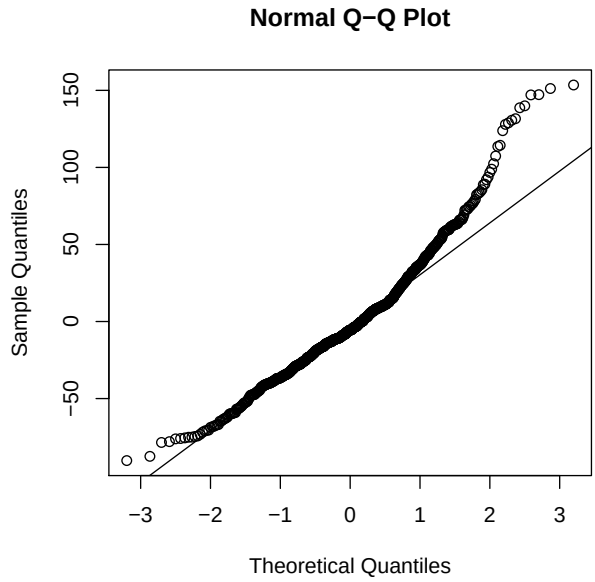
4 ☐ The most important of the three explanatory variables is the wind speed, since $|\hat{\beta}_{\text{im}}| < |\hat{\beta}_{\text{g}}| < |\hat{\beta}_{\text{w}}|$

5 ☐ It is estimated that when the wind speed increase, then the CO2 intensity decrease

Continues on page 15

Question V.3 (16)

In the validation of the model a Q-Q normal plot of the residuals are generated:



Which of the following conclusions can be drawn based on this plot?

1 ☐ The assumption that the errors ε_i are i.i.d. is not fulfilled

2 ☐ The assumption that the errors ε_i are i.i.d. is fulfilled

3 ☐ The assumption that the errors ε_i are normal distributed is not fulfilled

4 ☐ The assumption that the errors ε_i are normal distributed is fulfilled

5 ☐ It can be rejected that there is a linear relation between the CO2-intensity and the explanatory variables

Continues on page 16

Question V.4 (17)

In the matrix formulation of a linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

is the so-called design matrix $\boldsymbol{X}$. In the used model and data set, what are the dimensions of this matrix?

1 ☐ 2884 rows and 4 columns

2 ☐ 725 rows and 4 columns

3 ☐ 4 rows and 4 columns

4 ☐ 5 rows and 4 columns

5 ☐ 4 rows and 5 columns

Question V.5 (18)

Now the prediction for a new point is wanted. If the new forecasted values of the inputs are in the vector $\boldsymbol{x}_{\text{new}}$, and the model is formulated on matrix form, as in the previous question. How is the prediction of the CO2-intensity calculated?

1 ☐ $\hat{Y}_{\text{co2intensity,new}} = \boldsymbol{x}_{\text{new}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$

2 ☐ $\hat{Y}_{\text{co2intensity,new}} = \boldsymbol{x}_{\text{new}}\hat{\boldsymbol{\beta}} + \varepsilon_{\text{new}}$

3 ☐ $\hat{Y}_{\text{co2intensity,new}} = \text{V}(\boldsymbol{x}_{\text{new}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} + \varepsilon_{\text{new}})$

4 ☐ $\hat{Y}_{\text{co2intensity,new}} = \text{V}(\boldsymbol{x}_{\text{new}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y})$

5 ☐ $\hat{Y}_{\text{co2intensity,new}} = \text{V}(\boldsymbol{x}_{\text{new}}\hat{\boldsymbol{\beta}})$

Continues on page 17

Exercise VI

In a production line where lids are produced, one wants to investigate what the probability of errors is. For this purpose, 229 samples have been taken with 770 lids in each sample. The table below shows the result of sampling.

No. of defective	0	1	2	3	4	5	6	7	8	9
No. of samples	131	38	28	11	4	5	5	2	3	2

Thus, for example, 38 samples (with each 770 lids) have been observed with one defective lid in each. Assuming that the probability of error is the same in each of the 229 samples, the estimate of probability of error can be calculated to

$$\begin{aligned}\hat{p} =& \frac{38 + 2 \cdot 28 + 3 \cdot 11 + 4 \cdot 4 + 5 \cdot 5 + 6 \cdot 5 + 7 \cdot 2 + 8 \cdot 3 + 9 \cdot 2}{770 \cdot 229} \\ =& 0.00144\end{aligned}$$

Question VI.1 (19)

What is the average number of defective lids per sample?

1 ☐ 4.5

2 ☐ 22.9

3 ☐ 0.33

4 ☐ 1.109

5 ☐ 254

Question VI.2 (20)

If `defective` and `samples` denote first and second row, respectively, in the table above and `m` denotes the empirical mean of the number of defective lids per sample. Which of the following R commands will then calculate the empirical variance for the number of defective lids per sample?

1 ☐ `sum((defective - m)^ 2 * samples)/ (229 * 770)`

2 ☐ `sum((defective - m)^ 2) / 9`

3 ☐ `sqrt(sum((defective - m)^ 2 * samples)/ (229 * 770))`

4 ☐ `sum((defective - m)^ 2 * samples) / 228`

5 ☐ `sum((samples - m)^ 2) / 9`

Continues on page 18

Question VI.3 (21)

From which interval will IQR (inter quartile range) for the number of defective lids per sample be calculated?

1 ☐ $[3; 28]$

2 ☐ $[2.25; 6.75]$

3 ☐ $[0; 2]$

4 ☐ $[3.25; 23.75]$

5 ☐ $[2; 7]$

Question VI.4 (22)

Assuming the probability that a lid is defective is $p = 0.00144$, what is the expected number of samples with 5 defective lids, when there are a total of 229 samples each with 770 lids?

1 ☐ 0.00458

2 ☐ 0.0000225

3 ☐ 0.0173

4 ☐ 1.05

5 ☐ 3.53

Question VI.5 (23)

What is a 95% confidence interval for the probability of defect on a single lid?

1 ☐ $0.00144 \pm z_{0.975}\sqrt{\frac{0.00144(1-0.00144)}{770 \cdot 229}} = [0.00126; 0.00162]$

2 ☐ $0.00144 \cdot 229 \pm z_{0.975}\sqrt{\frac{0.00144(1-0.00144)\cdot 229}{770}} = [0.289; 0.370]$

3 ☐ $\frac{229}{770} \pm z_{0.975}\sqrt{\frac{229/770(1-229/770)}{770}} = [0.265; 0.330]$

4 ☐ $0.00144 \cdot \frac{770}{229} \pm z_{0.975}\sqrt{\frac{0.00144(1-0.00144)}{770 \cdot 229}} = [0.00466; 0.00502]$

5 ☐ $0.00144 \pm z_{0.975}\frac{\sqrt{229/770(1-229/770)}}{770} = [0.000276; 0.00260]$

Continues on page 19

Question VI.6 (24)

A new sampling is now planned. How many samples with each 770 lids should be taken if you want a margin of error of 0.0001 with a significance level $\alpha = 0.05$ and the observed fraction of defective (0.00144) is used as a scenario?

1 ☐ 384

2 ☐ 2413

3 ☐ 545

4 ☐ 506

5 ☐ 718

It is now decided that the distribution assumption is questionable and therefore a simulation-based 95% confidence interval for the defect probability is constructed. For this purpose, the following R code has been executed (all of which does not necessarily make sense) in the code `new.samp` indicates a vector with 131 zeros, 38 ones, etc.

```
k <- 100000
defective <- c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9)
samples <- c(131, 38, 28, 11, 4, 5, 5, 2, 3, 2)
new.samp <- rep(defective, samples)

sim <- replicate(k, sample(new.samp, replace = TRUE))

sim.p <- apply(sim, 2, sum) / (229 * 770)

quantile(sim.p, c(0.025, 0.05, 0.95, 0.975))

##        2.5%          5%         95%       97.5%
## 0.001139908 0.001185278 0.001707027 0.001758067

c(mean(sim.p), sd(sim.p))

## [1] 0.0014406186 0.0001587641
```

Continues on page 20

```
################################################
##
################################################
sim2 <- replicate(k, sample(samples, replace=TRUE))

sim.p2 <- apply(sim2, 2, sum) / 770

quantile(sim.p2, c(0.025, 0.05, 0.95, 0.975))

##       2.5%         5%        95%      97.5%
## 0.07012987 0.08831169 0.58961039 0.63896104

c(mean(sim.p2), sd(sim.p2))

## [1] 0.2965242 0.1553125
```

Question VI.7 (25)

The simulation-based confidence interval becomes:

1 ☐ $[0.00114; 0.00176]$

2 ☐ $[0.0883; 0.590]$

3 ☐ $[0.041; 0.552]$

4 ☐ $[0.0701; 0.639]$

5 ☐ $[0.00119; 0.00171]$

A similar sample (i.e. 229 samples with 770 lids in each) is now taken from another production line, the result is given in the table below

No. of defective	0	1	2	3	4	5	6	7	8	9
No. of samples	73	84	46	17	5	3	1	0	0	0

The estimate for the defect probability can be calculated in the same way as above and the result is $\hat{p}_2 = 0.00152$. One now wishes to investigate whether there is a difference in the defect probability in the 2 production lines.

Continues on page 21

Question VI.8 (26)

What is the usual test statistics and conclusion for the difference in defect probability in the 2 production lines (using the significance level $\alpha = 0.05$)?

1 ☐ $\frac{0.00152-0.00144}{\sqrt{0.00148(1-0.00148)\cdot\frac{2}{229\cdot 770}}} = 0.618$, and no difference can documented since $|0.618| < 1.96$

2 ☐ $\frac{0.00152-0.00144}{\sqrt{0.00148(1-0.00148)\cdot\frac{2}{229}}} = 0.0223$, and there is a significant difference since $0.0223 < 0.05$

3 ☐ $\frac{0.00152-0.00144}{\sqrt{0.00148(1-0.00148)\cdot\frac{2}{770}}} = 0.0408$, and there is a significant difference since $0.0408 < 0.05$

4 ☐ $\frac{0.00152-0.00144}{\sqrt{0.00148(1-0.00148)\cdot\left(\frac{1}{229}+\frac{1}{770}\right)}} = 0.0276$, and there is a significant difference since $|0.0276| < 1.96$

5 ☐ $\frac{0.00152-0.00144}{\sqrt{0.00148(1-0.00148)\cdot\frac{2}{229}}} = 0.0223$, and there is a significant difference since $|0.0223| < 1.96$

Regardless of the outcome of the previous questions, it is decided to make a more general comparison of the 2 distributions using tests in contingency tables. For this purpose, the table below has been prepared

Defective	line 1	line 2
0	131	73
1	38	84
2	28	46
3	11	17
$4-5$	9	8
>5	12	1

Continues on page 22

Question VI.9 (27)

What is the contribution to the usual test statistics for the 4-5 defective group (sum of contributions from both production lines)?

1 ☐ $\frac{1}{17}$

2 ☐ $\frac{9^2}{8.5} + \frac{8^2}{8.5}$

3 ☐ $\frac{2}{17}$

4 ☐ $\frac{8}{8.5} + \frac{9}{8.5}$

5 ☐ $\frac{1}{34}$

Question VI.10 (28)

The usual test statistic is now calculated to 48.865, what is the p-value and conclusion for the test of the two distributions being equal (use confidence level $\alpha = 0.05$)?

1 ☐ The p-value become $4.31 \cdot 10^{-7}$, and it cannot be rejected that the distributions are equal.

2 ☐ The p-value become $2.36 \cdot 10^{-9}$, and we can reject that the distributions are equal.

3 ☐ The p-value become 0.0201, and it cannot be rejected that the distributions are equal.

4 ☐ The p-value become $4.31 \cdot 10^{-7}$, and it cannot be rejected that the distributions are equal.

5 ☐ The p-value become 0.0201, and we can reject that the distributions are equal.

Continues on page 23

Exercise VII

Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, and assume that X and Y are independent.

Question VII.1 (29)

If we assume that $\mu_1 = \mu_2 = 0$, what is $Var(X^2 + Y^2)$ then?

1 ☐ $2\sigma_1^4 + 2\sigma_2^4$

2 ☐ $4\sigma_1^2 + 4\sigma_1^2$

3 ☐ $4\sigma_1^2 + 4\sigma_1^2 + \sigma_1^2\sigma_2^2$

4 ☐ $\sigma_1^4 + \sigma_2^4$

5 ☐ $2\sigma_1^2 + 2\sigma_1^2$

Question VII.2 (30)

If we now assume that $\sigma_1^2 = 2\sigma_2^2$, what is $P((X - \mu_1)^2 > (Y - \mu_2)^2)$ then?

1 ☐ $P(Z > 0)$, where Z follows a χ^2-distribution with 1 degree of freedom.

2 ☐ $P\left(Z > \frac{1}{2}\right)$, where Z follows a F-distribution with 1 og 1 degrees of freedom.

3 ☐ $P\left(Z > \frac{1}{2}\right)$, where Z follows a t-distribution with 2 degrees of freedom.

4 ☐ $P(Z > 2)$, where Z follows a χ^2-distribution with 1 degree of freedom.

5 ☐ $P\left(|Z| < \frac{1}{2}\right)$, where Z follows a t-distribution with 1 degree of freedom.

SÆTTET ER SLUT. God sommer!